Mining for novel cyclomaltodextrin glucanotransferases unravels the carbohydrate metabolism pathway via cyclodextrins in Thermoanaerobacterales

Sara Centeno-Leija1*, Laura Espinosa-Barrera2, Beatriz Velazquez-Cruz2, Yair Cárdenas-Conejo1, Raúl Virgen-Ortiz2, Georgina Valencia-Cruz3, Roberto A. Saenz4, Yerli Marín-Tovar5, Saúl Gómez-Manzo6, Beatriz Hernández-Ochoa7, Luz María Rocha-Ramírez8, Rocío Zatarain-Palacios9, Juan A. Osuna-Castro10, Agustín López-Munguía11 & Hugo Serrano-Posada12*

Carbohydrate metabolism via cyclodextrins (CM-CD) is an uncommon starch-converting pathway that thoroughly depends on extracellular cyclomaltodextrin glucanotransferases (CGTases) to transform the surrounding starch substrate to α-(1,4)-linked oligosaccharides and cyclodextrins (CDs). The CM-CD pathway has emerged as a convenient microbial adaptation to thrive under extreme temperatures, as CDs are functional amphipathic toroids with higher heat-resistant values than linear dextrans. Nevertheless, although the CM-CD pathway has been described in a few mesophilic bacteria and archaea, it remains obscure in extremely thermophilic prokaryotes (T_{opt} ≥ 70 °C). Here, a new monophyletic group of CGTases with an exceptional three-domain ABC architecture was detected by (meta)genome mining of extremely thermophilic Thermoanaerobacterales living in a wide variety of hot starch-poor environments on Earth. Functional studies of a representative member, CldA, showed a maximum activity in a thermoacidophilic range (pH 4.0 and 80 °C) with remarkable product diversification that yielded a mixture of α:β:γ-CDs (34:62:4) from soluble starch, as well as G3–G7 linear dextrans and fermentable sugars as the primary products. Together, comparative genomics and predictive functional analysis, combined with data of the functionally characterized key proteins of the gene clusters encoding CGTases, revealed the CM-CD pathway in Thermoanaerobacterales and showed that it is involved in the synthesis, transportation, degradation, and metabolic assimilation of CDs.

1Consejo Nacional de Ciencia y Tecnología, Laboratorio de Biología Sintética, Estructural y Molecular, Laboratorio de Agrobiotecnología, Tecnoparque CLQ, Universidad de Colima, Carretera Los Limones-Loma de Juárez, 28627 Colima, Colima, Mexico. 2Laboratorio de Biología Sintética, Estructural y Molecular, Laboratorio de Agrobiotecnología, Tecnoparque CLQ, Universidad de Colima, Carretera Los Limones-Loma de Juárez, 28627 Colima, Colima, Mexico. 3Centro Universitario de Investigaciones Biomédicas, Universidad de Colima, Avenida 25 de julio 965, Colonia Villa de San Sebastián, 28045 Colima, Colima, Mexico. 4Facultad de Ciencias, Universidad de Colima, Bernal Díaz del Castillo 340, 28045 Colima, Colima, Mexico. 5Laboratorio de Bioquímica Estructural, Departamento de Medicina Molecular y Bioprocesos, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Avenida Universidad 2001, Colonia Chamilpa, 62210 Cuernavaca, Mexico. 6Laboratorio de Bioquímica Genética, Instituto Nacional de Pediatría, Secretaría de Salud, 04530 Mexico City, Mexico. 7Laboratorio de Inmunólogia y Biología Celular, Hospital Infantil de México Federico Gómez, Secretaría de Salud, 06720 Mexico City, Mexico. 8Unidad de Investigación en Enfermedades Infecciosas, Hospital Infantil de México Federico Gómez, Dr. Márquez No. 162, Colonia Doctores, 06720 Delegación Cuauhtémoc, Mexico. 9Escuela de Medicina General, Universidad José Martí, Bosques del Decán 351, 28089 Colima, Colima, México. 10Facultad de Ciencias Biológicas y Agropecuarias, Universidad de Colima, Autopista Colima-Manzanillo, 28100 Tecomán, Colima, Mexico. 11Instituto de Biotecnología, Universidad Nacional Autónoma de México, Avenida Universidad 2001, Colonia Chamilpa, 62210 Cuernavaca, Morelos, Mexico. 12*email: scenteno0@ucol.mx; hserrano0@ucol.mx
Cyclodextrins (CDs) are cyclic α-(1,4)-linked oligosaccharides that commonly consist of six (α-CD), seven (β-CD) and eight (γ-CD) glucopyranose units, forming a unique truncated cone structure with a hydrophobic central cavity and a hydrophilic outer surface. The CD structure enables the formation of reversible binding of nonpolar guest molecules (e.g., organic, inorganic, or biological molecules) to increase their solubility, stability, and bioavailability. Since functional characterizations of CDs have been primarily addressed for biotechnological and pharmaceutical applications, their physiological purpose has not been thoroughly discussed. Nevertheless, the role of CDs seems to be related to resource competition in microbial communities, such as monopolizing substrate availability or mitigating the toxicity of surrounding organic substrates and volatiles, as well as carrying antimicrobial and signaling molecules. CDs also act as surfactants by increasing the bioavailability of hydrocarbons in microbial communities living in oil reservoirs. Furthermore, because the glass transition temperature (Tg) of CDs is higher than monosaccharides and linear dextrans, starch conversion to CDs is particularly valuable for survival in high-temperature environments.

