N-glycan chitobiose core biosynthesis by Agl24 strengthens the hypothesis of an archaeal origin of the eukaryal N-glycosylation

Benjamin H. Meyer[1,2,4 #], Ben A. Wagstaff[2], Panagiotis S. Adam[3],

Sonja-Verena Albers[4], and Helge C. Dorfmueller[2#]

[1] Molecular Enzyme Technology and Biochemistry, Environmental Microbiology and Biotechnology, Faculty of Chemistry University Duisburg-Essen, Germany;
[2] Division of Molecular Microbiology, School of Life Sciences, University of Dundee, Dundee, DD1 5EH UK;
[3] Group for Aquatic Microbial Ecology, Environmental Microbiology and Biotechnology, Faculty of Chemistry University Duisburg-Essen, Germany;
[4] Molecular Biology of Archaea, Faculty of Biology, University of Freiburg, 79104 Freiburg, Germany

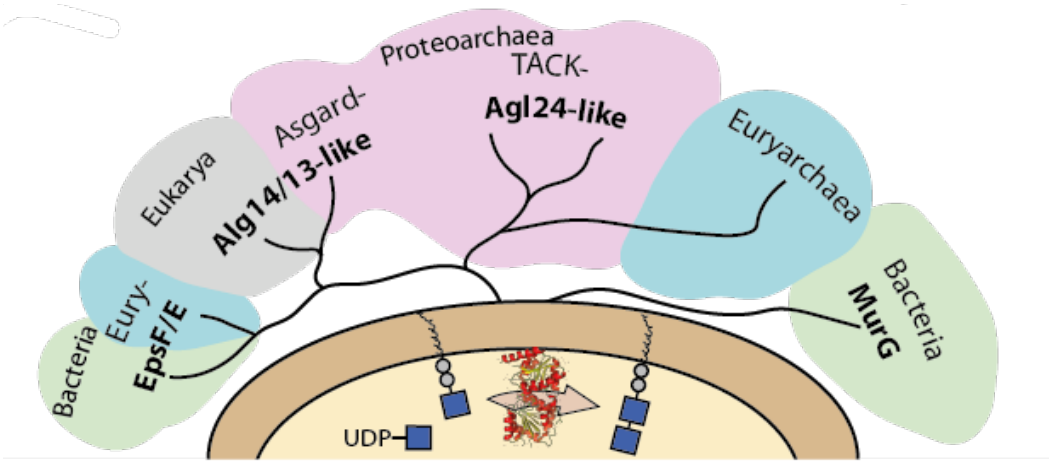
[#]Correspondence to B. H. Meyer, phone: +49 201 183 7084, e-mail: benjamin.meyer@uni-due.de and H. C. Dorfmueller, +44 1382 386203, e-mail: hczdorfmueller@dundee.ac.uk

Key words: **chitobiose, Archaea, N-glycosylation, N-glycan, Agl24, Alg13, Agl14, MurG**

Abbreviation: Agl, archaeal glycosylation enzyme; Alg, Asparagine linked glycosylation; CAZy, Carbohydrate-Active enZymes database; Dol, dolichol; Dol-P, dolichol-phosphate; DPANN, archaeal superphylum containing Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota, Nanohaloarchaeota, Woesearchaeota and Pacearchaeota; Glc, glucose; GlcA, glucuronic acid; GlcNAc, N-acetylglucosamine; GFP, green fluorescent protein; GT, glycosyltransferase; LLO, lipid-linked oligosaccharide; MALDI-MS, matrix-assisted laser desorption ionization mass spectrometry; OST, oligosaccharyltransferase; SDS-PAGE, sodium dodecyl sulfate–polyacrylamide gel electrophoresis; TACK, archaeal superphylum containing Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota; TMD, transmembrane domain; UDP, uridine diphosphate; Und, undecaprenol

Short title: Agl24 syntheses the N-glycan chitobiose core

Cover Art:



Abstract

Protein N-glycosylation is the most common posttranslational modifications found in all three domains of life. The crenarchaeal N-glycosylation begins with the synthesis of a lipid-linked chitobiose core structure, identical to that in eukaryotes. Here, we report the identification of a thermostable archaeal beta-1,4-N-acetylglucosaminyltransferase, named archaeal glycosylation enzyme 24 (Agl24), responsible for the synthesis of the N-glycan chitobiose core. Biochemical characterization confirmed the function as an inverting β-D-GlcNAc-(1→4)-α-D-GlcNAc-diphosphodolichol glycosyltransferase. Substitution of a conserved histidine residue, found also in the eukaryotic and bacterial homologs, demonstrated its functional importance for Agl24. Furthermore, bioinformatics and structural modeling revealed strong similarities between Agl24 and both the eukaryotic Alg14/13 and a distant relation to the bacterial MurG, which catalyze the identical or a similar process, respectively. Our data, complemented by phylogenetic analysis of Alg13 and Alg14, revealed similar sequences in Asgardarchaeota, further supporting the hypothesis that the Alg13/14 homologs in eukaryotes have been acquired during eukaryogenesis.

Highlights

- First identification and characterization of a thermostable β-D-GlcNAc-(1→4)-α-D-GlcNAc-diphosphodolichol glycosyltransferase (GT family 28) in Archaea.
- A highly conserved histidine, within a GGH motif in Agl24, Alg14, and MurG, is essential for function of Agl24.
- Agl24-like homologs are broadly distributed among Archaea.
- The eukaryotic Alg13 and Alg14 are closely related to the Asgard homologs, suggesting their acquisition during eukaryogenesis.

Introduction

Asparagine (N)-linked glycosylation is one of the commonest co- and posttranslational protein modifications found in all three domains of life (Larkin and Imperiali, 2011, Jarrell et al., 2014, Nothaft and Szymanski, 2010) In Eukarya, N-glycosylation is an essential process and is evolutionarily highly conserved from yeast to humans (Lehle et al., 2006). It is estimated that more than half of all eukaryotic proteins are glycoproteins (Apweiler et al., 1999, Zielinska et al., 2010). N-glycosylation is required for correct protein folding and stability, intra- and extracellular recognition of protein targets and for enzyme activity (Caramelo and Parodi, 2007, Helenius and Aebi, 2004, Varki, 1993). Therefore, the biological functions of protein glycosylation range from relatively minor to crucial for survival of an

62 organism. As such, congenital disorders of glycosylation in humans are often lethal or cause severe
63 phenotypes and diseases ranging from intellectual disability, growth retardation, cardiac anomalies,
64 and early death (Reily et al., 2019). Although it was long thought that N-glycosylation was restricted to
65 Eukarya, N-glycosylation pathways have now been characterized in all three domains of life. Bacterial
66 N-glycosylation seems to be restricted to delta/epsilon proteobacteria (Nothaft and Szymanski, 2013,
67 Whitfield et al., 2017). In contrast, all sequenced Archaea possess the N-glycan
68 oligosaccharyltransferase AglB, with the exceptions of *Aeropyrum pernix* and *Methanopyrus* species
69 (Jarrell et al., 2014, Nikolayev et al., 2020). The common feature of N-glycosylation is the assembly of
70 an oligosaccharide onto a lipid by the action of specific glycosyltransferases (GTs), which use either
71 nucleotide- or lipid-activated sugar donors. The assembled lipid-linked oligosaccharide (LLO) is flipped
72 across the cytoplasmic/ER membrane. The glycan is then either enlarged or directly transferred onto
73 a target protein by an oligosaccharyltransferase (OST). Target proteins contain a specific asparagine
74 residue found within the conserved sequon (N-X-S/T in Eukarya and Archaea or D/E-X_1-N-X_2-S/T in
75 Bacteria, where X is not proline) recognized by the OST.

76 Archaeal N-glycosylation displays a mosaic of bacterial and eukaryal features. Like Bacteria, Archaea
77 use only a single OST subunit, whereas Eukarya require an OST complex of nine subunits, to transfer
78 the LLO onto the target protein. In contrast, Archaea use the shorter version of the recognition sequon
79 (N-X-S/T) as well as dolichol-phosphate (Dol) as lipid, as in Eukarya, whereas Bacteria use
80 undecaprenyl-phosphate. Archaea also display unique N-glycosylation features, *e.g.* the assembly of
81 the N-glycan onto Dol-phosphate in the Euryarchaeota, while the Crenarchaeota use Dol-
82 pyrophosphate as in eukaryotes (Taguchi et al., 2016).

83 Currently, more than 15 different archaeal N-glycan structures have been characterized (Jarrell et al.,
84 2014, Albers et al., 2017). These archaeal N-glycans display unique structures and sugar compositions
85 and based on the limited characterization of archaeal N-glycans, it is plausible to assume that Archaea
86 possess an enormous range of structural and composition diverse N-glycans. However, closely related
87 species *e.g.,* of the thermophilic crenarchaeal Sulfolobales family, retain a conserved N-glycan core
88 structure but differentiated by their terminal sugar residues. The glycosylation pathway of *Sulfolobus*
89 *acidocaldarius*, belonging to Sulfolobales, has been partially characterized. Secreted proteins, such as
90 the cell wall surface-layer (S-layer) protein and motility filament forming protein ArlB, are N-
91 glycosylated with a tribranched hexasaccharide. This N-glycan has the composition: (Glc-β1-4-
92 sulfoquinovoseβ-1-3)(Man α-1-6)(Man α-1-4)GlcNAc-β1-4-GlcNAc-β-Asp (Peyfoon et al., 2010,
93 Zahringer et al., 2000). Interestingly, the N-glycan core structure is composed of a chitobiose (GlcNAc-
94 β1-4-GlcNAc), which is identical to the core structure of all eukaryal N-glycans. Members of the
95 Sulfolobales family are the only Archaea reported to display this eukaryal N-linked chitobiose core

96 structure (van Wolferen et al., 2020, Palmieri et al., 2013). Based on this identical N-glycan core
97 structure, it has been suggested that the *Sulfolobus* N-glycan biosynthesis pathway may be similar to
98 the eukaryal process (Meyer and Albers, 2013). Indeed, a homolog of the yeast Alg7 enzyme, which
99 initiates eukaryal N-glycosylation by transferring a GlcNAc-P from UDP-GlcNAc to Dol-P to generate
100 the precursor Dol-PP-GlcNAc, was identified in *S. acidocaldarius* as AlgH (Meyer et al., 2017). This AglH
101 showed not only a similar topology and highly conserved amino acid motifs to Alg7, but also
102 functionally complemented a conditional yeast Alg7 knockout (Meyer et al., 2017). This is a rare
103 example of cross-domain complementation that further strengthens the hypothesis of a common
104 origin of eukaryotic and archaeal N-glycosylation.

105 Here we report the identification of the β-1,4-N-acetylglucosamine-transferase Agl24, which shows
106 conserved amino acid motifs and apparent structural similarities with the eukaryal Alg13/14 and
107 bacterial MurG functional homologs. Activity assays, along with HPLC, MALDI, and NMR analyses
108 confirmed the function of Agl24 as catalyzing the second N-glycan biosynthesis step by generating the
109 lipid-linked chitobiose core. Our extensive bioinformatics analyses support the hypothesis that the
110 eukaryotic N-glycosylation pathway originates from an archaeal ancestor.

111

112 **Results**

113 **Identification of a Dol-PP-GlcNAc UDP-GlcNAc GlcNAc GT candidate in *S. acidocaldarius***

114

115 To identify the glycosyltransferase that conducts this first committed synthetic step in the N-
116 glycosylation process in *S. acidocaldarius*, we searched for similar biosynthesis processes to identify
117 potential homologs. The bacterial enzyme MurG transfers a GlcNAc residue from UDP-GlcNAc to the
118 C4 hydroxyl group of MurNAc-(pentapeptide)-PP-undecaprenol to produce the lipid-linked β-(1,4)
119 disaccharide known as lipid II (Ha et al., 1999, Men et al., 1998). Although MurG does not synthesize
120 chitobiose like *S. acidocaldarius*, the enzyme uses a similar lipid-linked acceptor and an identical sugar
121 donor (UDP-GlcNAc) and therefore presents a valid candidate for evaluation. MurG has been classified
122 by the Carbohydrate Active enZyme database (http://www.cazy.org/) as a family 28 GT(Lombard et
123 al., 2014). Up to now, only members of the archaeal families *Methanobacteriaceae* and
124 *Methanopyraceae*, which synthesize pseudomurein as their cell wall structure (Albers and Meyer,
125 2011, Meyer and Albers, 2020), were found to have a GT28 family homolog in the CAZy database
126 (Lombard et al., 2014).
127 Another candidate homolog is the eukaryal protein complex formed by Alg13 and Alg14 that transfers
128 GlcNAc from UDP-GlcNAc to the C4 hydroxyl of GlcNAc-PP-dolichol to produce the β-(1-4) chitobiose
129 core of the eukaryal N-glycan. Alg13 and Alg14 are classified as GT 1 family members but are distantly
130 related to GT28 family enzymes (Lombard et al., 2014). Alg13 and Alg14 form the C- and N-terminal
131 domains, respectively, of a functional heterodimeric GT enzyme (Fig 2); Alg14 contains the acceptor-
132 binding site and the membrane association domain, while Alg13 possesses the UDP-GlcNAc-binding
133 pocket. The interaction of these two enzymes is required for the catalytic activity, as the active site lies
134 in the cavity formed by the two proteins (Gao et al., 2005, Bickel et al., 2005, Chantret et al., 2005).
135 Although members of the GT 1 family are most prominent in Archaea, until now no GT1 family member
136 has been classified in Sulfolobales. There is no meaningful shared sequence identity between Alg14-
137 13 and MurG and they are subsequently annotated in different GT families. However, both enzymes
138 use UDP-GlcNAc to GlcNAcylate a lipid-linked sugar acceptor and therefore represent valid candidates
139 for a bioinformatics search in *S. acidocaldarius.*
140 Using the *E. coli* MurG protein sequence (WP_063074721.1; EC 2.4.1.227) as template for a Delta Blast
141 search (Boratyn et al., 2012), various homologs were identified in *S. acidocaldarius* with very low
142 sequence identity of only 10–17% (Table S1). The anchoring subunit Alg14 from *Saccharomyces*
143 *cerevisiae* S288C (NP_009626.1) was used as bait for a homology search but no protein candidate was
144 identified below the threshold of 0.005. However, when either the catalytic subunit Alg13 from
145 *Saccharomyces cerevisiae* S288C (NP_011468.1; EC 2.4.1.141) or the fusion of both Alg14 and Alg13
146 sequences was used as a search model, a single Sulfolobales homolog was identified with 17%