CDs are synthesized by enzymatic conversion using cyclomaltodextrin glucanotransferases (CGTases; EC 2.4.1.19) through an intramolecular transglycosylation of glucosyl intermediates using starch as substrate (cyclization activity). CGTases are enzymes that belong to subfamily 2 of the glycoside hydrolase family 13 (GH13). GH13 family proteins (malto/dextrin/starch-active enzymes) are the second-largest family of glycoside hydrolases and the principal group of enzymes related to α-amylases, encompassing 44 subfamilies. GH13 family enzymes belong to the clan GH-H along with GH70 (sucrose/starch/maltooligosaccharides-active enzymes) and GH77 (amylomaltases) families. All members of the clan GH-H display a catalytic TIM-barrel topology with at least four conserved sequence regions (CSR) from I to IV and display an α-retaining double displacement catalytic mechanism. The four CSR I-IV motifs of the GH13 family contain functionally important residues conserved among CGTases, including an acidic catalytic triad Asp/Glu/Asp, as well as an Arg residue located on the second CSR I-VI motif. The catalytic nucleophile Asp/Glu163. Most of the CSR motifs contain several residues that distinguish the specificities of CGTases from those of other GH13 members. Similar to α-amylases, CGTases can also produce linear oligosaccharides through hydrolysis, disproportionation, or coupling activities. According to the Carbohydrate-Active enZYmes (CAZy; http://www.cazy.org) database, GH13.2 contains 51 characterized enzymes of known sequence that have been isolated from bacteria and archaea, where ~80% belong to the well-characterized Gram-positive (G+) mesophilic Bacilli class bacteria, which are distinguished by synthesizing CDs as the primary catalytic product. The overall CGTase fold comprises a multidomain architecture of five domains (ABCDε; ~700 residues in total), where domain A adopts a TIM-barrel topology and domain B is found as a proteobacterial loop inserted into domain A. While domains A and B comprise the enzyme active site, the C and E domains adopt β-sandwich folding and contain maltose-binding sites (MBS) for substrate binding. Nevertheless, while the E domain belongs to the carbohydrate-binding module family 20 (ECBM20) and contains MBS1 and MBS2 involved in starch-binding, the C domain contains MBS3-5. Domain D also adopts β-sandwich folding, but its function is to structurally connect the ABC architecture to the F_ECBM20 domain. Both domains A and B include nine subsites (−7 to +2) that comprise the enzyme active site. Thus, the starch substrate is arranged in a ring-shaped structure at the active site of CGTases and cleavage at subsites −1 and +1 by the conserved acidic catalytic triad Asp/Glu/Asp, and AspI from CSR II, III, and IV, respectively. Simultaneously, residues at subsites +2, −2, and −3 address the four catalytic activities of CGTases described above, while residues at subsites −4 to −7 determine the CDs size specificity. Furthermore, a conserved aromatic central Tyr residue at subsite −1 is usually replaced by a nonaromatic residue in α-amylases. Notably, structure-based protein engineering has shown that mutations in the active site change the specificity, allowing the conversion of CGTases to α-amylases bacteria showed an unusual four-domain ABCECBM20 distribution with the D domain absent, and the classical F_ECBM20 domain is usually replaced in CGTases from archaea by a C-terminal F_E domain with an unclear structure-function relationship.

Carbohydrate metabolism via cyclodextrins (CM-CD) is an unusual microbial starch-converting pathway that involves synthesis, transportation, degradation, and metabolic assimilation of CDs. Notably, although the CM-CD pathway is well described for the hyperthermophilic archaea Thermococcus sp., Pyrococcus furiosus, and Archaeoglobus fulgidus, the descriptions from bacteria are limited to mesophilic G− Klebsiella oxytoca and G+ Bacillus subtilis. Extracellular CGTases are the key enzymes that catalyze the first step of the CM-CD pathway by converting the surrounding starch substrate to CDs. In G− bacteria, CDs are subsequently internalized by a periplasmic maltodextrinoporin (CDP). The entry of CDs into the cytoplasm of bacteria and archaea occurs via a type I ATP-dependent ABC sugar importer system MdxEFG-(X/MsmX)4, which internalizes both cyclic/maltodextrin molecules. Hence, sugar translocation into the cytoplasm is triggered by a dedicated MdxX ATPase in G− bacteria (CymD in K. oxytoca) or by a promiscuous MsmX ATPase in G+ bacteria. The following reaction in the CM-CD pathway is the cleavage of CDs by a cytoplasmic cyclodextrinase (CDase, EC 3.2.1.54), resulting in maltose/maltoligosaccharides that are further degraded to glucose-1-phosphate (G1P) by an α-glucan phosphorylase (GP; EC 2.4.1.1)4, finally. While glucose metabolism proceeds through the typical glycolytic pathway in K. oxytoca and B. subtilis, a modified Embden-Meyerhof-Parnas (EMP) glycolytic pathway is found in archaea. Although CM-CD is considered a secondary pathway for starch breakdown and conversion in K. oxytoca and B. subtilis, it is the main starch-converting pathway in sulfur-reducing hyperthermophilic archaea. Similarly, extremely thermophilic bacteria [T_{sp} ≥ 70°C] such as Deinococcales, Thermotogales, and Thermooceanobacteria that live in a wide variety of hot environments on Earth (e.g., hydrothermal and geothermal vents) are capable of metabolizing a broad range of carbohydrates, including starch. Nevertheless, because attention has been focused on the well-studied CGTases from Bacilli class bacteria, the identification and characterization
of CGTases from extremely thermophilic bacteria have remained vague and are limited to *Thermoanaerobacter* spp., *Carboxydocella* sp., and *Thermoanaerobacterium thermosulfurigenes*. Moreover, since the identification of CGTases for structure–function relationship studies has also been the central focus over the years, their functional role in a putative CM-CD pathway for extremely thermophilic bacteria remains obscure.

In this work, a novel group of CGTases from GH13_2 with an exceptional three-domain ABC architecture was detected by (meta)genome mining of microbial communities living in a wide variety of hot environments on Earth. Sequence analysis revealed that this group of CGTases belongs to the extremophilic *Thermoanaerobacterales* *Caldanaerobacter subterraneus* ssp., and *Thermoanaerobacter* spp. and shares ≤ 46% sequence identity with the CGTases characterized thus far. Sequence and comparative genomic analysis also showed that the three-domain ABC CGTase-encoding genes are exceptionally grouped in unrevealed gene clusters that encode the entire CM-CD pathway and several important proteins for prokaryotic cell functions. Together, functional studies of a representative member, CldA, combined with phylogenetic analysis revealed a new evolutionary path among CGTases and shed light on a nonclassical pathway for starch metabolism in *Thermoanaerobacterales*.

**Results**

**Database mining for novel thermophilic CGTase enzymes.** To identify putative CGTases involved in the CM-CD pathway of extremely thermophilic bacteria, a database mining approach was applied to ~130 public metagenomes of microbial communities from diverse thermophilic environments (Tables S1 and S2). Notably, a low number of putative CGTases were detected (14 hits in total; Table S1), which seems to be related to the rarity of the CM-CD pathway in extremely thermophilic bacteria living in starch-poor environments. Nevertheless, a CGTase-encoding gene (*cldA*) from Obsidian Pool hot spring metagenomic data at Yellowstone National Park was distinguished (Tables S1 and S2). Sequence analysis revealed that CldA consists of 524 residues and shares ≤ 42% sequence identity (100% query coverage) with the 51 characterized enzymes from GH13_2. A BLAST search in the nonredundant GenBank database revealed another three CldA-like sequences that share 98% average sequence identity with CldA (100% query coverage) and are annotated as hypothetical glycosidase/α-amylase enzymes in eight available genomes from several *Thermoanaerobacterales* subspecies of G+ thermophilic *Caldanaerobacter subterraneus* (Table S3). Although *C. subterraneus* subspecies (T<sub>opt</sub> of 60–85 °C) are found in various extremophilic environments, they natively live in the Obsidian Pool hot spring at Yellowstone National Park. Sequence analysis also revealed that CldA exhibits a 21-residue N-terminal signal peptide, MRKKNFKAFVSAALIIFSGS, which contains a positively charged tail, RKNFK, followed by a hydrophobic core region that ends with the conserved Cys22 (boldface residues) typical for the cleavage site of signal peptidase type II (SPII). In agreement with this observation, the extracellular glycoside hydrolases of the GH13 family from G+ bacteria are translocated from the cytoplasmic membrane through the general secretion (Sec) system. Because CldA and CldA-like enzymes displayed an unusual short-form sequence compared to conventional five-domain CGTases (Figs. 1A and S1), a functional domain analysis was conducted. Remark-
ably, CldA showed an atypical three-domain ABC distribution compared to CGTases with either conventional five-domain ABCDE$_{GH13}$ or four-domain ABCE$_{GH13}$ distribution (Fig. 1A). Thus, the mature form of CldA consists of catalytic AB domains (residues 22–434) and the starch-binding C domain (residues 434–524) at the C-terminal region (Figs. 1A and S1). Sequence alignment of CldA with the 51 characterized CGTases from GH13_2 revealed the presence of CSR I–VII motifs from the GH13 family (Fig. 1B), including the highly conserved catalytic triad Asp250, Glu279, and Asp351 from CSR II, III, and IV, respectively (Fig. 1B), which is involved in glycoside bond cleavage. Furthermore, both the conserved aromatic central Phe216 residue from CSR V (which is usually replaced by a nonaromatic residue in α-amylases) and the pair of hydrophobic residues Trp204/Met281, which are critical in sugar chain circularization for CD formation, were observed (Fig. 1B). Interestingly, while Met281 belongs to CSR III, Trp204 is found in a 199GSISN motif.

Functional characterization of CldA. The recombinant CldA enzyme was successfully produced in _Escherichia coli_ to evaluate CGTase activity. The mature form of CldA consists of 511 residues with a calculated molecular mass of 58.4 kDa, including a C-terminal His$_6$-tag sequence without the N-terminal signal peptide. Protein purification was performed by a heat treatment procedure and nickel-affinity chromatography followed by size-exclusion chromatography (SEC)-dynamic light scattering (DLS) coupled experiments (Fig. S2A), resulting in a purification yield of ~45 mg CldA from 1 L of culture. Purified recombinant CldA showed a molecular mass of 58.5 kDa in the SEC-DLS analysis with an optimal monodispersity (Mw/Mn = 1.02), showing that the biological assembly is monomeric (Fig. S2A). CldA also showed a molecular mass of ~58 kDa on SDS-PAGE (Fig. S2B) and a theoretical isoelectric point (pI) of 5.7. CldA displayed cyclization activity over a broad range of temperatures from 40 to 100 °C and pH ranges from 4 to 8 (Fig. 2A), using soluble starch as the substrate. Furthermore, CldA reached more than 65% relative cyclization activity at acidic pH (4–5) and high temperatures (70–90 °C) (Fig. 2A). CldA also displayed a half-life ($t_{1/2}$) of 25.5 min at 80 °C and extraordinary thermostability at 70 °C ($t_{1/2}$ = 63.4 h) (Fig. S3). CD production was monitored over time by incubating CldA with 50 g L$^{-1}$ soluble starch at 75 °C and pH 4. The production of α-, β-, and γ-CDs increased over time, achieving the maximum yield of total CDs (2.72 ± 0.06 g L$^{-1}$) after 2 h of incubation (Figs. 2B and S3). The proportion of α- and β-CDs (34:62) was relatively conserved over time with minor γ-CD production (Fig. 2B,C), revealing that the CldA enzyme is a β-CGTase. Nevertheless, while CldA displayed a specific β-cyclization activity of 51.26 ± 6.3 U mg$^{-1}$, it exhibited an unusual high hydrolytic activity of 40.40 ± 5.4 U mg$^{-1}$. According to the latter, CldA yielded as the primary products those related to the hydrolysis of soluble starch, such as linear oligosaccharides with different degrees of polymerization (G3–G7) and the fermentable sugars maltose (G2) and glucose (G1) (Figs. 2C and S4). All products synthesized by the action of CldA from soluble starch were confirmed by HPLC and mass spectrometry analysis (Figs. S4 and S5).

Discovery of a novel group of three-domain ABC CGTases. To identify additional three-domain ABC CGTases, a database mining approach was also applied to ~30 public metagenomes of microbial communities from the Obsidian Pool hot spring (Table S2), using the CldA sequence as a template. The database mining approach revealed a homologous CGTase-encoding gene (thmA) that codifies for a 526-residue CGTase (Table S3) sharing 80% sequence identity with CldA (100% query coverage) (Fig. S1). Functional domain analysis showed that ThmA is a three-domain ABC CGTase exhibiting the highly conserved Asp252/Glu281/Asp353 catalytic triad, the conserved aromatic central Phe218, and the pair of hydrophobic residues Trp206/Met283 (Fig. 1B). A BLAST search in the GenBank database of the ThmA enzyme showed 100% sequence identity with _Caldanaerobacter_ sp. strain from the _Thermoanaerobacter_ ethanolicus. Furthermore, 14 putative ThmA-like sequences...
encoded in 16 genomes from several Thermoanaerobacterales subspecies of G+ thermophilic Thermoa naerobacter spp., were also found (Table S3). A subsequent BLAST search in the GenBank database confirmed that the 19 three-domain ABC (CldA/ThmA)-like CGTases (Table S3) belong to C. subterraneus ssp. and Thermoa naerobacter spp., respectively. Furthermore, CldA and ThmA share only 38% average sequence identity with three characterized five-domain ABCDECBM20 CGTases (100% query coverage for ABC domains) from Thermoa naerobacter spp., confirming that both three-domain CldA/ThmA CGTases are not truncated forms from conventional five-domain CGTases. Accordingly, to determine the evolutionary relationship among this novel group of three-domain CGTases with all characterized CGTases from GH13_2, a phylogenetic analysis was conducted, including seven α-amylases from GH13 as an outgroup. The analysis showed that the CGTases were distributed in five phylogenetic groups that presented a bootstrap value of 100% (Fig. 3). The four-domain ABCECBM20 CGTases (blue), five-domain ABCDEarch CGTases (orange), and conventional five-domain ABCDECBM20 CGTases from the well-studied G+ Bacilli class bacteria were observed in three different clades. Nevertheless, it has been shown that the five-domain ABCDECBM20 configuration is not unique to CGTases from G+, as has been observed in the thermophilic CGTase from archaea Thermococcus sp. B1001 and the halophilic CGTase from archaea Haloferax mediterranei. A fourth clade comprises maltogenic starch-acting enzymes (blue dashed line) and α-amylases (black branch) from GH13_2 and GH13, respectively, are also shown in two different clades. Note that while the α-amylases from Aspergillus oryzae and Cordyceps farinosa belong to the GH13_1 subfamily, the α-amylases from bacteria showed an unassigned GH13 subfamily. Bootstrap values (1000 iterations) are indicated for each node. Only bootstrap values above 50% were shown. The tree was drawn using iTOL v4 (http://itol.embl.de).