147 sequence identity (Table S1). Thus, the most likely candidate for the β-1,4-N-acetylglucosamine
148 transferase was found to be Saci1262 (Uniprot: Q4J9C3), now named Agl24, a hypothetical 327 amino
149 acid protein. In contrast to Euryarchaeota, where the genes of the N-glycosylation enzymes are
150 clustered with *aglB,* clustering of these genes is uncommon in Crenarchaeota (Kaminski et al., 2013,
151 Nikolayev et al., 2020), which makes the identification of GTs involved in the N-glycosylation process
152 more challenging, as in the Sulfolobales. Interestingly, the gene *saci1262 / agl24* is located only eight
153 genes downstream of the OST *aglB in S. acidocaldarius.* (Fig S1).
154
155
156 **Detailed bioinformatics comparison of Agl24 with Alg13/14 and MurG**
157
158 The overall topology of Agl24 is identical to MurG, which consist of two domain folds separated by a
159 deep cleft corresponding to a GT-B fold structure (Ha et al., 2000, Hu et al., 2003) (Fig 1C). Sequence
160 and domain comparisons of the eukaryal Alg13 and Alg14 to MurG and Agl24 reveal that Alg14
161 corresponds to the N-terminal part of MurG/Agl24, while Agl13 resembles the C-terminal part of
162 MurG/Agl24 (Fig 1A). Both N-terminal (aa1–152) and C-terminal (aa153–327) parts of Agl24 show very
163 low sequence identity to Alg14 and Alg13, with 15% and 16%, respectively. Interestingly, in lower
164 eukaryotes, such as Leishmania and Trypanosoma, Alg14 and Alg13 enzymes are present as a single
165 protein fusion, similar to *S. acidocaldarius* (Averbeck et al., 2007). The question arises whether the
166 ancestral enzyme was fused and split into two domains during evolution. Strikingly, the opposite
167 domain orientation is found *e.g.,* in *Entamoeba histolytica* (Bickel et al., 2005) and *Dictyostelium*
168 *discoideum*. An interesting difference in these proteins from the three domains of life is the presence
169 of an N-terminal transmembrane domain (TMD) in most of eukaryal Alg14 proteins (aa 5–25), which is
170 missing in bacterial MurGs and archaeal Agl24 enzymes. The TMD is predicted to facilitate the
171 association of Agl14 with the ER membrane and to participate in recruitment of Dol-PP-GlcNAc to the
172 membrane (Lu et al., 2012). The introduction of this TMD appears to be derived later in the evolution
173 of eukaryotes, and the interaction with the ER or cytoplasmic membrane is not solely dependent on a
174 TMD, as the presence of hydrophobic patches and protein–protein interactions with membrane
175 proteins are also proposed to facilitate this interaction. Indeed, in lower Eukarya, *e.g.* in *D. discoideum*
176 and *E. histolytica,* having the exchanged domain orientation, lacking an N-terminal TMD (Averbeck et
177 al., 2007).

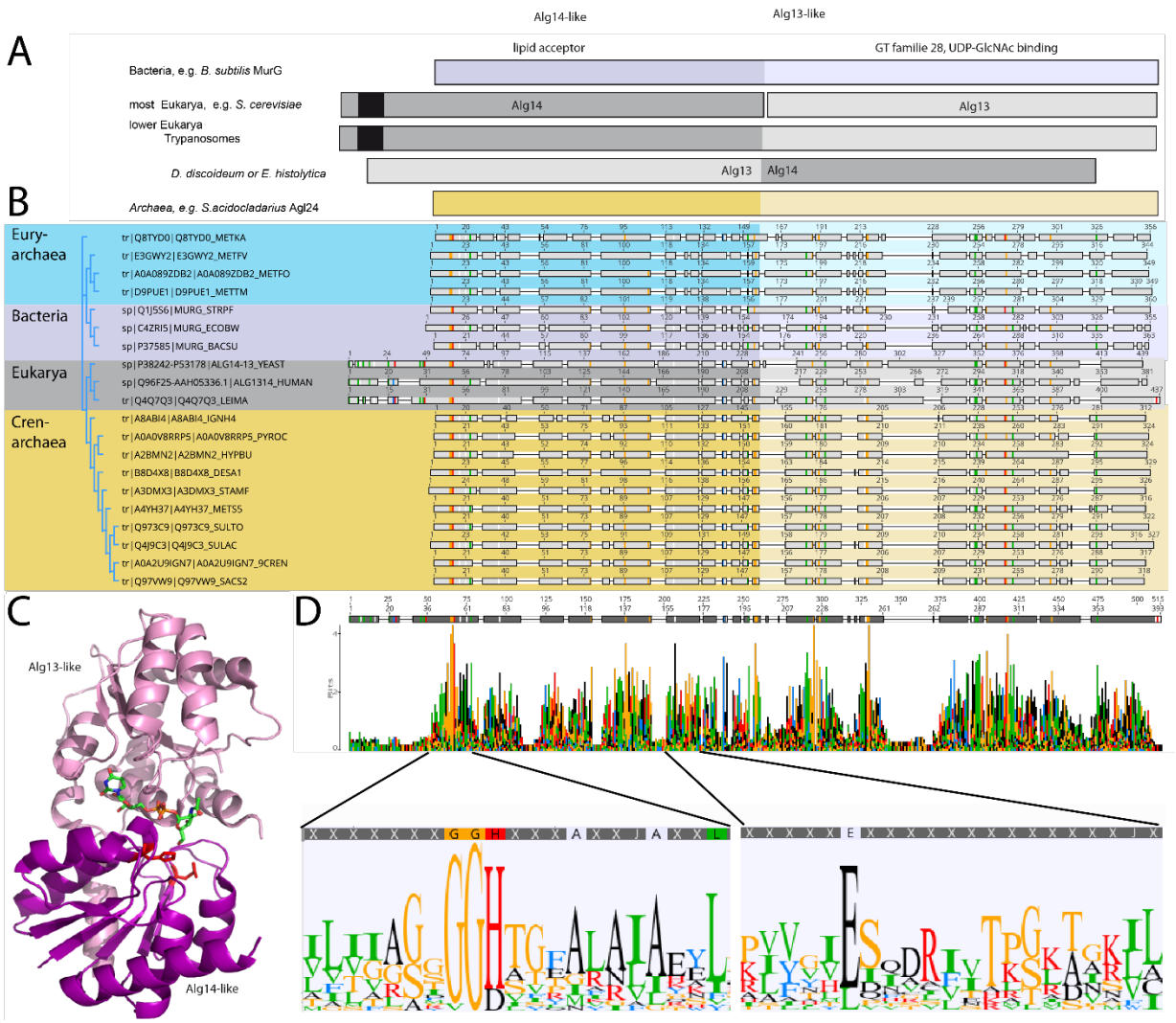
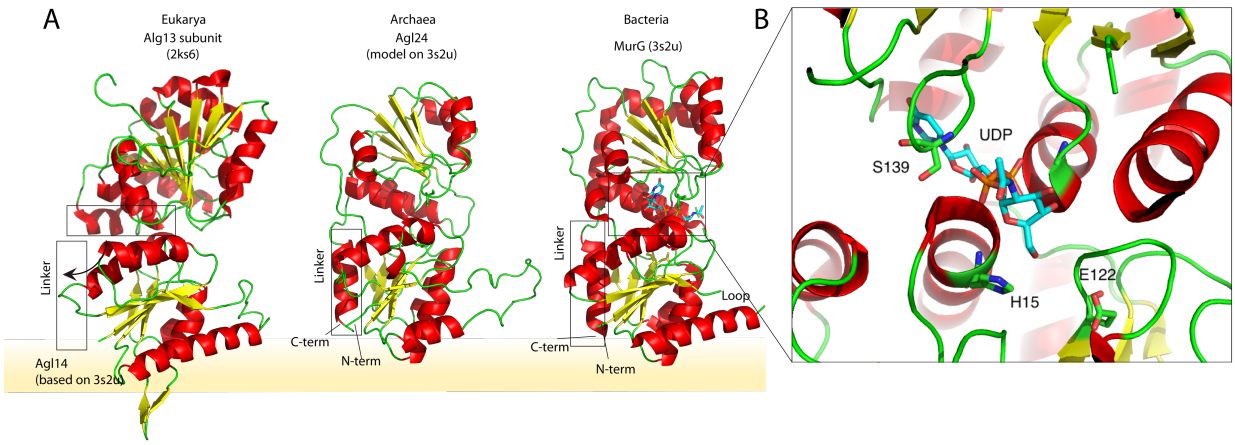


Fig. 1 Comparison of the archaeal Agl24, bacterial MurG, and eukaryal Alg14-13

A) Simplified comparison of the Agl14-like (darker) or Alg13-like (light color) domain in eukaryal Alg14-13, archaeal Agl24, and bacterial MurG ortologs. Transmembrane domain is depicted in black. **B)** Protein sequences were aligned with the Clustal Omega (Sievers et al., 2018) (for the full alignment see Fig S2). Sequences derived from i) three bacterial MurG: *E.coli* (C4ZRI5), *Streptococcus pyogenes* (Q1J5S6), and *Bacillus subtilis* (P37585); ii) eukaryal Alg14-13: *Leishmania major* (Q4Q7Q3), an artificial Alg14-13 fusion from *Saccharomyces cerevisiae* (P38242-P53178) and *Homo sapiens* (Q96F25-Q9NP73-2); iii) the crenarchaeal Agl24 homologs: *Ignicoccus hospitalis (*A8ABI4), *Pyrodictium occultum* (A0A0V8RRP5), *Hyperthermus butylicus (*A2BMN2), *Desulfurococcus amylolyticus* (B8D4X8), *Staphylothermus marinus (*A3DMX3), *Metallosphaera sedula* (A4YH37), *Sulfurisphaera tokodaii* (Q973C9), *Sulfolobus acidocaldarius* (Q4J9C3), *Acidianus brierleyi* (A0A2U9IGN7), and *Saccharolobus solfataricus (*Q97VW9). Selected sequences from pseudomurein producing Euryarchaea with higher sequence similarity to the MurG are aligned: *Methanopyrus kandleri* (Q8TYD0), *Methanothermus fervidus (*E3GWY2), *Methanobacterium formicicum* (A0A089ZDB2), *Methanothermobacter marburgensis* (D9PUE1). Conserved amino acids (65% threshold) are highlighted in color corresponding to their chemical properties of the amino acids: green polar amino acids (G, S, T, Y, C, Q, N), blue basic (K, R, H), red acidic (D, E), and black hydrophobic amino acids (A, V, L, I, P, W, F, M). End of the Agl14-like domain and start of Alg13-like domains are indicated by the change from dark to light background color. **C)** Ribbon presentation of MurG (PDB: 3s2u) bound to UDP-GlcNAc (sticks, green carbon, blue nitrogen, red oxygen) and orientation towards the cell membrane (bottom). Agl13-like domain (turquoise) and Agl14-like domain (violet) are labelled. Conserved His and Glu residues in Agl14 are shown as red sticks. **D)** Strict consensus (65%) and weblogo from the sequence alignment shown in B). Height of each bar correspond to the observed frequency, highly conserved motifs or amino acids are enlarged below, including conserved motifs G(x)GGH15 (Loop I) and E_{114} (full Weblogo see Fig S2).

The amino acid sequence alignment of the representative bacterial MurG, eukaryal Alg14-13, and archaeal homologs revealed conservation of specific patches and individual residues (Fig 1B and 1D). Despite the low sequence identity, a similar structural topology of Agl24 to Agl14/13 and MurG is not surprising, as GTs in general show only a limited number of structural folds. GTs mainly possess a GT-A or GT-B fold, whereas oligosaccharyltransferases and a number of recently characterized GTs display a GT-C, GT-D, or GT-E folds, respectively (Mestrom et al., 2019). Crystal structures of the bacterial

204 MurG (PDB:3s2U, 1F0K, 1NLM) and the C-terminal half of the eukariotic Alg13 (PDB:2jzc) show typical
205 GT-B structural characteristics (Brown et al., 2013, Ha et al., 2000, Hu et al., 2003, Wang et al., 2008,
206 Raman et al., 2010) (Fig 2). The N-terminal part of Alg13 contains a Rossmann-like fold with a mixed
207 parallel and antiparallel β-sheet rather than the conventional Rossmann fold found in all GT-B
208 enzymes, indicating a unique topology among glycosyltransferases(Wang et al., 2008). The N-terminal
209 helix of Agl14 and MurG is proposed to be the membrane association domain (Chantret et al., 2005,
210 Lu et al., 2011). We have generated an Agl24 structural model to investigate potential features via
211 SWISS-MODEL (Waterhouse et al., 2018) (Fig 2). Previous studies have compared the structures of
212 MurG and Agl13, including a model for Alg14 (Gao et al., 2008). The formation of the Agl14/Alg13
213 complex is mediated by a short C-terminal α-helix of Alg13 in cooperation with the last three amino
214 acids of Alg14 (Gao et al., 2008). In the Alg13 crystal structure this C-terminal helix is associated with
215 itself, as Agl14 is missing. The orientation is predicted to be similar to the linker in MurG and the change
216 is indicated by the black arrow (in Fig 2A) (Gao et al., 2008). The structural model of Agl24 is also
217 predicted to contain a C-terminal helix, which interacts with the N-terminal part of the protein (Fig 2A).
218 We revealed a conserved GGxGGH$_{14}$ motif (Agl24) within the N-terminal sequences of MurG and Alg14
219 (Fig 1D, S2, and 2B). This motif is located in the cavity between the two different structural domains
220 next to the substrate binding pocket. The crystal structure of MurG (PDB: 3s2u) in complex with UDP-
221 GlcNAc shows the close proximity of the sugar donor to the conserved residues (Fig 2B) (Brown et al.,
222 2013). In proximity to the GGxGGH$_{14}$ motif is the conserved glutamic acid (E$_{114}$) (Fig 1D, S2, and 2B).
223 Interestingly, both these conserved residues are absent in the archaeal MurG-like GT28 family
224 orthologs of pseudomurein-producing Archaea, where aspartic acid (D) and leucine (L) are more
225 frequently found replacing the H$_{14}$ and E$_{114}$ (Fig 1D and S2). We propose that these residues might be
226 required to accommodate the different acceptor molecule.



227

228 **Fig 2: A) Structural comparison of the eukaryal Agl13 (PDB:2ks6) with a structural model of Agl14, the bacterial MurG (PDB: 3s2u), and the**
229 **structural model of the crenarchaeal Agl24.** Structural models were built with SWISS-MODEL (Waterhouse et al., 2018). Detailed results are
230 listed in Supplementary Table S2. The interactions of the N-terminal helices with the membrane are depicted with a yellow background.
231 **B)** Magnified and 90˚ rotated view into the active site of MurG (PDB: 3s2u) in complex with UDP-GlcNAc (sticks). The catalytic site is located
232 in the cleft of both domains, with conserved His$_{15}$ and Glu $_{(122)}$ residues are highlighted in red.

233 **Agl24 is essential in *S. acidocaldarius***

234 Due to the overall low sequence similarity of Agl24 to Alg14-13 and MurG, we proposed to
235 demonstrate the predicted function of Agl24 *in vivo* by generating a deletion mutant of *Agl24* in the
236 genome of *S. acidocaldarius*. This deletion mutant should arrest the N-glycosylation process after the
237 generation of Dol-PP-GlcNAc and should contain either non-glycosylated glycoproteins or N-
238 glycosylated proteins containing only a single GlcNAc residue linked to the asparagine in the conserved
239 N-glycosylation motif. The genomic integration by homologous recombination of the plasmid
240 pSVA1312 via either the up- or downstream region of the *agl24* and the selection *pyrEF* genes was
241 confirmed by PCR (Fig S3A). However, we were not successful in generating the marker-less in-frame
242 *agl24* deletion mutant. An alternative approach was conducted, aimed to delete the *agl24* gene by a
243 single homologous recombination step by the integration of the *pyrEF* selection cassette. All attempts
244 to generate this gene disruption mutant failed. Therefore, this strongly suggests that the *agl24* gene is
245 essential in *S. acidocaldarius*, at least under the conditions tested. To exclude that the other GTs,
246 identified in the bioinformatic homology search (Table S1), are involved in the N-glycosylation, marker-
247 less deletion mutants of *saci1907*, *saci1921*, *saci0807, saci1094*, *saci1201*, *saci1821*, and *Saci1249*
248 were successfully generated. Only the deletion of *saci0807* showed a significant effect on the N-
249 glycosylation of the S-layer proteins, which have been characterized to encode the GT Agl16 that
250 transfers the terminal glucose residue to the N-glycan (Meyer et al., 2013).
251
252 **Agl24 transfers a single GlcNAc residue onto a GlcNAc acceptor**
253
254 In agreement with the observed membrane association in *S. acidocaldarius* (Fig S4), Agl24 was also
255 found in the membrane fraction during its purification from *E. coli*. Recombinant Agl24, produced in *E.*
256 *coli* (Fig S5), was assayed to test the function of Agl24 using the predicted nucleotide sugar donor UDP-
257 GlcNAc and either one of two synthetic acceptor substrates designed to mimic the native lipid-linked
258 acceptor: $C_{13}H_{27}$-PP-GlcNAc (acceptor 1) or phenyl-O-$C_{11}H_{22}$-PP-GlcNAc (acceptor 2) (Fig 3C). The
259 reaction product was purified and characterized. The MALDI–MS spectra obtained confirmed that
260 Agl24 transfers a single GlcNAc to both acceptor substrates when incubated with UDP-GlcNAc (Fig 5B
261 and D) (Fig 3A). The Agl24 assay revealed that the product peaks are shifted by 203 Da to m/z = 811
262 $[M-1H+2Na]^+$ and m/z = 833 $[M-2H+3Na]^+$ (Fig 3B), corresponding to one dehydrated GlcNAc (203 Da).
263 The same shift was observed for acceptor 2 (Fig 3D). We also investigated Agl24 promiscuity towards
264 utilizing UDP-Glucose as the sugar donor, but no mass shift was observed (Fig 3E), indicating that Agl24
265 uses exclusively UDP-GlcNAc.

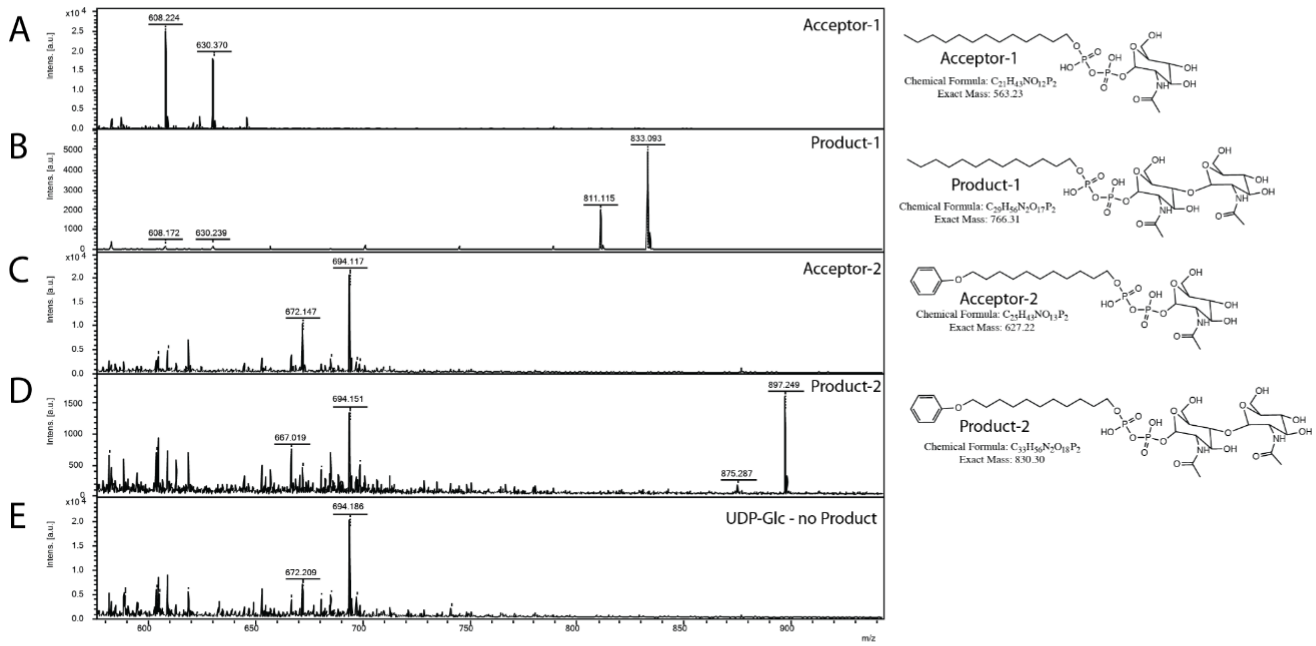


Fig 3: MALDI-MS spectra of the *in vitro* Agl24 reaction assessing the N-acteylglucosamine transferase activity. Spectra obtained from the enzymatic reactions containing Agl24-GFP and (**A**) acceptor-1, **(B)** acceptor-1 and UDP-GlcNAc, and Agl24-GFP, **(C)** acceptor-2, **(D)** acceptor-2, and UDP-GlcNAc, **(E)** acceptor-2 and UDP-Glc. The conversion from acceptor 1 (608 m/z $[M-1H+2Na]^+$ and 630 m/z $[M-2H+3Na]^+$) or acceptor 2 (672 m/z $[M-1H+2Na]^+$ and 694 m/z $[M-2H+3Na]^+$) to the product (811 m/z $[M-1H+2Na]^+$ and 833 m/z $[M-2H+3Na]^+$) or (875 m/z $[M-1H+2Na]^+$ and 897 m/z $[M-2H+3Na]^+$) was observed only when UDP-GlcNAc was used as nucleotide sugar donor.

Activity of Agl24

Since *S. acidocaldarius* is a thermophilic microorganism, with an optimal growth temperature of 75°C, the temperature dependency of the Agl24 activity was investigated using our established mass spectrometry assay. The substrates are stable at the conditions tested and MALDI analysis of the negative control lacking the enzyme revealed only the acceptor-1 mass (Fig 4A). At elevated temperatures, the peak intensity from the product increased while the peak intensity of the acceptor molecule was reduced (Fig 4B-F). Highest activity was detected at 70°C, close to the optimal growth temperature of *S. acidocaldarius*. Furthermore, the addition of EDTA did not affect the activity of Agl24, demonstrating that Agl24 is a metal independent GT (Fig S7, 5D). This result is in agreement with the lack of a conserved Asp-X-Asp (DxD) motif within Agl24 sequence, which has been shown to be important to coordinate the metal ions in A-fold GTs, whereas GT-B GTs are metal ion-independent(Lairson et al., 2008, Gloster, 2014).

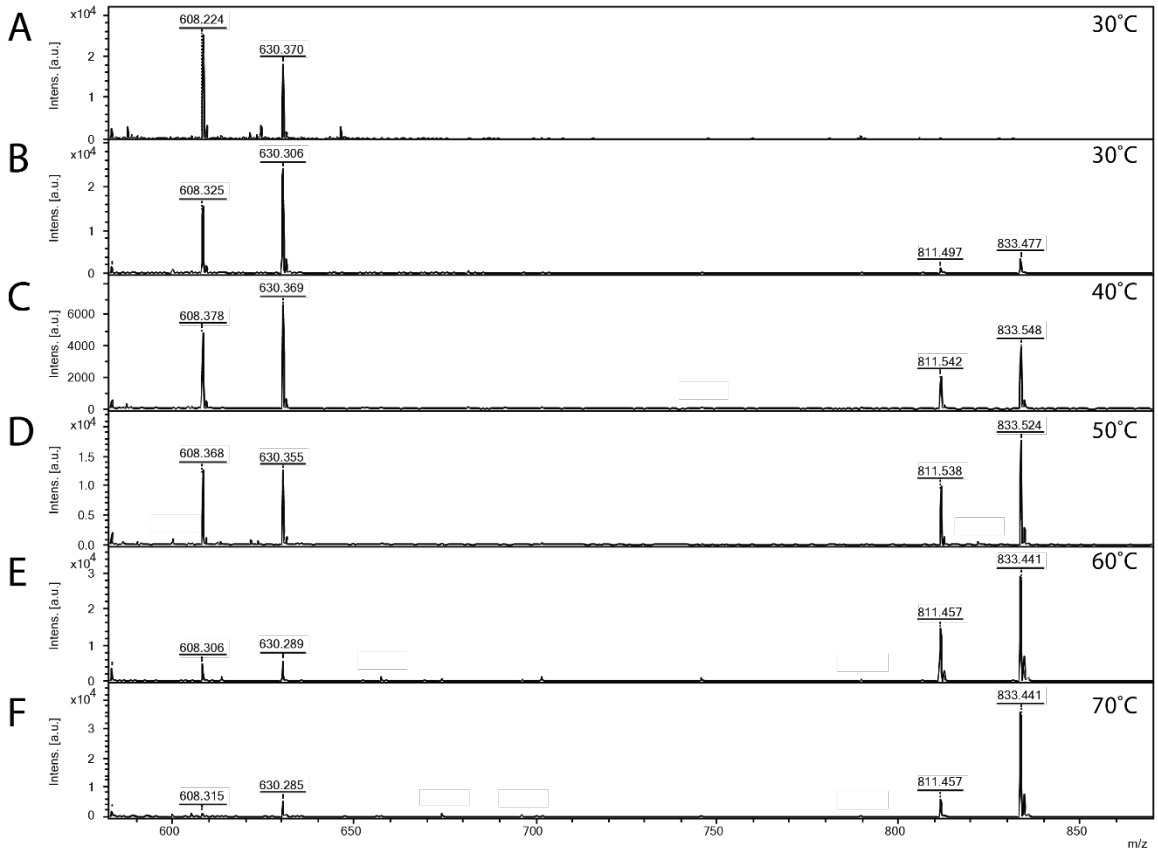


Fig 4: MALDI-MS spectra of the *in vitro* Agl24 reaction at different temperatures. Spectra obtained from the purified enzymatic reaction mix with only the acceptor-1 (**A**), acceptor-1, UDP-GlcNAc, and Agl24-GFP WT at 30°C (**B**), 40°C (**C**), 50°C (D), 60°C (**E**), and 70°C (F). Conversion of the acceptor-1 (608 m/z [M-1H+2Na] and 630 m/z [M-2H+3Na]) towards the product (811 m/z [M-1H+2Na] and 833 m/z [M-2H+3Na]) is significantly dependent on the applied temperature. Above 50°C the depletion of the acceptor towards the gain of product is clearly visible.

Conserved His14 is essential for Agl24 function

Two conserved amino acid residues were targeted by mutagenesis to investigate their role in the Agl24 enzyme. Alanine substitution of histidine residue H_{14} within the conserved GGH_{14} motif, found in all Alg14 (GSGGH), MurG (GGxGGH), and Agl24 (GGH) homologs, was inactive (Fig 5B). This demonstrated that this highly conserved residue, located next to the nucleotide-binding site, is important for the enzyme function. This GGxGGH motif, and a subsequent second glycine rich motif (GGY), have been proposed to enable MurG to be involved in interaction with the diphosphate group of the lipid acceptor, as these two motifs resemble phosphate-binding loops of nucleotide-binding proteins (Baker et al., 1992, Carugo and Argos, 1997). Alanine substitution of the conserved His residue in MurG resulted in undetectable activity and loss of ability to complement a temperature sensitive MurG mutant (Crouvoisier et al., 2007, Hu et al., 2003). In contrast, the substitution of the second conserved residue E_{114}, opposite the nucleotide-binding site, had no effect on the activity of Agl24 (Fig 5C). In MurG this glutamic acid residue is found in a conserved $H_{124}EQN_{127}$ motif (*E. coli*), proposed to coordinate the lipid acceptor (Crouvoisier et al., 2007, Hu et al., 2003). The MurG $E_{125}A$ variant was able to complement the thermosensitive MurG *E.coli* strain, but the remaining activity was 860-fold lower than that of the wild-type MurG (Crouvoisier et al., 2007).