**Figure 3.** Phylogenetic analysis of novel three-domain ABC CGTases. Evolutionary relationships were determined by the maximum likelihood method based on the WAG + G model using the full amino acid sequences of 78 CGTases, including the 48 characterized CGTases from GH13_2 recognized in the CAZy database, 19 three-domain ABC (CldA/ThmA)-like CGTases, and 11 putative CGTases. The sequences of 7 α-amylases from GH13 were used as an outgroup. The conventional five-domain ABCDECBM20 CGTases (blue), five-domain ABCDEarch CGTases (orange), four-domain ABCDECBM20 CGTases (red), and the novel group of 19 three-domain ABC CGTases, (CldA/ThmA)-like enzymes from thermophilic C. subterraneus ssp. and Thermoa naerobacter spp. (magenta) were observed in four different clades. The ABCDECBM20 maltogenic starch-acting enzymes (blue dashed line) and α-amylases (black branch) from GH13_2 and GH13, respectively, are also shown in two different clades. Note that while the α-amylases from Aspergillus oryzae and Cordyceps farinosa belong to the GH13_1 subfamily, the α-amylases from bacteria showed an unassigned GH13 subfamily. Bootstrap values (1000 iterations) are indicated for each node. Only bootstrap values above 50% were shown. The tree was drawn using iTOL v4 (http://itol.embl.de).
clusters where the (CldA/ThmA)-like-encoding genes are located. Strikingly, a gene cluster of 30 genes (cld) of the 1130 total gene clusters encompassing the core genome from the *Caldanaerobacter* genus was identified in the complete assembled scaffolds from *C. subterraneus* (Fig. 4, Table S4). Sequence analysis of the cld gene cluster predicts several proteins of the CM-CD pathway: a putative type I ATP-dependent ABC transporter, 

![Figure 4. Comparative view of the gene clusters involved in the CM-CD pathway.](image-url)

- **Figure 4.** Comparative view of the gene clusters involved in the CM-CD pathway. Note the genetic organization of the CM-CD gene clusters from *K. oxytoca* (cym), *Thermococcus* sp. (cgt), *B. subtilis* (cyc), *C. subterraneus* sp. (cld), *Thermoanaerobacter* spp. (thm), and *Thermoanaerobacterium* spp. (thb). Additionally, note the protein-encoding genes involved in the four steps of the CM-CD pathway. (i) Synthesis: CGTases (12), PolIIIα (13), PHP (14), and thb (15). (ii) Translocation/Internalization: MdxE (2), MdxF (3), and MdxG (4) in blue. While the MdxX (5) and CDP (6) from G− *K. oxytoca* (cym) are also blue, the putative msnX-encoding gene is not included. (iii) Degradation: CDase (7), GA (8), and GP (9) in green. (iv) Metabolic assimilation: Pgi (10), Pka (11), and PykF (12) in orange. AmyB (33) and the AmyEDC transporter system (34–36) from *Thermoanaerobacterium* spp. (thb), and the putative transcriptional regulator of the ABC transporter system from *cym/cyc* (37–38) are shown. Note the five groups of protein-encoding genes that are essential for several prokaryotic cell functions: (i) HPr (13), PolIIIα (25), and the CBS domain/Bateman module (24) for carbon catabolite regulation, bacterial genome replication, and sensing cellular energy status, metal ion concentration, and ionic strength. (ii) MurB (14), PHP (15), RapZ (16), RodZ (17), and WhiA (18) for cell wall biogenesis, sporulation, and cell division. (iii) Feruloyl esterase (22), 2-phospho-D-lactate transferase (19), the enzyme system (R)-2-hydroxyglutaryl-CoA dehydratase (20, 21), and 4-hydroxy benzoyl-CoA thioesterase (23) for oxidative stress defense, degradation of aromatic compounds, and fatty acid metabolism. (iv) Signal-transducing protein PII (26), methylenetr hdrahydrof olate reductase (29), methionine synthase (30), PepT (27) and the anaerobic transcriptional activator fnr (28) for amino acid metabolism. (v) TrnA(m^5U54)methyltransferase (31) and MATE (32) for tRNA maturation and detoxification. Genes of unknown function are in gray. Abbreviations are listed in Table S4.
putative type I ATP-dependent ABC transporter system, MdxEFG (CldEFG), and the three cytoplasmic enzymes CDase, GP, and GA. Nevertheless, while the thm gene cluster contains three-domain ABC ThmA-like CGTases, the thb gene cluster contains a conventional five-domain ABCDECGTM20 CGTase. In addition, although the Fgi-encoding gene was absent in the thb gene cluster, the critical enzymes for the EMP pathway were encoded in both the thm and thb gene clusters (Fig. 4, Table S4).

Remarkably, sequence analysis of the cld/thm/thb gene clusters also revealed the presence of 18 protein-encoding genes that are essential for prokaryotic cell functions (Fig. 4, Table S4), such as the functionally and structurally characterized phosphotransferase HPt (PDB ID: 3LE5), which is a key enzyme for carbon catabolite regulation in *C. subterraneus* ssp. *tengcongensis*72, *Thermoanaerobacter* spp.73, and *Thermoanaerobacterium* spp.74, as well as a DNA polymerase III (PolIIa, EC 2.7.7.7) responsible for bacterial genome replication75, which is preceded by a putative CBS domain/Bateman module involved in sensing cellular energy status, metal ion concentration, and ionic strength76,77. The second group of putative proteins of the cld/thm/thb gene clusters is involved in cell wall biogenesis, sporulation, and cell division: (i) UDP-N-acetylmuramate dehydrogenase (MurB, EC 1.3.1.98) is involved in the biosynthesis of bacterial cell wall peptidoglycan79, (ii) histidinol phosphatase (PHP) is required in the phosphorelay system to regulate the biosynthesis of cell wall-associated polysaccharides80, (iii) RapZ regulator is implicated in the RNA-mediated regulatory network of glucosamine biogenesis81, (iv) the transmembrane RodZ protein is a key protein in cell elongation (elongasome) and cell division82,83, and (v) the sporulation transcription WhiA regulates cell differentiation84,85. The third group of proteins is essential for oxidative stress defense, degradation of aromatic compounds, and fatty acid metabolism: (i) the functionally characterized feruloyl esterase (EC 3.1.1.73) from *C. subterraneus* ssp. *tengcongensis*, which can hydrolyze esterified phenolic acids from xylan and pectin86, (ii) 2-phospho-l-lactate transferase (EC 2.7.8.28) involved in the biosynthesis of redox coenzyme F420 which is important for the redox transformations of cell wall lipids, degradation of aromatic xenobiotic compounds, and neutralization of oxidative and nitrosative stress87,88, (iii) the two components E1 (activator) and E2 (dehydratase) of the enzyme system (R)-2-hydroxyglutaryl-CoA dehydratase (EC 4.2.1.167), which is involved in glutamate metabolism via butyrate fermentation in G+ bacteria89, and (iv) putative 4-hydroxy benzoyl-CoA thioesterase, which can hydrolyze fatty acyl-CoA thioesters90. The fourth group of putative proteins is implicated in amino acid metabolism: (i) signal-transducing protein PII involved in the regulation of nitrogen metabolism via glutamine/glutamate cycle91, (ii) methylenetetrahydrofolate reductase (EC 2.1.1.13), and methionine synthase (EC 1.5.1.20), which are both involved in methionine biosynthesis via methylenetetrahydrofolate (methyl-THF), and (iii) treptide aminopeptidase T (PepT; EC 3.4.11.4), which is preceded by its anaerobic transcriptional activator fnr91 and is only included in the cld/thm gene clusters. Finally, the putative tRNA(m5U54)methyltransferase (EC 2.1.1.190) and a multiantimicrobial extrusion protein (MATE), which might be involved in tRNA maturation and detoxification, respectively92,93, are also encoded in the cld/thm/thb gene clusters. Although G− *K. oxytoca*, archaea *Thermococcus* sp., and G+ *B. subtilis* arranged the proteins involved in the CM-CD pathway in three similar gene clusters, *cym*, *cgt*, and *cyc*, respectively (Fig. 4), none of the latter protein-encoding genes for prokaryotic cell functions and the proteins for the EMP pathway are encoded near their CM-CD gene clusters. The proteins encoded in the cld/thm gene clusters (Fig. 4) are shown in Table S4.