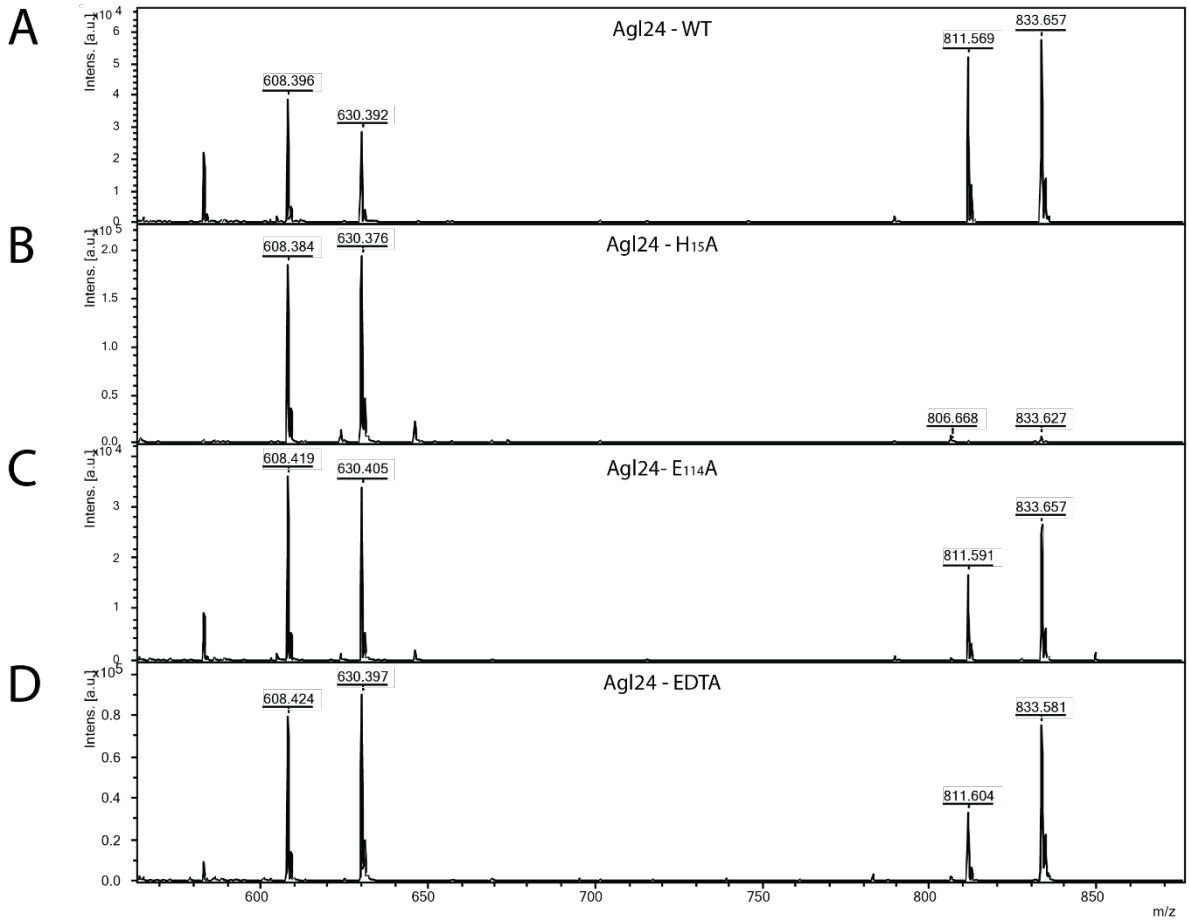


Fig 5: MALDI-MS spectra of the *in vitro* Agl24 reaction, the generated point mutations H15A and E114A and EDTA control. Spectra obtained from the purified enzymatic reaction mix with acceptor-1(608 m/z [M-1H+2Na] and 630 m/z [M-2H+3Na]), UDP-GlcNAc and (**A**) Agl24-GFP WT, (**B**) Agl24-H_{15}A-GFP, (**C**) Agl24-E_{114}A-GFP, and (**D**) Agl24-GFP WT with addition of 10 mM EDTA. The activity was significantly reduced in the H_{15}A mutants, while a similar amount of product (811 m/z [M-1H+2Na] and 833 m/z [M-2H+3Na]) was obtained in the E_{114}A mutant.

Agl24 is an inverting β-1,4-N-acetylglucosamine-transferase

For kinetic analysis of Agl24, a HPLC assay was used to monitor conversion of the mono-GlcNAcylated-PP-lipid acceptor-2 substrate to the bi-GlcNAcylated product (Fig S6). A K_m value for acceptor-2 could not be determined using this system due to significant substrate inhibition above 0.5 mM, however, a $K_m^{(app)}$ for UDP-GlcNAc was determined: ${}^{(UDP\text{-}GlcNAc)}K_m$ = 1.37 ± 0.13 mM, V_{max} = 32.5 ± 0.9 pmol min^{-1}. These results are in agreement with the K_m values of other GT enzymes, which typically exhibit K_m affinities for their respective nucleotide sugars in the high micromolar range, reflecting the estimated intracellular concentrations of the nucleotide sugar (Varki et al., 1999). A ^{1}H NMR analysis confirmed that the enzymatic product contained two GlcNAc residues with different linkage types based on the presence of two differently-coupled anomeric protons (Fig 6 and S8). One anomeric proton appeared as a doublet of doublets (5.35 ppm, $J_{H1\text{-}H2}$ = 3.1 Hz, $J_{H1\text{-}P}$ = 7.2 Hz) typical of an α-linked GlcNAc residue (Table 1). Another anomeric signal appeared as a doublet with a large *J* coupling value (4.49 ppm, $J_{H1\text{-}H2}$ = 8.5 Hz), indicative of a β-linked GlcNAc residue. The substrate acceptor 2 contains an α-linked GlcNAc; this strongly suggested the terminal GlcNAc introduced by Agl24 was β-linked. Experiments

using 1D total correlation spectroscopy (TOCSY) deciphered the detailed proton signals from each of the two sugar rings (Fig S8), and 2D COSY (Fig S9). In addition, HSQC experiments (Fig S10) were used to assign the identity of each proton and carbon signal. The C4 signal of the α-GlcNAc residue is shifted to 79.6 ppm, strongly suggesting that the terminal β-GlcNAc was linked at this position. This was further supported by a relative increase in shift of the H4 proton of the α-GlcNAc residue compared to the un-modified acceptor (Zorzoli et al., 2019). Relative shifts of other protons reported for the acceptor aligned with our experimental data. In conclusion, a combination of ^{1}H NMR and 1D and 2D TOCSY, COSY and HSQC experiments confirmed that Agl24 is an inverting β-1,4-GlcNAc transferase.

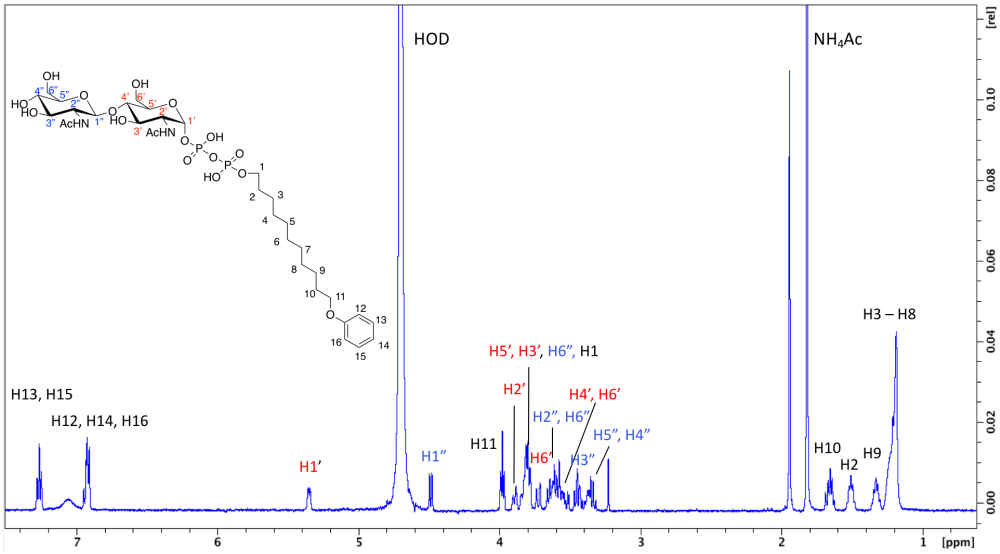


Fig 6 – ^{1}H NMR spectra of purified Agl24 enzymatic product. Spectra were acquired in D_2O at 293K on a Bruker AVANCE III HD 500 MHz NMR Spectrometer fitted with a 5-mm QCPI cryoprobe. Chemical shifts are reported with respect to the residual HDO signal at δ_H 4.70 ppm.

Table 1 – ^{1}H and ^{13}C chemical shifts of the *in vitro* Agl24 reaction product.

Lipid		**1**	**2**	**3 to 8**	**9**	**10**	**11**	**12, 14**	**13, 15**	**16**
	^{1}H NMR	3.79	1.51	1.16-1.26	1.33	1.66	3.98	6.92	7.26	6.92
	^{13}C NMR	66.8	30.3	28.8	28.7	28.7	68.8	115.6	130.1	121.8
α-GlcNAc		**1'**	**2'**	**3'**	**4'**	**5'**	**6'a**	**6'b**		
	^{1}H NMR	5.35	3.896	3.8	3.58	3.83	3.73	3.57		
	^{13}C NMR	94.25	53.65	70	79.65	71.69	60.15			
β-GlcNAc		**1"**	**2"**	**3"**	**4"**	**5"**	**6"a**	**6"b**		
	^{1}H NMR	4.49	3.64	3.45	3.34	3.38	3.8	3.6		
	^{13}C NMR	101.55	56.14	74.01	70.38	76.27	61.15			

The eukaryotic GTs Alg14 and Alg13 are closely related to Asgard homologs

In order to analyze the phylogenetic relationship between the eukaryotic N-glycosylation GTs Alg13 and Alg14 with archaeal and bacterial homologs, an extensive phylogenetic analysis was performed.

345 The eukaryotic Alg13 and Alg14 sequences cluster with homologs from the Asgard phyla
346 Thorarchaeota and Odinarchaeota, as well as with sequences from Verstraetearchaeota,
347 Micrarchaeota, Geothermarchaeota, and an unclassified archaeon (Fig7 A and B). Even though the
348 sister clades of Eykaryotes are Thorarchaeota and Odinarchaeota for Alg13 and Alg14 respectively, the
349 only the latter relationship is strongly supported. No closely-related Alg13- and Alg14-like sequences
350 were found in other Asgard phyla, suggesting that the remaining Asgard, such as Lokiarchaeota and
351 Heimdallarchaeota, could be using a different enzyme for synthesizing N-glycans.
352 Next to the Alg13- or Alg14-like cluster, there is a second distinct cluster of EpsF- or EpsE-like
353 sequences, respectively. The function of EpsF and EpsE has been studied in Lactobacillales, where
354 these enzymes are involved in exopolysaccharide production (De Vuyst et al., 2001, Kolkman et al.,
355 1997). Here, similar to Alg14/13, the combined activity of EpsE and EpsF links either a glucose (Glc)
356 from UDP-Glc (van Kranenburg et al., 1999, Kleerebezem et al., 1999), or a galactose (Gal) from UDP-
357 Gal via a β-1,4 linkage to lipid-linked glucose acceptors (van Kranenburg et al., 1999), creating either a
358 lipid-linked cellobiose or a lactose. The archaeal Eps-like sequences are mainly found in diverse
359 Euryarchaeota, whereas the bacterial sequences are scattered among different phyla, possibly as a
360 result of extensive horizontal gene transfers.
361 The remaining clades in the Alg13 and Alg14 phylogenies consist of sequences from the archaeal TACK
362 superphylum and some Euryarchaeota, mainly methanogens (Fig 7A and B). The Agl24 identified in this
363 study is found among them, within a Cren- and Bathyarchaeota branch. Close homologs of the Agl24
364 are found in all presently available genomes of the Crenarchaeota, in the orders Fervidicoccales,
365 Acidilobales, Desulfurococcales and Sulfolobales, with the exception of any homolog in the members
366 of the order Thermoproteales (Fig S11, Sup Data 1).
367 The remaining phyla of the TACK superphylum (Aig-, Bathy-, Geotherm-, Kor-, Thaum-, Nezha-, Mars-,
368 Brock-, and Verstraetearchaeota) along with some additional Crenarchaeota, including Geoarchaeota,
369 and an Odinarchaeota sequence (Fig 7A and B), form a separate branch containing Agl24-like
370 sequences. Next to this branch is a clade of methanogens and Altiarchaeales. Due to the large number
371 of very distant Alg13/14 homologs with low similarity, resulting in poor alignments, for the phylogenies
372 in Fig 7 we sought to imitate and expand the clades presented previously (Lombard, 2016).
373 Nevertheless, in the homology searches used to expand the previous set (Lombard, 2016), some of the
374 hits formed well-supported branches in preliminary phylogenies (Sup Data 1). By including some of
375 them in expanded Alg13/EpsF and Alg14/EpsE datasets, none of them branch within the Alg13/EpsF-
376 like or Alg14/EpsE-like clades. Instead, they are fused Agl24-like sequences, mainly from Euryarchaeota
377 and various DPANN phyla. Either way, the Cren-/Bathyarchaeota Agl24 clade along with a small
378 Nanohaloarchaeota clade appear to be the closest relatives of the Alg13/EpsF-like and Alg14/EpsE-like
379 proteins (Sup Data 1).

The fact that we are able to trace Agl24-like homologs to the ancestor of TACK and several Euryarchaeota lineages indicates that, despite multiple instances of lateral gene transfer, they are very ancient and diverged from the bacterial MurG at the separation of the two domains, as proposed by Lombard *et al.* (Lombard, 2016). Given that both MurG (Lombard, 2016) and the various Agl24-like homologs are fused genes, a major split event seems to have occurred once at the base of the Alg/Eps-like clade (Fig 8A and B, red arrow), with some independent, more recent, fusions found throughout eukaryotes and Archaea (*e.g.* the Thorarchaeota RLI55694.1). Finally, recovering the monophyly of eukaryotes in our phylogenies corroborates the proposition by Lombard *et al.* (Lombard, 2016) that Alg13 and Alg14 were present in the last eukaryotic common ancestor (LECA). Since the sister clade of eukaryotes is in both cases an Asgard phylum (although not the Heimdallarchaeota), it is possible that these genes were inherited vertically during eukaryogenesis in a two-domain tree of life (Eme et al., 2017).

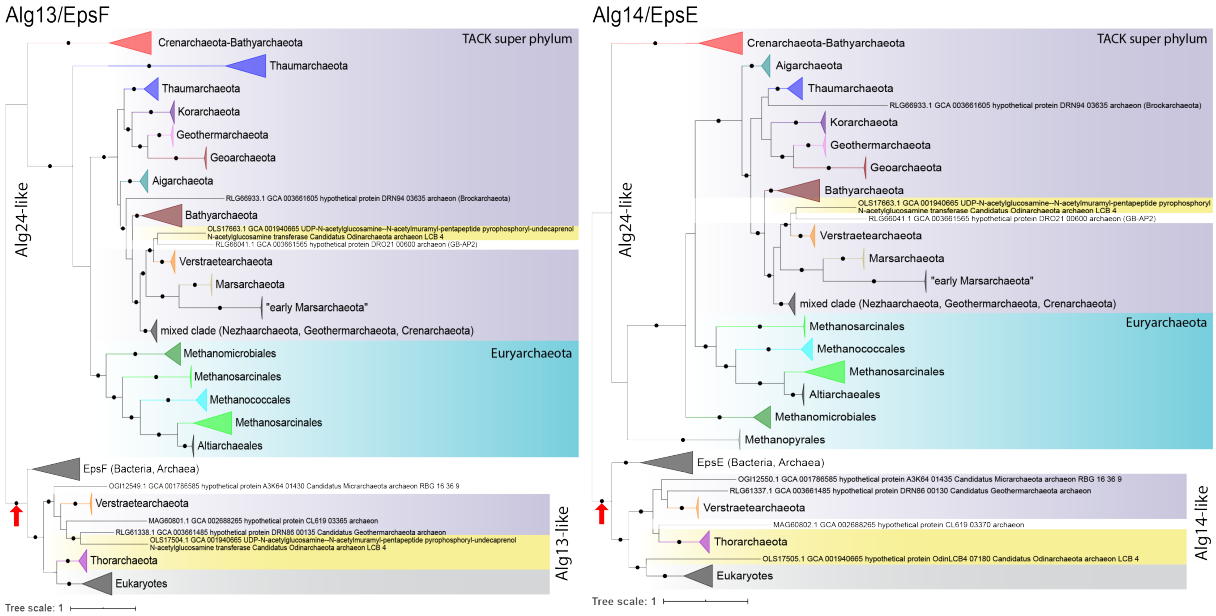


Fig 7: Single gene phylogenies of Alg13/EpsF and Alg14/EpsE. Dots indicate strongly supported branches (ultrafast bootstrap >=95 and aLRT SH-like support >=80). Higher-level classifications are highlighted in specific background colors: purple: archaeal TACK superphylum, turquoise: Euryarchaeota, yellow: Asagradarchaeota, grey: Eukaryotes. Agl24 of *S. acidocaldarius* is found within the red colored clade at the top of the phylogenetic tree. Red arrow: indicates major split event to form the Alg13/EpsF-like and Alg14/EpsE-like enzymes.

Discussion

In this study we identified the first archaeal and thermostable β-1,4-N-acetylglucosaminyltransferase, named archaeal glycosylation enzyme 24 (Agl24), responsible for the synthesis of the lipid-linked chitobiose on which the N-glycan of *S. acidocaldarius* is assembled. The reported essentiality of Agl24 is in agreement with the indispensable properties of the OST AglB and AglH, catalyzing either the last or first step of the N-glycosylation process in *S. acidocaldarius* (Meyer and Albers, 2014, Meyer et al.,

405 2017). The essentiality of the N-glycosylation has also been confirmed by a transposon study in
406 *Sulfolobus islandicus*, a close relative to *S. acidocaldarius*, in which all homologous genes, including
407 *agl24*, lack any transposon insertions (Zhang et al., 2018). However, the fundamental nature of
408 archaeal N-glycosylation cannot be generalized since deletion mutants of the
409 oligosaccharyltransferase AglB can be obtained in Euryarchaeota, *e.g.,* in *Methanococcus maripaludis*
410 and *Haloferax volcanii* (Abu-Qarn et al., 2007, Chaban et al., 2006). The requirement for N-
411 glycosylation might rely on many factors including the N-glycan composition and structure as well as
412 the modification frequency on proteins; this frequency is reported to be higher in thermophilic,
413 compared to mesophilic, Archaea (Meyer and Albers, 2013, Jarrell et al., 2014).
414 The comparison of Agl24 with eukaryotic Alg14/13 and the bacterial MurG enzymes revealed structural
415 and amino acid similarities. Although the sequence identity of Agl24 with the eukaryotic and bacterial
416 enzymes is extremely low (~16%), the structural prediction of Agl24 indicated a MurG–like GT-B fold.
417 Interestingly, the amino acids of the nucleotide sugar donor binding site of MurG (Hu et al., 2003) are
418 not conserved in Agl24 and Alg14, although all enzymes use UDP-GlcNAc as the sugar donor (Fig S2,
419 boxed). Strikingly, a conserved GGH_{14} motif within an N-terminal loop of MurG, Alg14-13, and Agl24
420 (Fig 1 and S2), which protrudes into the cavity formed by the two structural folds, could be identified.
421 In Bacteria this $G_{13}GTGGH_{18}$ loop is extended by a second $G_{103}GY_{105}$ loop. As both loops are suggested
422 to be reminiscent of the phosphate-binding loops of nucleotide-binding proteins, their involvement in
423 the binding of the diphosphate group of the acceptor lipid has been proposed (Crouvoisier et al., 2007).
424 In agreement with our results, the alanine substitutions of the invariant bacterial H_{14} led to an
425 extremely low or undetectable activity of MurG as well as the loss of ability to complement a
426 temperature sensitive MurG mutant (Crouvoisier et al., 2007, Hu et al., 2003). However, up to now, no
427 structure of MurG or Agl14-13 with the bound lipid acceptor molecule has been published and the
428 interaction of the imidazole ring of the histidine residue with the acceptor molecule remained to be
429 shown. Our alignment also reveals a conserved glutamic acid in the bacterial, eukaryotic, and archaeal
430 sequences. In bacteria, this glutamic acid is found in a conserved $H_{124}EQN_{127}$ loop (*E. coli* MurG), which
431 has been proposed to form a key part of the acceptor-binding site (Hu et al., 2003). In agreement with
432 our study, an $E_{125}A$ substitution significantly affected the function of MurG, although a marked
433 decrease in lipid acceptor binding (1760-fold) and nucleotide substrate binding (210-fold) was
434 observed (Hu et al., 2003, Crouvoisier et al., 2007).
435 Although lacking a TMD, Agl24 associates with the membrane, indicating interaction with membrane
436 enzymes, similar to the interaction of MurG or Alg13/14 with their membrane interaction partners
437 MraY and Alg7 respectively. Based on the structural similarities of these complexes, a shared common
438 evolutionary origin of the first enzymes of eukaryotic N-glycosylation and the bacterial peptidoglycan
439 biosynthesis has been hypothesized (Bugg and Brandish, 1994, Burda and Aebi, 1999, Bouhss et al.,

440 2008). As *S. acidocaldarius* possesses a homolog of Alg7, termed AlgH (Meyer et al., 2017), and a
441 bioinformatics study has indicated that Alg7 might have emerged out of the Crenarchaeota (Lombard,
442 2016), AglH presents the most likely candidate for the interaction partner of Agl24.
443 A phylogenetic analysis shed light onto the distribution of the identified Agl24 in Archaea, and
444 determined its relationship with eukaryotic and bacterial homologs. This revealed a very ancient and
445 broad distribution of Agl24-like sequences in Archaea. Importantly, close homologs of Agl24 are only
446 found within the Crenarchaeota and Bathyarchaeota. Indeed, all Crenarchaeota, with the exception of
447 the order Thermoproteales, possess an Agl24 homolog, strongly suggesting that these Crenarchaeota
448 contain a chitobiose N-glycan core structure. In contrast, the N-glycans of members of the
449 Thermoproteales would be assembled using a different enzyme. Indeed, the N-glycan of *Pyrobaculum*
450 *caldifontis*, belonging to the order Thermoproteales, differs from the N-glycan in Sulfolobales (Fujinami
451 et al., 2017). Here, the N-glycan core is composed of two di-N-acetylated β-1,4-linked GlcNAc residues
452 with the second sugar carboxylated at C6 ($GlcA(NAc)_2$-β-1,4-$(Glc(NAc)_2$. A di-N-acetylated glucuronic
453 acid sugar is also found in the N-glycan from *Methanococcus voltae,* which is transferred by the GT
454 AglC onto the Dol-P linked GlcNAc (Chaban et al., 2009). Thus, a homolog of AglC is likely to be involved
455 in the N-glycan biosynthesis in *P. caldifontis*.
456 Our phylogenetic analysis revealed that the closest related enzymes of the eukaryotic Alg14 and Alg13
457 are found within the Asgard superphylum. Since the discovery of the Archaea as the third domain of
458 life next to Bacteria and Eukarya (Woese and Fox, 1977), the comparison of archaeal molecular
459 processes has gradually revealed a strong resemblance to those found in eukaryotes (Zillig et al., 1989,
460 Huet et al., 1983). Furthermore, recent phylogenetic analyses indicated that eukaryotes originated
461 close to the TACK superphylum (Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota)
462 within the Asgard superphylum (Spang et al., 2015, Zaremba-Niedzwiedzka et al., 2017, Petitjean et
463 al., 2015, Koonin, 2015). The current hypothesis for the formation of the first eukaryotic cell proposes
464 a membrane protrusion of an ancestral archaeon that engulfed a bacterium to increase the
465 interspecies interaction surface (Baum and Baum, 2014, Imachi et al., 2020). The eukaryotic N-
466 glycosylation would therefore be inherited from this ancient archaeal ancestor, which is in agreement
467 with the high conservation of the initial N-glycosylation biosynthesis steps in all eukaryotes (Helenius
468 and Aebi, 2004, Samuelson et al., 2005, Schwarz and Aebi, 2011). In addition, the AglB in members of
469 the Asgard and TACK superphyla shows considerable similarity to its eukaryotic homologue Stt3, and
470 is clearly distinct from the AglB of the DPANN superphylum and the Euryarchaeota (Nikolayev et al.,
471 2020). Most eukaryotic Stt3 contain a double sequon motif for N-glycosylation near the substrate-
472 binding site, which is hypothesized to contribute to the LLO (Shrimal and Gilmore, 2019, Wild et al.,
473 2018). Interestingly, this double sequon is also present in the AglB of the members of the Asgard and
474 TACK superphyla but absent in the AglB of Euryarchaeota (Shrimal and Gilmore, 2019). In addition,

homologs of non-catalytic subunits of the eukaryotic Stt3 complex, *e.g.,* ribophorin, Ost3/6, Ost5, and Wbp1, have been detected in some Arsgardarchaeota (Zaremba-Niedzwiedzka et al., 2017). However, their specific function in the N-glycosylation has to be shown, especially as ribophorin is only found in higher mammals (Wilson and High, 2007) and therefore unlikely to exist in an early eukaryotes. Nevertheless, the assembly of the N-glycan linkage in the TACK superphylum is conducted via a pyrophosphate linkage to the Dol, identical to that in Eukarya. This is in contrast to Euryarchaeota, which use a mono-phosphate linkage. A summary of the similarities and differences of the archaeal N-glycosylation to Eukarya and Bacteria are given in Table S4. All these observations strengthen the hypothesis that the eukaryotic N-glycosylation has emerged from an ancient archaeon.

In the future, we believe that the detailed characterization of the N-glycosylation process in members of the TACK and Asgard superphyla will lead to the elucidation of further molecular similarities and unique properties to the eukaryotic N-glycosylation process and will provide further support for the origin of the eukaryotic N-glycosylation.

489 **Materials and methods**

490 **Strains and growth conditions**

491 The strain *Sulfolobus acidocaldarius* MW001 (Δ*pyrE*) (Wagner et al., 2012) and all derived modified
492 strains (Table S3) were grown in Brock medium at 75°C, pH 3, adjusted using sulfuric acid. The medium
493 was supplemented with 0.1% w/v NZ-amine and 0.1% w/v dextrin as carbon and energy source (Brock
494 et al., 1972). Selection gelrite (0.6%) plates were supplemented with the same nutrients with the
495 addition of 10 mM $MgCl_2$ and 3 mM $CaCl_2$. For second selection plates 10 mg ml^{-1} uracil and 100 mg
496 ml^{-1} 5-fluoroorotic acid (5-FOA) were added. For the growth of the uracil auxotrophic mutants, 10 mg
497 ml^{-1} uracil was added to the medium. Cell growth was monitored by measuring the optical density at
498 600 nm. Protein expression in *S. acidocaldarius* was conducted in medium supplemented with 0.1%
499 w/v NZ-amine and 0.1% w/v L-arabinose as carbon and energy source. All *E. coli* strains DH5α, BL21,
500 or ER1821 were grown in LB media at 37°C in a shaking incubator at 200 rpm. According to the
501 antibiotic resistance in the transformed vector(s), media were supplemented with the antibiotics
502 carbenicillin (amp) at 100 µg ml^{-1} and/or kanamycin (kan) at 50 µg ml^{-1}.
503

504 **Construction of deletion plasmids**

505 The predicted function of *Agl24* was verified via the generation of the lipid-linked chitobiose core of
506 the *N*-glycan. Therefore, a marker-less deletion mutant of this gene was constructed in the *S.*
507 *acidocaldarius* MW001, as previously described (Wagner et al., 2012). Briefly, the strain MW001,
508 auxotrophic for uracil biosynthesis, was transformed with the plasmid pSVA1312. Therefore, two
509 ~1000 bp DNA fragments, one from the upstream and one from the downstream regions of *Agl24*
510 (*saci1262*) gene, were PCR amplified. Restriction sites *Apa*I and *Bam*HI were introduced at the 5´ ends
511 of the upstream forward primer (4168) and of the downstream reverse primer (4165), respectively.
512 The upstream reverse primer (4163) and the downstream forward primer (4164) were each designed
513 to incorporate 15 bp of the reverse complement strand of the other primer, resulting in a 30 bp overlap
514 stretch. The overlapping PCR fragments were purified, digested with *Apa*I and *Bam*HI, and ligated in
515 the digested plasmid pSVA407, containing *pyrEF* (Wagner et al., 2012).

516 **Generation of a linear *Agl24*$_{up}$-*pyrEF*-*Agl24*$_{down}$ fragment for the direct *Agl24*::*pyr*EF replacement**

517 To further underline the essential role of *Agl24* in *S. acidocaldarius*, a disruption of the *Agl24* gene by
518 direct homologous integration of the *pryEF* cassette was performed. For this approach, 387 bp of the
519 *Agl24* upstream region, the full 1525 bp of the *pyrEF* cassette, and 1011 bp of the *Agl24* downstream
520 region were PCR amplified using the primers: 4168+6336, 4115+4116, and 6337+6338, respectively.
521 At the 5´ ends of the upstream forward primer (4168) and of the downstream reverse primer (6338)

522 restriction sites *Apa*I and *Bam*HI were introduced, respectively. The *pyrEF* forward primer (4115) was
523 designed to incorporate 40 bp of the upstream reverse primer (6336) resulting in a 40 bp overlapping
524 stretch. The *pyrEF* reverse primer (4116) and the downstream forward primer (6337) weres designed
525 to incorporate the reverse complement strand of the other primer, resulting in a 46 bp overlapping
526 stretch. The upstream, *pyrEF*, and downstream fragments were fused by an overlapping PCR, using the
527 3' ends of each fragment as primers. The 2904 bp overlap PCR fragment gained, was amplified using
528 the outer primers (4168 and 6338), digested with *Apa*I and BamHI and ligated into the p407 vector
529 predigested with the same restriction enzymes. The resulting plasmid, pSVA3338 (*Agl24*$_{up}$-*pyrEF*-
530 *Agl24*$_{down}$), was transformed into *E. coli* DH5α and selected on LB-plates containing 50 mg ml^{-1}
531 ampicillin. The accuracy of the plasmid was verified by sequencing. Before transformation in
532 *S. acidocaldarius,* the plasmid was digested with *Apa*I and *Bam*HI to create the linear *Agl24*$_{up}$-*pyrEF*-
533 *Agl24*$_{down}$ fragment.