**Discussion**

Traditionally, the five-domain ABCDECGTM20 organization has been considered the central architecture of CGTases, with the only few exceptions for five-domain ABCDEarch CGTases from archaea and four-domain ABCE CGTases from G−, highlighting the recurrence of both the ABC core structure and the ECBM20/Earch domain in the overall CGTase fold. Here, a database mining approach allowed the identification of a novel group of three-domain ABC (CldA/ThmA)-like CGTases from G− thermophilic sp., respectively, which exhibit a unique CGTase domain distribution that is different from that seen in all other CGTases characterized thus far (Fig. 1A). Notably, although the (CldA/ThmA)-like enzymes displayed a distinctive active site for CGTases with the presence of all CSR I-VII motifs from the GH13 family (Fig. 1B), other CGTases characterized thus far differ in their overall CGTase fold. Here, a database mining approach allowed the identification of a novel group of three-domain ABC (CldA/ThmA)-like CGTases from G− thermophilic sp., respectively, which exhibit a unique CGTase domain distribution that is different from that seen in all other CGTases characterized thus far (Fig. 1A). Notably, although the (CldA/ThmA)-like enzymes displayed a distinctive active site for CGTases with the presence of all CSR I-VII motifs from the GH13 family (Fig. 1B), the three-domain ABC architecture is not commonly associated with conventional CGTases. The functional characterization of a representative member, the three-domain ABC CldA, revealed that regardless of whether β-CD is synthesized as the major cyclization product from the starch substrate under the assay conditions, cyclization does not appear to be the main activity of the enzyme (Fig. 2). Accordingly, the production of fermentable sugars, dextrins, and functional CDs from the starch substrate by the action of extracellular (CldA/ThmA)-like CGTases seems to be a reasonable adaptation to diversify products and increase the probability of survival in extremely hot environments with low starch and nutrient concentrations.

The identification of this novel group of enzymes showed for the first time that the three-domain ABC organization represents the minimal functional core structure for CGTases and confirmed previous studies suggesting that the C-terminal region of CGTases has been acquired through evolutionary processes95,96. Indeed, while the raw starch-binding ECBM20 domain is observed in several GH families97,98, both the Ecore domain with an unknown structure–function relationship and the connecting D domain are unique to CGTases95. Interestingly, the three-domain CGTases clustered together in a new monophyletic group that diverged as a novel evolutionary path among conventional CGTases. Hence, while the four-domain CGTases from G− separated early from the rest of CGTases, the three-domain CGTases and both groups of five-domain CGTases diverged later from a common ancestor. This observation also indicates that three-domain CGTases are not truncated forms from either of the two groups of five-domain CGTases, and the minimal ABC framework of the (CldA/ThmA)-like enzymes from Thermoanaerobacteriales is not the common ancestor of all CGTases (Fig. 3).
In addition to the phylogenetic analysis, the presence of this novel group of three-domain CGTases suggests a role in starch metabolism. Nevertheless, Thermoaerobacteriales are obligate anaerobic Clostridia class bacteria with low genomic G+C content capable of thriving in various hot environments on Earth, such as geothermal fields, submarine hydrothermal vents, and oil reservoirs, which are expected to be starch-poor environments.

Consequently, genomic gene clustering analysis against 246 Thermoaerobacteriales genomes allowed the identification of only three gene clusters involved in the CM-CD pathway, \( \text{cld} \), \( \text{thm} \), and \( \text{thb} \), from the Thermoaerobacteracea family (\( \text{C. subterraneus} \) ssp. and \( \text{Thermoanaerobacter} \) spp.) and from Thermoaerobacteriales family III (\( \text{Thermoanaerobacterium} \) spp.), respectively, confirming the rarity of the pathway. Thus, while the three-domain (CldA/ThmA)-like-encoding genes belong to the \( \text{cld} \) and \( \text{thm} \) gene clusters, respectively, the \( \text{thb} \) gene cluster contains a conventional five-domain CGTase-encoding gene (Fig. 4). Based on comparisons with \( \text{G}^- \text{K. oxytoca} \), archaea Thermococcus spp., and \( \text{G}^+ \text{B. subtilis} \), which arranged the proteins involved in the CM-CD pathway in three similar gene clusters, cym, cgt, and cyc, respectively (Fig. 4), the first step of the CM-CD pathway in Thermoaerobacteriales involves converting the surrounding starch substrate to CDs catalyzed by secreted three- and five-domain CGTases (Fig. 5). As previously established by X-ray crystallography studies, the resulting CDs are then internalized into the periplasm by a transmembrane \( \beta \)-barrel CDP in \( \text{G}^- \text{K. oxytoca} \) (CymA, PDB ID: 4V3G), which mediates the passive diffusion of CDs through the perturbation of electrostatic interactions of the N-terminal region with the \( \beta \)-barrel wall of CDP. Therefore, the 15 N-terminal residues of CymA are expelled from the barrel through a ligand-expelled gate mechanism, allowing the diffusion of CDs into the periplasmic...
cluster101. However, the database mining analysis revealed that the AmyB-like enzyme from ABC transporter systems from G+ bacteria. Notably, additional data mining analysis revealed that MalEFG, from G+, occurs through a conformational change of the two permease subunits MdxFG triggered by the ATPase activity of MdxX/MsmX. Accordingly, the MdxEFG-X transporter system from G− of CDs into the cytoplasm through the EMP pathway (Fig. 5). Thus, while the linearization of CDs by CDase produces G1 and G2 molecules for the EMP pathway, dextrins (Gn>3) are either converted into G1 or G1P (with features that distinguish sugar-binding proteins from type I ATP-dependent ABC transporter systems). Thus, a transmembrane CDP in G+ bacteria, differences between MdxE proteins from G+ and G− bacteria are typical of CDs into the cytoplasm (Fig. 5). As expected, the putative type I ATP-dependent ABC transporter system, CldEFG-MsmX, from the Thermoanaerobacterales division (Fig. 4, Table S4), suggesting that extracellular heat-resistant CGTases could play a leading role in the synthesis of CDs might have a physiological role as functional amphipathic toroids. Interestingly, the entire CM-CD pathway is encoded along with several essential proteins for G+ cell functions, such as DNA replication, carbon catabolite regulation, tRNA maturation, biogenesis, sporulation, and cell division (Fig. 4, Table S4), suggesting that extracellular heat-resistant CGTases could play a leading role in the metabolism of Thermoanaerobacteriales. Moreover, the presence of protein-encoding genes related to extreme thermophilic metabolism, such as oxidative stress defense, degradation of aromatic compounds, fermentation, and fatty acid and amino acid metabolism (Fig. 4, Table S4), also indicates that the physiological role of heat-resistant CGTases in product diversification seems to be a convergent adaptation to survive in hot starch-poor environments. Accordingly, the relevance of CGTases during starch metabolism can be supported by early observations of Thermoanaerobacteriales spp. in which the secreted thermophilic α-amylase/amylolypullulanase AmyB was found to hydrolyze a variety of α-1,4- and α-1,6-glucans acting together with an ABC maltose/maltotriose importer (amyECD). Notably, the amyBEDC gene cluster is located immediately upstream of the conventional five-domain CGTase (formerly named AmyA) encoded in the ABEDC gene cluster (Fig. 4). Accordingly, the relevance of CGTases during starch metabolism can be supported by early observations of Thermoanaerobacteriales spp. in which the secreted thermophilic α-amylase/amylolypullulanase AmyB was found to hydrolyze a variety of α-1,4- and α-1,6-glucans acting together with an ABC maltose/maltotriose importer (amyECD). Notably, the amyBEDC gene cluster is located immediately upstream of the conventional five-domain CGTase (formerly named AmyA) encoded in the ABEDC gene cluster (Fig. 4). Thus, amyB and the five-domain CGTase seem to play a cooperative role, as it has been shown that the transcription of the amyBEDC gene cluster and the CGTase-encoding gene is induced by maltose or starch as carbon sources. Likewise, the deduced promoter sequences of ABEDC genes, 5′-TGGACT-17 bp-TAATAAT, and 5′-TTTCCA-17 bp-CATATT, showed similarity to the σ-dependent consensus promoters of the amyBEDC gene cluster. However, the database mining analysis revealed that the AmyB-like enzyme from C. subterraneus ssp.
and *Thermoanaerobacter* spp. is not encoded near the *cld/thm* gene clusters (Fig. 4), indicating that the secreted three-domain CGTases are the main starch-acting enzymes of both gene clusters, highlighting their importance for product diversification on these microorganisms.