534 **Generation of Agl24 Expression plasmids for the production of Agl24 in *S. acidocaldarius* and *E. coli***

535 Cellular localization studies were carried out using N- and C-terminal fusion proteins of Agl24. The
536 *Agl24* gene was amplified from genomic DNA of *S. acidocaldarius* introducing *NcoI* and *Pst*I restriction
537 sites at the 5'ends of the primer pair 4176 and 4177, respectively. The PCR fragment was cloned into
538 pSVA1481, yielding vector pSVA1336 containing an inducible arabinose promoter and *Agl24*-Strep-
539 His$_{10}$. The plasmid was digested with *Nco*I and *Eag*I and the insert *Agl24*-Strep-His$_{10}$ was ligated into
540 pSVA1481 containing an inducible maltose promoter, yielding pSVA1337. The N-terminal Strep-His$_{10}$-
541 TEV-Agl24 was cloned using the primer pair 4180 and 4181, which incorporated the restriction sites
542 *NcoI* and *NotI*, respectively. The Insert was ligated into pSVA2301 containing an inducible maltose
543 promoter yielding pSVA1339.
544 For heterologous expression of Agl24 in *E. coli,* the plasmid pHD0499 was generated. The GFP-His$_8$-
545 tagged *Agl24* gene was constructed by amplifying the full-length *Agl24* gene from *S. acidocaldarius*
546 genomic DNA with the primer pair A596 and A597. The PCR fragment was cloned in-frame into the
547 vector pWaldo (Waldo et al., 1999) using *Xho*I and *Kpn*I restriction sites, generating a C-terminal fusion
548 with a TEV cleavage site, GFP, and His$_8$ tag. For the generation of Agl24 mutants of the conserved
549 mutants H$_{15}$ and E$_{114}$, a quick-change mutagenesis PCR was applied using the primers: A597 and A684
550 or A685 and A686, respectively.
551

552 **Transformation and selection of the deletion mutant in *S. acidocaldarius***

553 Generation of competent cells was performed based on the protocol of Kurosawa and Grogan
554 (Kurosawa and Grogan, 2005). Briefly, *S. acidocaldarius* strain MW001 was grown to an OD$_{600}$ between

555 0.1 and 0.3 in Brock medium supplemented with 0.1% w/v NZ-amine and 0.1% dextrin. Cooled cells
556 were harvested by centrifugation (2000 x g at 4°C for 20 min). The cell pellet was washed three times
557 successively in 50 ml, 10 ml and 1 ml of ice-cold 20 mM sucrose (dissolved in demineralized water)
558 after mild centrifugation steps (2000 x g at 4°C for 20 min). The final cell pellet was resuspended in 20
559 mM sucrose at an OD_{600} of 10.0 and stored in 50 µl aliquots at -80°C. 400–600 ng of methylated
560 pSVA1312 or the linearized *Agl24*$_{up}$-*pryEF-Agl24*$_{down}$ fragment, was added to a 50 µl aliquot of
561 competent MW001 cells and incubated for 5 min on ice before transformation in a 1 mm gap
562 electroporation cuvette at 1250 V, 1000 Ω, 25 mF using a Bio-Rad gene pulser II (Bio-Rad, X, USA).
563 Directly after transformation 50 µl of a 2x concentrated recovery solution (1% sucrose, 20 mM b-
564 alanine, 20 mM malate buffer pH 4.5, 10 mM $MgSO_4$) was added to the sample and incubated at 75°C
565 for 30 min under mild shaking conditions (150 rpm). Before plating, the sample was mixed with 100 µl
566 of heated 2x concentrated recovery solution and twice 100 µl were spread onto gelrite plates
567 containing Brock medium supplemented with 0.1% NZ-amine and 0.1% dextrin. After incubation for
568 5–7 days at 75°C, large brownish colonies were used to inoculate 50 ml of Brock medium containing
569 0.1% NZ-amine and 0.1% dextrin, which were incubated for 3 days at 78°C. Cultures confirmed by PCR
570 to contain the genomically integrated plasmid were grown in Brock medium supplemented with 0.1%
571 NZ-amine and 0.1% dextrin to an OD of 0.4. Aliquots of 40 µl were spread on second selection plates,
572 supplemented with 0.1% NZ-amine and 0.1% dextrin and 10 mg ml^{-1} uracil, were incubated for 5–7
573 days at 78°C. Newly formed colonies were streaked on fresh second selection plates to ensure single
574 colony selection before each colony was screened by PCR for the genomic absence, presence, or
575 modification, of the *Agl24* gene.

576 **Expression of Agl24 in *S. acidocaldarius***

577 Plasmid transformation into *S. acidocaldarius* cells was performed as described above for the deletion
578 mutants. Cells were spread onto gelrite plates containing Brock medium supplemented with 0.1% NZ-
579 amine and 0.1% dextrin. After incubation for 5–7 days at 75°C large brownish colonies were used to
580 inoculate 50 ml of Brock medium containing 0.1% NZ-amine and 0.1% dextrin and incubated for 3 days
581 of 78°C. Presence of the expression plasmid was confirmed by PCR. For induction of expression the
582 strains were grown either in Brock medium supplemented with either 0.1% NZ-amine and 0.1% dextrin
583 ($MAL_{promotor}$) or with 0.1% NZ-amine and 0.1% L-arabinose ($ARA_{promotor}$) to an OD_{600} of 1.0.
584

585 **Expression and purification of recombinant Agl24 protein from *E. coli***

586 For heterologous Agl24 protein expression, 6 x 1 Liter LB medium was inoculated with 10 ml from an
587 overnight BL21 culture previously transformed with pHD0499. Cells were grown at 16˚C overnight in
588 auto-induction medium (Studier, 2014) containing 30 µg ml^{-1} kanamycin. The cells were harvested by

589 centrifugation at 4,200 x g for 25 minutes and used to isolate the membrane fractions. All subsequent
590 purification steps were carried out at 4˚C. Cells were fractionated by passing four times through an
591 Avestin C3 High Pressure Homogeniser (Biopharma, UK), followed by a 20 min low speed spin at 4,000
592 x g. The resulting supernatant was centrifuged at 200,000 x g for 2 h to obtain the membrane fraction
593 and 2–3 g of membranes were routinely used for isolation of Agl24-GFP-His_8 proteins. Samples were
594 solubilized in 18 ml Buffer 1 (500 mM NaCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4 2.7 mM KCl, pH 7.4, 20
595 mM imidazole, 4 mM TCEP) with the addition of 1% n-dodecyl-β-maltoside (DDM) for 2 h at 4˚C. The
596 sample was two-fold diluted with Buffer 1 and centrifuged at 200,000 x g for 2 h. The supernatant was
597 loaded onto a Ni-Sepharose 6 Fast Flow (GE Healthcare) column with 1 ml of prewashed Ni-NTA-beads.
598 The column was washed with 20 ml of wash buffer (500 mM NaCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4
599 2.7 mM KCl, pH 7.4, 20 mM imidazole, 0.4 mM TCEP, 0.03% DDM) and eluted with 5 x 1ml elution
600 buffer (500 mM NaCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO4 2.7 mM KCl, pH 7.4, 250 mM imidazole, 0.4
601 mM TCEP, 0.03% DDM). Elution fractions were combined and imidazole removed using a HiPrep 26/10
602 desalting column (GE Healthcare) equilibrated with Buffer (1x PBS, 0.03% DDM, 0.4 mM TCEP). Protein
603 concentration was determined using a Bradford reaction (Bio-Rad) and purity was confirmed by SDS–
604 PAGE analysis. The concentrated fractions were separated by SDS–PAGE and stained with Coomassie
605 blue. Protein identity was confirmed by tryptic peptide mass fingerprinting, and the level of purity and
606 molecular weight of the recombinant protein was determined by matrix-assisted laser
607 desorption/ionization time-of-flight mass spectrometry (MALDI–TOF). The analysis was provided by
608 the University of Dundee 'Fingerprints' Proteomics Facility.

609 **Agl24 activity**

610 Activity was measured in a 100 µl reaction volume containing 1 mM UDP-D-GlcNAc, 1 mM acceptor
611 molecule, 5 mM $MgCl_2$, and 5 µg of purified Agl24 in TBS Buffer (150 mM NaCl, 50 mM Tris-HCl, pH
612 7.6); reaction was performed in a PCR cycler at 60°C for 12 h.

613 **MS analysis**

614 MALDI–TOF was used to analyze the acceptors and products of the Agl24 *in vitro* assay. The 100ul
615 reaction samples were purified over 100 mg Sep-Pak C18 cartridges (Waters) pre-equilibrated with 5%
616 EtOH. The bound samples were washed with 800 µl of H_2O and 800 µl of 15% EtOH, eluted in two
617 fractions with (a) 800 µl of 30% EtOH and (b) 800 µl of 60% EtOH. The two elution fractions were dried
618 in a SpeedVac vacuum concentrator and resuspended in 20 µl of 50% MeOH. A 1 µl sample was mixed
619 with 1 µl of 2,5-dihydroxybenzoic acid matrix (15 mg ml^{-1} in 30:70 acetonitrile, 0.1% TFA), and 1 µl was
620 added to the MALDI grid. Samples were analyzed by MALDI in an Autoflex Speed mass spectrometer
621 set up in reflection positive ion mode (Bruker, Germany).

622 **HPLC analysis**

623 With the exception of kinetics reactions, which are detailed below, Agl24 reactions were analyzed
624 using a HPLC assay with, typically, 50 µl samples containing 2.5 mM UDP-GlcNAc, 1.5 mM lipid
625 acceptor, and 1.8 µg of purified Agl24 in a TBS buffer containing 5 mM $MgCl_2$ (150 mM NaCl, 50 mM
626 Tris-HCl, pH 7.5). Reactions were left at 60˚C (or alternative temperatures) for the desired time period
627 and quenched with one equivalent of acetonitrile to precipitate Agl24. Following filtration to remove
628 precipitate, reactions were injected onto an XBridge BEH Amide OBD Prep column (130 Å, 5 µM, 10 x
629 250 mm) at a flow rate of 4 ml min^{-1} using a Dionex UltiMate 3000 system (Thermo Scientific) fitted
630 with a UV detector optimized to detect the *O*-phenyl functional group of the acceptor molecule at 270
631 nm. Each run was 35 minutes using running Buffer A (95% acetonitrile, 10 mM ammonium acetate, pH
632 8) and Buffer B (50% acetonitrile, 10 mM ammonium acetate, pH 8). A linear gradient from 20–80%
633 buffer B was performed over 20 min, followed by an immediate drop back to 20% buffer B for the
634 remaining 15 minutes of the run to re-equilibrate the column to starting conditions. The Agl24 reaction
635 substrates and products typically eluted 8–11 minutes into a run. For kinetic analyses, assays were
636 performed in triplicate (with the exception of 1 mM conc. which was only performed in duplicate) and
637 altered to contain a fixed concentration of lipid-acceptor (0.5 mM) whilst varying the concentration of
638 UDP-GlcNAc (0.5, 0.75, 1, 1.5, 3, 6, 9, 12 mM). Reactions were run for 10 hours before being quenched
639 and, after completion, areas of substrate and product peaks were calculated to determine the reaction
640 conversion. Conversion over 10 hours was converted to pmol per minute, and the resulting data were
641 analyzed using GraphPad Prism v8 to generate a Michaelis–Menten curve and resulting kinetic data.

642 **NMR Analysis**

643 The HPLC-purified Agl24 products (0.5mg–2mg) were dried using a Christ RVC 2-25 speed vacuum
644 fitted with a Christ CT 02-50 cold-trap to remove excess acetonitrile, then freeze dried (Alpha 1-2
645 LDplus, Christ) to remove residual water. Products were subsequently dissolved in 600 µl of D_2O and
646 NMR spectra were recorded at 293K. The spectra were acquired on a Bruker AVANCE III HD 500 MHz
647 NMR Spectrometer equipped with a 5 mm QCPI cryoprobe. For 1D TOCSY experiments, H1' was
648 irradiated at 5.35 ppm, H2' was irradiated at 3.90 ppm, and H1" was irradiated at 4.49 ppm (Fig S8). A
649 combination of ^{1}H ^{1}H COSY, 1D TOCSY, and ^{1}H ^{13}C HSQC experiments were used to fully assign the ^{1}H
650 and ^{13}C signals for the Agl24 reaction product. Full ^{1}H and ^{13}C chemical shift assignments can be found
651 in Table 1 and are recorded with respect to the residual HDO signal at 4.7 ppm.

652 **Phylogenetic analysis**

653 The eukaryotic Alg13-Alg14 and the bacterial EpsF-EpsE sequences were taken from the datasets in
654 (Lombard, 2016), since our own databanks (25118 bacterial and 1611 eukaryotic genomes) contained

too many and too divergent hits that could possibly render downstream analyses unfeasible. For Archaea, HMM searches were impossible, since they recovered too many and too divergent possible homologs to handle, while DIAMOND searches were not sensitive enough. Ultimately, we searched for homologs with DIAMOND (--more-sensitive -k 0 & default e-value cutoff) (Buchfink et al., 2015) using multiple seeds from (Lombard, 2016) to cover the different clades of archaeal homologs: 1) the *S. acidocaldarius* Agl24 from this study; 2) the Alg13/EpsF-like and Alg14/EpsE-like homologs from *Methanothrix soehngenii* (NCBI: AEB67939.1 and AEB67938.1 respectively); 3) AIC14646.1 from *Nitrososphaera viennensis* (TACK clade); 4) ABX11947.1 from *Nitrosopumilus maritimus* (additional Alg13/EpsF-like Thaumarchaeota clade); 5) AAB99267.1 from *Methanocaldococcus jannaschii* (methanogen clade). For the first two, we repeated the homology search with the same parameters using all the hits from the previous round as seeds to ensure we did not miss any sequences from our primary clades of interest. The hits from each search were aligned with MAFFT L-INS-I v7.453 (Katoh and Standley, 2013) and preliminary phylogenies were reconstructed with IQ-TREE 2.0.5 (Minh et al., 2020) (-m TEST -alrt 1000). From these phylogenies we isolated the sequences forming well-supported branches corresponding to the ones in (Lombard, 2016), albeit with increased diversity plus some additional clades of potential interest. The individual datasets (archaeal, bacterial, eukaryotic) were then fused into single Alg13/EpsF-like and Alg14/EpsE-like datasets that were aligned as above and trimmed with ClipKIT (kpigappy) (Steenwyk et al., 2020). Then phylogenies were inferred with IQ-TREE 2.0.5 (Minh et al., 2020) under the model selected by ModelFinder (Kalyaanamoorthy et al., 2017), and ultrafast bootstrap (Hoang et al., 2018), aLRT SH-like (Guindon et al., 2010), and approximate Bayesian (Anisimova et al., 2011) branch support tests (-m MFP -bb 1000 -alrt 1000 -abayes). The trees were visualized in iTol (Letunic and Bork, 2019).