In summary, this is the first identification of a novel group of CGTases with an uncommon three-domain ABC organization, which further established a new evolutionary path among CGTases. These novel enzymes were detected in two gene clusters, *cld* and *thm*, from extremely thermophilic *Thermoanaerobacter*ales *C. subterraneus* spp. and *Thermoanaerobacter* spp., as part of a CM-CD pathway involved in the synthesis, transportation, degradation, and metabolic assimilation of CDs from starch. These findings were extended to *Thermoanaerobacter*ales *Thermoanaerobacterium* spp., which also showed a CM-CD pathway not previously described but governed by a conventional five-domain CGTase encoded in the *thb* gene cluster. In contrast to the secondary role of the CM-CD pathway in mesophilic bacteria, the remarkable product diversification catalyzed by the three-domain CGTases suggests that they could play a critical role in the carbohydrate metabolism of *C. subterraneus* spp. and *Thermoanaerobacter* spp. Future X-ray crystal structure determination, structure-based protein engineering, and kinetic studies of ClDa will offer an opportunity to gain insights into this particular pathway and the structure–function relationship of this novel group of enzymes.

**Materials and methods**

**Data mining for CGTases.** Metagenomes were analyzed from the Joint Genome Institute (JGI) IMG/M database103, which contains more than 15,014 metagenomes from different environments (last search, July 2021). Putative CGTases were detected by a BLASTn search in ~ 130 publicly assembled metagenomes in the IMG/M platform using an E-value cutoff of 1.0e-5. The metagenomes were filtered for those containing different terms from hyperthermophilic ecological niches (e.g., geothermal fumarole, geyser, hot spring, or hydrothermal vent) in the "Genome Name/Sample Name" description (Table S1). The protein query sequences consisted of the complete amino acid sequences of experimentally characterized CGTases, including CGTase from *G. theta* (NCBI ID: AAA25059.1)104, a CGTase from the thermophilic archaea *P. furiosus* distribution (NCBI ID: AAA25059.1)104, and a CGTase from the thermophilic archaea *P. furiosus* (NCBI ID: ABA33720.1)105. Putative CGTases that shared > 45% sequence identity with query sequences were excluded to increase novelty. The best hits were analyzed manually to evaluate the complete scaffold templates and discard truncated sequences. NCBI's Batch Web CD-Search Tool against the Conserved Domain Database (CDD/SPARCLE)106 was employed to predict the functional domains of selected hits. Hence, a putative CGTase with a unique three-domain ABC distribution (named ClDa) was identified in a scaffold containing ~ 50 genes in a metagenome of thermophilic microbial communities from Obsidian Pool hot spring at Yellowstone National Park (Wyoming, USA) (Table S1). Therefore, a second database mining approach was applied to identify additional three-domain ABC CGTases. The ClDa sequence was then submitted to BLASTn against 30 publicly assembled metagenomes deposited in the IMG/M platform103 that belong to several microbial communities from the Obsidian Pool hot spring at Yellowstone National Park (Table S2). A second putative three-domain ABC CGTase (named ThmA) was identified in three metagenomes from the Obsidian Pool hot spring (Table S2). Redundant sequences and truncated genes were discarded. The ClDa/ThmA sequences, along with the 51 sequences of characterized enzymes from GH13 and GH14 families, were used as queries in a FASTA file and subjected to multiple alignments using Clustal Omega with default parameters107. Manual refinement of the multiple alignments was performed to detect key conserved catalytic residues from CGTases108. Finally, a third database mining approach was conducted to identify additional (ClDa/ThmA)-like CGTases. Hence, the ClDa/ThmA sequences were submitted to BLASTn against publicly assembled genomes deposited in the GenBank database from *Caldanaerobacter* spp. (NCBI Taxonomy ID: 2149539) and *Thermoanaerobacter* spp. (NCBI Taxonomy ID: 68295). Several (ClDa/ThmA)-like sequences were obtained (Table S3), listed in a FASTA file, and subjected to the bioinformatics pipeline described above. The sequence logo was generated by WebLogo109. The three-domain ABC CGTase ClDa was selected for further recombinant production and functional studies.