Author contributions:

B. H. M., S. V. A., and H. C. D. conceptualization; B. H. M. data curation, initial BLAST searches, deletion analysis, protein alignments, protein expression, protein purification, alanine substitutions, activity analyses, MALDI analyses, visualization; B. A. W. HPLC activity analyses and NMR analyses; P. S. A. phylogenetic analyses, B. H. M.., B. A. W., P. S. A. and H. C. D. investigation; B. H. M. and H. C. D. validation; B. H. M. and H. C. D. methodology; B. H. M. writing-original draft; B. H. M., P. S. A., B. A. W. , S. V. A., and H. C. D. writing-review and editing; H. C. D. supervision; H. C. D. and S. V. A. funding acquisition

Acknowledgments

687 We thank Vladimir I. Torgov, Vladimir V. Veselovsky, and Leonid L. Danilov from the N. D. Zelinsky
688 Institute of Organic Chemistry, Russian Academy of Sciences, Moscow, Russia for sharing the synthetic
689 acceptor molecules. We thank Prof. Alexander J. Probst from the University Duisburg-Essen for fruitful
690 discussions and Mr. Tom Snelling for assay assistance.

691 **Funding**

692 The HCD laboratory (BHM, BAW and HCD) is supported by The Wellcome Trust and Royal Society Grant
693 109357/Z/15/Z and the University of Dundee Wellcome Trust Fund 105606/Z/14/Z. SVA and BHM have
694 been supported by the SFG grant SFB 1381 (Deutsche Forschungsgemeinschaft (German Research
695 Foundation) under project no. 403222702-SFB 1381)

696 For the purpose of Open Access, the authors have applied a CC BY public copyright license to any
697 Author Accepted Manuscript version arising from this submission.

References

698 ABU-QARN, M., YURIST-DOUTSCH, S., GIORDANO, A., TRAUNER, A., MORRIS, H. R., HITCHEN, P.,
699 MEDALIA, O., DELL, A. & EICHLER, J. 2007. *Haloferax volcanii* AglB and AglD are involved in *N*-
700 glycosylation of the S-layer glycoprotein and proper assembly of the surface layer. *J Mol Biol,*
701 374**,** 1224-36.
702 ALBERS, S., EICHLER, J. & AEBI, M. 2017. Archaea. *In:* VARKI, A., CUMMINGS, R. D., ESKO, J. D., STANLEY,
703 P., HART, G. W., AEBI, M., DARVILL, A. G., KINOSHITA, T., PACKER, N. H., PRESTEGARD, J. H.,
704 SCHNAAR, R. L. & SEEBERGER, P. H. (eds.) *Essentials of Glycobiology.* Cold Spring Harbor (NY):
705 Cold Spring Harbor Laboratory Press
706 ALBERS, S. V. & MEYER, B. H. 2011. The archaeal cell envelope. *Nat Rev Microbiol,* 9**,** 414-426.
707 ANISIMOVA, M., GIL, M., DUFAYARD, J. F., DESSIMOZ, C. & GASCUEL, O. 2011. Survey of Branch Support
708 Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based
709 Approximation Schemes. *Syst Biol,* 60**,** 685-699.
710 APWEILER, R., HERMJAKOB, H. & SHARON, N. 1999. On the frequency of protein glycosylation, as
711 deduced from analysis of the SWISS-PROT database. *Biochim Biophys Acta,* 1473**,** 4-8.
712 AVERBECK, N., KEPPLER-ROSS, S. & DEAN, N. 2007. Membrane topology of the Alg14 endoplasmic
713 reticulum UDP-GlcNAc transferase subunit. *J Biol Chem,* 282**,** 29081-8.
714 BAKER, P. J., BRITTON, K. L., RICE, D. W., ROB, A. & STILLMAN, T. J. 1992. Structural consequences of
715 sequence patterns in the fingerprint region of the nucleotide binding fold. Implications for
716 nucleotide specificity. *J Mol Biol,* 228**,** 662-71.
717 BAUM, D. A. & BAUM, B. 2014. An inside-out origin for the eukaryotic cell. *BMC Biol,* 12**,** 76.

718 BICKEL, T., LEHLE, L., SCHWARZ, M., AEBI, M. & JAKOB, C. A. 2005. Biosynthesis of lipid-linked
719 oligosaccharides in *Saccharomyces cerevisiae*: Alg13p and Alg14p form a complex required for
720 the formation of GlcNAc(2)-PP-dolichol. *J Biol Chem,* 280**,** 34500-6.
721 BORATYN, G. M., SCHAFFER, A. A., AGARWALA, R., ALTSCHUL, S. F., LIPMAN, D. J. & MADDEN, T. L.
722 2012. Domain enhanced lookup time accelerated BLAST. *Biol Direct,* 7**,** 12.
723 BOUHSS, A., TRUNKFIELD, A. E., BUGG, T. D. & MENGIN-LECREULX, D. 2008. The biosynthesis of
724 peptidoglycan lipid-linked intermediates. *FEMS Microbiol Rev,* 32**,** 208-33.
725 BROCK, T. D., BROCK, K. M., BELLY, R. T. & WEISS, R. L. 1972. *Sulfolobus* - new genus of sulfur-oxidizing
726 bacteria living at low pH and high temperature. *Arch Microbiol,* 84**,** 54-&.
727 BROWN, K., VIAL, S. C., DEDI, N., WESTCOTT, J., SCALLY, S., BUGG, T. D., CHARLTON, P. A. & CHEETHAM,
728 G. M. 2013. Crystal structure of the *Pseudomonas aeruginosa* MurG: UDP-GlcNAc substrate
729 complex. *Protein Pept Lett,* 20**,** 1002-8.
730 BUCHFINK, B., XIE, C. & HUSON, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat*
731 *Methods,* 12**,** 59-60.
732 BUGG, T. D. & BRANDISH, P. E. 1994. From peptidoglycan to glycoproteins: common features of lipid-
733 linked oligosaccharide biosynthesis. *FEMS Microbiol Lett,* 119**,** 255-62.
734 BURDA, P. & AEBI, M. 1999. The dolichol pathway of *N*-linked glycosylation. *Biochim Biophys Acta,*
735 1426**,** 239-57.
736 CARAMELO, J. J. & PARODI, A. J. 2007. How sugars convey information on protein conformation in the
737 endoplasmic reticulum. *Semin Cell Dev Biol,* 18**,** 732-42.
738 CARUGO, O. & ARGOS, P. 1997. NADP-dependent enzymes. I: Conserved stereochemistry of cofactor
739 binding. *Proteins,* 28**,** 10-28.
740 CHABAN, B., LOGAN, S. M., KELLY, J. F. & JARRELL, K. F. 2009. AglC and AglK are involved in biosynthesis
741 and attachment of diacetylated glucuronic acid to the *N*-glycan in *Methanococcus voltae*. *J*
742 *Bacteriol,* 191**,** 187-95.
743 CHABAN, B., VOISIN, S., KELLY, J., LOGAN, S. M. & JARRELL, K. F. 2006. Identification of genes involved
744 in the biosynthesis and attachment of *Methanococcus voltae N*-linked glycans: insight into *N*-
745 linked glycosylation pathways in Archaea. *Mol Microbiol,* 61**,** 259-68.
746 CHANTRET, I., DANCOURT, J., BARBAT, A. & MOORE, S. E. 2005. Two proteins homologous to the N-
747 and C-terminal domains of the bacterial glycosyltransferase MurG are required for the second
748 step of dolichyl-linked oligosaccharide synthesis in *Saccharomyces cerevisiae*. *J Biol Chem,* 280**,**
749 9236-42.
750 CROUVOISIER, M., AUGER, G., BLANOT, D. & MENGIN-LECREULX, D. 2007. Role of the amino acid
751 invariants in the active site of MurG as evaluated by site-directed mutagenesis. *Biochimie,* 89**,**
752 1498-508.

753 DE VUYST, L., DE VIN, F., VANINGELGEM, F. & DEGEEST, B. 2001. Recent developments in the
754 biosynthesis and applications of heteropolysaccharides from lactic acid bacteria. *Int Dairy J,*
755 11**,** 687-707.
756 EME, L., SPANG, A., LOMBARD, J., STAIRS, C. W. & ETTEMA, T. J. G. 2017. Archaea and the origin of
757 eukaryotes. *Nat Rev Microbiol,* 15**,** 711-723.
758 FUJINAMI, D., TAGUCHI, Y. & KOHDA, D. 2017. Asn-linked oligosaccharide chain of a crenarchaeon,
759 *Pyrobaculum calidifontis,* is reminiscent of the eukaryotic high-mannose-type glycan.
760 *Glycobiology*.
761 GAO, X. D., MORIYAMA, S., MIURA, N., DEAN, N. & NISHIMURA, S. 2008. Interaction between the C
762 termini of Alg13 and Alg14 mediates formation of the active UDP-N-acetylglucosamine
763 transferase complex. *J Biol Chem,* 283**,** 32534-41.
764 GAO, X. D., TACHIKAWA, H., SATO, T., JIGAMI, Y. & DEAN, N. 2005. Alg14 recruits Alg13 to the
765 cytoplasmic face of the endoplasmic reticulum to form a novel bipartite UDP-N-
766 acetylglucosamine transferase required for the second step of N-linked glycosylation. *J Biol*
767 *Chem,* 280**,** 36254-62.
768 GLOSTER, T. M. 2014. Advances in understanding glycosyltransferases from a structural perspective.
769 *Curr Opin Struct Biol,* 28**,** 131-41.
770 GUINDON, S., DUFAYARD, J. F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New
771 Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the
772 Performance of PhyML 3.0. *Syst Biol,* 59**,** 307-321.
773 HA, S., CHANG, E., LO, M. C., MEN, H., PARK, P., GE, M. & WALKER, S. 1999. The kinetic characterization
774 of *Escherichia coli* MurG using synthetic substrate analogues. *J Am Chem Soc,* 121**,** 8415-8426.
775 HA, S., WALKER, D., SHI, Y. G. & WALKER, S. 2000. The 1.9 angstrom crystal structure of *Escherichia coli*
776 MurG, a membrane-associated glycosyltransferase involved in peptidoglycan biosynthesis.
777 *Protein Sci,* 9**,** 1045-1052.
778 HELENIUS, A. & AEBI, M. 2004. Roles of N-linked glycans in the endoplasmic reticulum. *Annu Rev*
779 *Biochem,* 73**,** 1019-49.
780 HOANG, D. T., CHERNOMOR, O., VON HAESELER, A., MINH, B. Q. & VINH, L. S. 2018. UFBoot2:
781 Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol,* 35**,** 518-522.
782 HU, Y., CHEN, L., HA, S., GROSS, B., FALCONE, B., WALKER, D., MOKHTARZADEH, M. & WALKER, S. 2003.
783 Crystal structure of the MurG:UDP-GlcNAc complex reveals common structural principles of a
784 superfamily of glycosyltransferases. *Proc Natl Acad Sci U S A,* 100**,** 845-9.
785 HUET, J., SCHNABEL, R., SENTENAC, A. & ZILLIG, W. 1983. Archaebacteria and eukaryotes possess DNA-
786 dependent RNA polymerases of a common type. *EMBO J,* 2**,** 1291-4.

787 IMACHI, H., NOBU, M. K., NAKAHARA, N., MORONO, Y., OGAWARA, M., TAKAKI, Y., TAKANO, Y.,
788 UEMATSU, K., IKUTA, T., ITO, M., MATSUI, Y., MIYAZAKI, M., MURATA, K., SAITO, Y., SAKAI, S.,
789 SONG, C., TASUMI, E., YAMANAKA, Y., YAMAGUCHI, T., KAMAGATA, Y., TAMAKI, H. & TAKAI,
790 K. 2020. Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature,* 577**,** 519-525.
791 JARRELL, K. F., DING, Y., MEYER, B. H., ALBERS, S. V., KAMINSKI, L. & EICHLER, J. 2014. N-linked
792 glycosylation in Archaea: a structural, functional, and genetic analysis. *Microbiol Mol Biol Rev,*
793 78**,** 304-341.
794 KALYAANAMOORTHY, S., MINH, B. Q., WONG, T. K. F., VON HAESELER, A. & JERMIIN, L. S. 2017.
795 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods,* 14**,** 587-
796 589.
797 KAMINSKI, L., LURIE-WEINBERGER, M. N., ALLERS, T., GOPHNA, U. & EICHLER, J. 2013. Phylogenetic-
798 and genome-derived insight into the evolution of N-glycosylation in Archaea. *Mol Phylogenet*
799 *Evol,* 68**,** 327-335.
800 KATOH, K. & STANDLEY, D. M. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
801 Improvements in Performance and Usability. *Mol Biol Evol,* 30**,** 772-780.
802 KLEEREBEZEM, M., VAN KRANENBURG, R., TUINIER, R., BOELS, I. C., ZOON, P., LOOIJESTEIJN, E.,
803 HUGENHOLTZ, J. & DE VOS, W. M. 1999. Exopolysaccharides produced by *Lactococcus lactis*:
804 from genetic engineering to improved rheological properties? *Antonie Van Leeuwenhoek* 76**,**
805 357-365.
806 KOLKMAN, M. A. B., VANDERZEIJST, B. A. M. & NUIJTEN, P. J. M. 1997. Functional analysis of
807 glycosyltransferases encoded by the capsular polysaccharide biosynthesis locus of
808 *Streptococcus pneumoniae* serotype 14. *J Biol Chem,* 272**,** 19502-19508.
809 KOONIN, E. V. 2015. Archaeal ancestors of eukaryotes: not so elusive any more. *BMC Biol,* 13.
810 KUROSAWA, N. & GROGAN, D. W. 2005. Homologous recombination of exogenous DNA with the
811 *Sulfolobus acidocaldarius* genome: properties and uses. *FEMS Microbiol Lett,* 253**,** 141-9.
812 LAIRSON, L. L., HENRISSAT, B., DAVIES, G. J. & WITHERS, S. G. 2008. Glycosyltransferases: structures,
813 functions, and mechanisms. *Annu Rev Biochem,* 77**,** 521-55.
814 LARKIN, A. & IMPERIALI, B. 2011. The expanding horizons of asparagine-linked glycosylation.
815 *Biochemistry,* 50**,** 4411-26.
816 LEHLE, L., STRAHL, S. & TANNER, W. 2006. Protein glycosylation, conserved from yeast to man: a model
817 organism helps elucidate congenital human diseases. *Angew Chem Int Ed Engl,* 45**,** 6802-18.
818 LETUNIC, I. & BORK, P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments.
819 *Nucleic Acids Res,* 47**,** W256-W259.
820 LOMBARD, J. 2016. The multiple evolutionary origins of the eukaryotic N-glycosylation pathway. *Biol*
821 *Direct,* 11**,** 36.

822 LOMBARD, V., GOLACONDA RAMULU, H., DRULA, E., COUTINHO, P. M. & HENRISSAT, B. 2014. The
823 carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res,* 42**,** D490-5.
824 LU, J., TAKAHASHI, T., OHOKA, A., NAKAJIMA, K., HASHIMOTO, R., MIURA, N., TACHIKAWA, H. & GAO,
825 X. D. 2012. Alg14 organizes the formation of a multiglycosyltransferase complex involved in
826 initiation of lipid-linked oligosaccharide biosynthesis. *Glycobiology,* 22**,** 504-16.
827 LU, J., TAKAHASHI, T., OHOKA, A., NAKAJIMA, K. I., HASHIMOTO, R., MIURA, N., TACHIKAWA, H. & GAO,
828 X. D. 2011. Alg14 organizes the formation of a multi-glycosyltransferase complex involved in
829 initiation of lipid-linked oligosaccharide biosynthesis. *Glycobiology*.
830 MEN, H. B., PARK, P., GE, M. & WALKER, S. 1998. Substrate synthesis and activity assay for MurG. *J Am*
831 *Chem Soc,* 120**,** 2484-2485.
832 MESTROM, L., PRZYPIS, M., KOWALCZYKIEWICZ, D., POLLENDER, A., KUMPF, A., MARSDEN, S. R.,
833 BENTO, I., JARZEBSKI, A. B., SZYMANSKA, K., CHRUSCIEL, A., TISCHLER, D., SCHOEVAART, R.,
834 HANEFELD, U. & HAGEDOORN, P. L. 2019. Leloir Glycosyltransferases in Applied Biocatalysis:
835 A Multidisciplinary Approach. *Int J Mol Sci,* 20.
836 MEYER, B. H. & ALBERS, S.-V. 2020. Archaeal Cell Walls. *In eLS , John Wiley & Sons, Ltd (Ed.).* .
837 MEYER, B. H. & ALBERS, S. V. 2013. Hot and sweet: protein glycosylation in Crenarchaeota. *Biochem*
838 *Soc Trans,* 41**,** 384-392.
839 MEYER, B. H. & ALBERS, S. V. 2014. AglB, catalyzing the oligosaccharyl transferase step of the archaeal
840 N-glycosylation process, is essential in the thermoacidophilic crenarchaeon *Sulfolobus*
841 *acidocaldarius*. *Microbiologyopen,* 3**,** 531-43.
842 MEYER, B. H., PEYFOON, E., DIETRICH, C., HITCHEN, P., PANICO, M., MORRIS, H. R., DELL, A. & ALBERS,
843 S. V. 2013. Agl16, a thermophilic glycosyltransferase mediating the last step of N-glycan
844 biosynthesis in the thermoacidophilic crenarchaeon *Sulfolobus acidocaldarius*. *J Bacteriol,* 195**,**
845 2177-2186.
846 MEYER, B. H., SHAMS-ELDIN, H. & ALBERS, S. V. 2017. AglH, a thermophilic UDP-N-acetylglucosamine-
847 1-phosphate:dolichyl phosphate GlcNAc-1-phosphotransferase initiating protein N-
848 glycosylation pathway in *Sulfolobus acidocaldarius*, is capable of complementing the eukaryal
849 Alg7. *Extremophiles,* 21**,** 121-134.
850 MINH, B. Q., SCHMIDT, H. A., CHERNOMOR, O., SCHREMPF, D., WOODHAMS, M. D., VON HAESELER,
851 A. & LANFEAR, R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic
852 Inference in the Genomic Era. *Mol Biol Evol,* 37**,** 1530-1534.
853 NIKOLAYEV, S., COHEN-ROSENZWEIG, C. & EICHLER, J. 2020. Evolutionary considerations of the
854 oligosaccharyltransferase AglB and other aspects of N-glycosylation across Archaea. *Mol*
855 *Phylogenet Evol,* 153**,** 106951.

856 NOTHAFT, H. & SZYMANSKI, C. M. 2010. Protein glycosylation in bacteria: sweeter than ever. *Nat Rev*
857 *Microbiol,* 8**,** 765-78.
858 NOTHAFT, H. & SZYMANSKI, C. M. 2013. Bacterial protein N-glycosylation: new perspectives and
859 applications. *J Biol Chem,* 288**,** 6912-20.
860 PALMIERI, G., BALESTRIERI, M., PETER-KATALINIC, J., POHLENTZ, G., ROSSI, M., FIUME, I. & POCSFALVI,
861 G. 2013. Surface-exposed glycoproteins of hyperthermophilic *Sulfolobus solfataricus* P2 show
862 a common glycosylation profile. *J Proteome Res*.
863 PETITJEAN, C., DESCHAMPS, P., LOPEZ-GARCIA, P. & MOREIRA, D. 2015. Rooting the Domain Archaea
864 by Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota.
865 *Genome Biol Evol,* 7**,** 191-204.
866 PEYFOON, E., MEYER, B., HITCHEN, P. G., PANICO, M., MORRIS, H. R., HASLAM, S. M., ALBERS, S. V. &
867 DELL, A. 2010. The S-layer glycoprotein of the crenarchaeote *Sulfolobus acidocaldarius* is
868 glycosylated at multiple sites with chitobiose-linked *N*-glycans. *Archaea,* 2010.
869 RAMAN, S., LANGE, O. F., ROSSI, P., TYKA, M., WANG, X., ARAMINI, J., LIU, G., RAMELOT, T. A., ELETSKY,
870 A., SZYPERSKI, T., KENNEDY, M. A., PRESTEGARD, J., MONTELIONE, G. T. & BAKER, D. 2010.
871 NMR structure determination for larger proteins using backbone-only data. *Science,* 327**,**
872 1014-8.
873 REILY, C., STEWART, T. J., RENFROW, M. B. & NOVAK, J. 2019. Glycosylation in health and disease. *Nat*
874 *Rev Nephrol,* 15**,** 346-366.
875 SAMUELSON, J., BANERJEE, S., MAGNELLI, P., CUI, J., KELLEHER, D. J., GILMORE, R. & ROBBINS, P. W.
876 2005. The diversity of dolichol-linked precursors to Asn-linked glycans likely results from
877 secondary loss of sets of glycosyltransferases. *Proc Natl Acad Sci U S A,* 102**,** 1548-53.
878 SCHWARZ, F. & AEBI, M. 2011. Mechanisms and principles of N-linked protein glycosylation. *Curr Opin*
879 *Struct Biol,* 21**,** 576-82.
880 SHRIMAL, S. & GILMORE, R. 2019. Oligosaccharyltransferase structures provide novel insight into the
881 mechanism of asparagine-linked glycosylation in prokaryotic and eukaryotic cells.
882 *Glycobiology,* 29**,** 288-297.
883 SPANG, A., SAW, J. H., JORGENSEN, S. L., ZAREMBA-NIEDZWIEDZKA, K., MARTIJN, J., LIND, A. E., VAN
884 EIJK, R., SCHLEPER, C., GUY, L. & ETTEMA, T. J. G. 2015. Complex archaea that bridge the gap
885 between prokaryotes and eukaryotes. *Nature,* 521**,** 173-9.
886 STEENWYK, J. L., BUIDA, T. J., LI, Y., SHEN, X.-X. & ROKAS, A. 2020. ClipKIT: a multiple sequence
887 alignment-trimming algorithm for accurate phylogenomic inference. 2020.06.08.140384.
888 STUDIER, F. W. 2014. Stable expression clones and auto-induction for protein production in *E. coli*.
889 *Methods Mol Biol,* 1091**,** 17-32.

890 TAGUCHI, Y., FUJINAMI, D. & KOHDA, D. 2016. Comparative analysis of archaeal lipid-linked
891 oligosaccharides that serve as oligosaccharide donors for Asn glycosylation. *J Biol Chem,* 291**,**
892 11042-54.
893 VAN KRANENBURG, R., VOS, H. R., VAN SWAM, I. I., KLEEREBEZEM, M. & DE VOS, W. M. 1999.
894 Functional analysis of glycosyltransferase genes from *Lactococcus lactis* and other gram-
895 positive cocci: Complementation, expression, and diversity. *J Bacteriol,* 181**,** 6347-6353.
896 VAN WOLFEREN, M., SHAJAHAN, A., HEINRICH, K., BRENZINGER, S., BLACK, I. M., WAGNER, A., BRIEGEL,
897 A., AZADI, P. & ALBERS, S.-V. 2020. Species-Specific Recognition of *Sulfolobales* Mediated by
898 UV-Inducible Pili and S-Layer Glycosylation Patterns. *mBio,* 11**,** e03014-19.
899 VARKI, A. 1993. Biological roles of oligosaccharides - all of the theories are correct. *Glycobiology,* 3**,** 97-
900 130.
901 VARKI, A., CUMMINGS, R., ESKO, J., FREEZE, H., HART, G., MARTH, J. & EDITORS 1999. Chapter 17,
902 Glycosyltransferases. *Essentials of Glycobiolog.* Cold Spring Harbor (NY): Cold Spring Harbor
903 Laboratory Press.
904 WAGNER, M., VAN WOLFEREN, M., WAGNER, A., LASSAK, K., MEYER, B. H., REIMANN, J. & ALBERS, S.
905 V. 2012. Versatile genetic tool Box for the crenarchaeote *Sulfolobus acidocaldarius*. *Front*
906 *Microbiol,* 3**,** 214.
907 WALDO, G. S., STANDISH, B. M., BERENDZEN, J. & TERWILLIGER, T. C. 1999. Rapid protein-folding assay
908 using green fluorescent protein. *Nat Biotechnol,* 17**,** 691-5.
909 WANG, X., WELDEGHIORGHIS, T., ZHANG, G., IMPERIALI, B. & PRESTEGARD, J. H. 2008. Solution
910 structure of Alg13: the sugar donor subunit of a yeast N-acetylglucosamine transferase.
911 *Structure,* 16**,** 965-75.
912 WATERHOUSE, A., BERTONI, M., BIENERT, S., STUDER, G., TAURIELLO, G., GUMIENNY, R., HEER, F. T.,
913 DE BEER, T. A. P., REMPFER, C., BORDOLI, L., LEPORE, R. & SCHWEDE, T. 2018. SWISS-MODEL:
914 homology modelling of protein structures and complexes. *Nucleic Acids Res,* 46**,** W296-W303.
915 WHITFIELD, C., SZYMANSKI, C. M. & AEBI, M. 2017. Eubacteria. *In:* VARKI, A., CUMMINGS, R. D., ESKO,
916 J. D., STANLEY, P., HART, G. W., AEBI, M., DARVILL, A. G., KINOSHITA, T., PACKER, N. H.,
917 PRESTEGARD, J. H., SCHNAAR, R. L., SEEBERGER, P. H. & EDITORS (eds.) *Essentials of*
918 *Glycobiology.* Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press
919 WILD, R., KOWAL, J., EYRING, J., NGWA, E. M., AEBI, M. & LOCHER, K. P. 2018. Structure of the yeast
920 oligosaccharyltransferase complex gives insight into eukaryotic N-glycosylation. *Science,* 359**,**
921 545-550.
922 WILSON, C. M. & HIGH, S. 2007. Ribophorin I acts as a substrate-specific facilitator of N-glycosylation.
923 *J Cell Sci,* 120**,** 648-57.

924 WOESE, C. R. & FOX, G. E. 1977. Phylogenetic structure of the prokaryotic domain: the primary
925 kingdoms. *Proc Natl Acad Sci U S A,* 74**,** 5088-5090.
926 ZAHRINGER, U., MOLL, H., HETTMANN, T., KNIREL, V. A. & SCHAFER, G. 2000. Cytochrome $b_{558/566}$ from
927 the archaeon *Sulfolobus acidocaldarius* has a unique Asn-linked highly branched
928 hexasaccharide chain containing 6-sulfoquinovose. *Eur J Biochem,* 267**,** 4144-4149.
929 ZAREMBA-NIEDZWIEDZKA, K., CACERES, E. F., SAW, J. H., BACKSTROM, D., JUZOKAITE, L., VANCAESTER,
930 E., SEITZ, K. W., ANANTHARAMAN, K., STARNAWSKI, P., KJELDSEN, K. U., STOTT, M. B.,
931 NUNOURA, T., BANFIELD, J. F., SCHRAMM, A., BAKER, B. J., SPANG, A. & ETTEMA, T. J. 2017.
932 Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature,* 541**,** 353-358.
933 ZHANG, C., PHILLIPS, A. P. R., WIPFLER, R. L., OLSEN, G. J. & WHITAKER, R. J. 2018. The essential genome
934 of the crenarchaeal model *Sulfolobus islandicus*. *Nat Commun,* 9**,** 4908.
935 ZIELINSKA, D. F., GNAD, F., WISNIEWSKI, J. R. & MANN, M. 2010. Precision mapping of an *in vivo N*-
936 glycoproteome reveals rigid topological and sequence constraints. *Cell,* 141**,** 897-907.
937 ZILLIG, W., KLENK, H. P., PALM, P., PUHLER, G., GROPP, F., GARRETT, R. A. & LEFFERS, H. 1989. The
938 phylogenetic relations of DNA-dependent RNA polymerases of archaebacteria, eukaryotes,
939 and eubacteria. *Can J Microbiol,* 35**,** 73-80.
940 ZORZOLI, A., MEYER, B. H., ADAIR, E., TORGOV, V. I., VESELOVSKY, V. V., DANILOV, L. L., UHRIN, D. &
941 DORFMUELLER, H. C. 2019. Group A, B, C, and G Streptococcus Lancefield antigen biosynthesis
942 is initiated by a conserved alpha-d-GlcNAc-beta-1,4-l-rhamnosyltransferase. *J Biol Chem,* 294**,**
943 15237-15256.