**Gene cloning and protein production.** A synthetic gene coding for the mature form of ClDa, codon-optimized for *E. coli* expression, was prepared by Integrated DNA Technologies (Iowa, USA). The synthetic cldA gene was cloned into the NdeI and NotI sites of the pET-22b(+)-expression vector (Novagen), which contains a sequence coding for six histidines at the C-terminus. The identity of the resulting plasmid pCldA was evaluated by restriction analysis and confirmed by DNA sequencing. Electroporantion of *E. coli* BL21(DE3)pLysS cells were transformed with pCldA and grown on Luria–Bertani (LB) agar plates containing 100 μg mL⁻¹ ampicillin at 37 °C. A single colony of BL21(DE3)pLysS/pCldA was picked to inoculate 5 mL LB medium overnight with 100 μg mL⁻¹ ampicillin at 37 °C, aliquoted in a sterile solution of 40% (v/v) glycerol and maintained at − 80 °C. For recombinant ClDa production, a fraction of a frozen cell aliquot was taken and cultured for 12 h at 37 °C and 200 rev min⁻¹ in 50 mL LB medium containing 200 μg mL⁻¹ ampicillin. This preinoculum was used to inoculate 1 L 2xYT medium with 200 μg mL⁻¹ ampicillin at an initial optical density at 600 nm (OD₆₀₀) of 0.05 at 37 °C and 200 rev min⁻¹. After induction by adding a final concentration of 0.1 mM IPTG to the medium (OD₆₀₀ of ~ 0.6), the temperature was lowered to 22 °C, and the culture was grown for 12 h at 150 rev min⁻¹. The cells were harvested by centrifugation (7500g, 10 min, 4 °C) and resuspended in 10 mL buffer A [50 mM sodium phosphate pH 8.0, 500 mM NaCl, 2% (v/v) glycerol, 20 mM imidazole] containing EDTA-free complete protease inhibitor cocktail mini tablet (Roche Molecular Biochemicals) and 1 μg mL⁻¹ DNase. The cell suspension was sonicated on ice for 30 min with an amplitude of 25–29%, and the resulting solution was heated for 20 min at 60 °C to precipitate the thermolabile protein fraction of *E. coli*. After the heating step, the lysate was centrifuged (19,000g, 45 min, 4 °C), and the supernatant containing recombinant thermophilic His₆-tagged ClDa was recovered.
Protein purification and SEC-DLS analysis. The supernatant containing CldA was filtered with a 0.22 µm pore filter and applied onto a 5 mL Ni²⁺-chelating HisTrap HP column (GE Healthcare) equilibrated with ten-bed volumes of buffer A using an ÄKTA Pure 25 M1 FPLC system with UNICORN software (GE Healthcare). The column was then washed with eight-bed volumes of buffer A to remove contaminants. Bound CldA enzyme was eluted with a linear gradient of 20–500 mM imidazole using buffer B [50 mM sodium phosphate pH 8.0, 500 mM NaCl, 2% (v/v) glycerol, 500 mM imidazole] at a flow rate of 5 mL min⁻¹ and analyzed by SDS-PAGE with Coomassie staining. A single peak at ~300 mM imidazole containing the CldA enzyme was collected, concentrated, and dialyzed against several volumes of buffer C [50 mM Tris–HCl pH 7.5, 100 mM NaCl] in an ultrafiltration cell (Amicon Ultracel filter, 30 kDa molecular-weight cutoff). The SEC-DLS analysis was performed using an ÄKTA Pure 25 M1 FPLC system coupled to a dynamic light scattering (DLS) detector using a Malvern Zetasizer µV DLS instrument. A concentrated sample of CldA at 10 mg mL⁻¹ was filtrated with a 0.22 µm pore filter and loaded onto a 120 mL HiLoad 16/600 Superdex 75 pg column (GE Healthcare) equilibrated with buffer C. CldA was then eluted with the same buffer in an SEC-DLS coupled experiment using a quartz flow cell of 8 µL (Malvern) at a flow rate of 1.0 mL min⁻¹. The SEC-DLS system was previously calibrated with a standard of bovine serum albumin (BSA, Sigma-Aldrich) at 17 mg mL⁻¹ in buffer C. Data acquisition and analysis of SEC-DLS measurements were carried out using the OmniSEC 5.12 software (Malvern). A highly purified and monodisperse peak corresponding to the CldA monomer (58.5 kDa) was collected, concentrated, and dialyzed against several volumes of buffer D [50 mM Tris–HCl pH 7.5] using a 30 kDa cutoff ultrafiltration cell for enzyme activity assays. Protein concentration was determined by the Bradford assay using BSA as a standard.

Enzyme activity assay. The reaction mixture (1 mL) at 75 °C consisted of 50 mM sodium acetate pH 4.0, 5% (w/v) soluble starch (Sigma-Aldrich, Product Number: S9765), 10 mM CaCl₂, and 1 µg (1.71e⁻⁵ µmol) purified CldA enzyme. The initial rates were measured using a 96-well microplate reader (Multiskan Sky Microplate Spectrophotometer, Thermo Fisher Scientific). The temperature dependence of β-CGTase activity was determined in the 40–100 °C range. The optimum pH was determined by incubating the enzyme in different 50 mM buffer solutions ranging from pH 3.0 to 9.0. Hence, glycine–HCl buffer was used at pH 3.0, acetate buffer at pH 4.0 to 5.0, phosphate buffer at pH 6.0 to 7.0, Tris–HCl buffer at pH 8.0, and glycine–NaOH buffer at pH 9.0. The β-CGTase activity was determined spectrophotometrically by the phenolphthalein method described elsewhere with minor modifications. Accordingly, 250 mL of working phenolphthalein solution was prepared by adding ~249 mL of 125 mM sodium carbonate pH 10.5 to 1 mL of 3 mM phenolphthalein solution in ethanol. The reaction was stopped by adding 175 µL of 1 mM NaOH to 50 µL aliquots of the reaction mixture. The latter solution was then mixed and vortexed with 100 µL of working phenolphthalein solution and analyzed by the decrease in absorbance at λ = 550 nm owing to β-CD-phenolphthalein complex formation. The β-CD concentration was determined in the 0.1–2 µM range by the ε₅₀₅₀ = 4040 (ε₅₀₅₀ = molar absorptivity at 550 nm) (Sigma-Aldrich). One unit of β-CGTase activity was defined as the amount of enzyme that produced 1 µmol β-CD per min under the defined conditions. The hydrolytic activity was measured as the liberation of reducing sugars from soluble starch by the 3,5-dinitrosalicylic acid (DNS) method using a standard curve constructed with commercial maltose (Sigma-Aldrich). One unit of hydrolytic activity was defined as the amount of enzyme that produced 1 µmol of reducing sugars per min under the defined conditions.

Product analysis. The formation of CDs from starch was determined by incubating 1.71e⁻² µM CldA with 5% (w/v) soluble starch in 50 mM sodium acetate pH 4.0 and 10 mM CaCl₂ at 75 °C for 4 h. Aliquots of 1 mL were taken at regular intervals, and the reactions were stopped by adding two volumes of cold HPLC water. The mixtures were centrifuged (16,000g, 15 min) and filtered through a 0.22 µm polyvinylidene difluoride (PVDF) membrane filter (GVs Life Sciences). Products formed were analyzed using a Waters Alliance HPLC system (Model e2695 Separations Module, USA) employing a Waters XBridge BEH amide column (5 µm, 150 mm × 4.6 mm) heated at 30 °C. Samples were processed at an eluent of acetonitrile:water (65:35, v/v) with a flow rate of 0.5 mL min⁻¹ using a refractive index (RI) detector (Model 2414, Waters) also heated at 30 °C. Data acquisition and treatment were performed with Empower software v.2.0 (Waters). Mass spectrometry analysis of products from 5% (w/v) soluble starch by the action of CldA was obtained from a mixture at 2 h using a QTOF Xevo G2-S (Waters). A direct infusion into the mass spectrometer was used at a flow rate of 5 µL min⁻¹. The ionization conditions were as follows: (i) the electrospray source was operated in positive ion mode, and the source and desolvation temperatures were 100 and 250 °C, respectively; (ii) desolvation and cone gas at a flow rate of 800 and 50 L h⁻¹, respectively; (iii) capillary and cone voltage of 2500 and 10 V, respectively; (iv) acquisition mass range from 50 to 1500 m/z. For HPLC and mass spectrometry analysis, high-purity oligosaccharides from G3 to G7 (Toronto Research Chemical) and G1-G2, α-, β-, and γ-CDs (Sigma-Aldrich) were used as standards.

Phylogenetic analyses. The phylogenetic tree was based on the alignment of 78 amino acid sequences of several CGTases, including the 48 characterized CGTases from GH13_2 recognized in the CAZy database, 19 three-domain ABC (CldA/ThmA)-like CGTases, and 11 putative CGTases (NCBI ID: SP250858.1, WP_102522592.1, WP_077677735.1, WP_013906564.1, WP_048146067.1, WP_048164181.1, WP_071127407.1, WP_115251089.1, WP_078681113.1, WP_069649891.1, and WP_069657150.1) to improve the fit of some clades. The sequences of 7 α-amylases from GH13 were used as an outgroup. Three starch-acting enzymes from WP_115251089.1, WP_078681113.1, WP_013906564.1, WP_048146067.1, WP_048164181.1, WP_071127407.1, WP_102522592.1, WP_077677735.1, WP_013906564.1, WP_048146067.1, WP_048164181.1, WP_071127407.1, WP_115251089.1, WP_078681113.1, WP_069649891.1, and WP_069657150.1) to improve the fit of some clades. The phylogenetic tree was constructed using MEGA X and visualized in Fig. 3. The sequences of seven α-amylases from GH13 were used as an outgroup. Three starch-acting enzymes from WP_115251089.1, WP_078681113.1, WP_013906564.1, WP_048146067.1, WP_048164181.1, WP_071127407.1, WP_102522592.1, WP_077677735.1, WP_013906564.1, WP_048146067.1, WP_048164181.1, WP_071127407.1, WP_115251089.1, WP_078681113.1, WP_069649891.1, and WP_069657150.1) were excluded from the analysis since they are not CGTases. Two phylogenetic trees were built using the full amino acid sequence (Fig. 3) or solely the amino acid sequence of the minimal functional core ABC (Fig. S6) for all 85 sequences mentioned above. The alignment of all amino acid sequences was conducted with the ClustalW algorithm using default parameters. The
evolutionary relationship of CGTases was inferred with the maximum likelihood method\textsuperscript{111}, setting the best-fit model of amino acid substitution (WAG + G)\textsuperscript{112}. The bootstrap method (1000 replicates) was applied to assess the confidence in the phylogenetic analysis. All the implemented algorithms are included in the Molecular Evolutionary Genetics Analysis (MEGA 6.06) package\textsuperscript{115}. The consensus tree was visualized and edited in Interactive Tree Of Life ITOL v4 (http://itol.embl.de)\textsuperscript{113}.

Data mining for CM-CD gene clusters. The cld gene clusters where the cldA-like-encoding genes are located were delimited in the complete assembled scaffold from \textit{C. subterraneus} spp. \textit{ponsicinius} KB-1 (NCBI ID: AXXD01000002, location 50928–86345), \textit{C. subterraneus} spp. \textit{ponsicinius} 38_43 (NCBI ID: LGEY10000002, location 21575–56994) and \textit{T. tengcongensis} M84 (NCBI ID: AE008691.1, location 1749287–1786305). Partial scaffolds of the cld gene clusters were also found in five other genomes from \textit{C. subterraneus} spp. (Tables S3, S4). Furthermore, the cld gene clusters involved in the CM-CD pathway were submitted to BLASTn against 246 available genomes deposited in the GenBank database from Thermoanaerobacteriales order (NCBI Taxonomy ID: 68295). Accordingly, the genomes from \textit{Carboxythermus} (NCBI Taxonomy ID: 129957), \textit{Thermacetogenium} (NCBI Taxonomy ID: 140458), \textit{Gelria} (NCBI Taxonomy ID: 189326), \textit{Desulfogulva} (NCBI Taxonomy ID: 418453), \textit{Tepidanaerobacter} (NCBI Taxonomy ID: 499228), \textit{Fervidocarboxybus} (NCBI Taxonomy ID: 555078), \textit{Caldanaerobius} (NCBI Taxonomy ID: 862261), \textit{Broeckia} (NCBI Taxonomy ID: 1648500), \textit{Calorimonas} (NCBI Taxonomy ID: 2606906), \textit{Thermodesulfimonas} (NCBI Taxonomy ID: 1914252), and \textit{Moorella} (NCBI Taxonomy ID: 44260), as well as \textit{Thermodesulfobacteria} \textit{Thermodesulfobium} (NCBI Taxonomy ID: 227388), Thermoanaerobacteriales family III \textit{Anaeroecellum} (NCBI Taxonomy ID: 339355), \textit{Caldicellulosiruptor} (NCBI Taxonomy ID: 44000), \textit{Thermovenabulum} (NCBI Taxonomy ID: 159730), \textit{Thermoselimonibacter} (NCBI Taxonomy ID: 291988), \textit{Caldanaerovirga} (NCBI Taxonomy ID: 591374), \textit{Syntrophacetica} (NCBI Taxonomy ID: 862071), and \textit{Thermoanaerobacteriales} family IV \textit{Methylha} (NCBI Taxonomy ID: 252965) were analyzed for CM-CD gene clusters. Similarly, both \textit{thm} and \textit{thb} gene clusters were detected using an expanded searching cross-families algorithm in the Pathways Resource Integration Center (PATRIC v3.6.8) database\textsuperscript{114}. Hence, the \textit{thm} gene cluster was detected in the complete scaffold of several species from \textit{Thermoanaerobacter}, including \textit{Thermoanaerobacter pseudethanolicus} ATCC 33232 (NCBI ID: CP000924), \textit{Thermoanaerobacter indiensis} BSB-33 (NCBI ID: ARDJ00000000), \textit{Thermoanaerobacter brockii} spp. \textit{fumii} Ako-1 (NCBI ID: CP002466), and \textit{Thermoanaerobacter} sp. strains X513 (NCBI ID: CP002210), X514 (NCBI ID: CP000923), UBA8867 (NCBI ID: DOPY00000000), and X561 (NCBI ID: ACXP00000000). Meanwhile, the \textit{thb} gene cluster was detected in the complete assembled scaffold from \textit{Thermoanaerobacterium aoteaoreano} SCUT27 (NCBI ID: AYSN00000000), \textit{Thermoanaerobacterium saccharolyticum} JW/SL-YS485 (NCBI ID: CP003184), \textit{Thermoanaerobacterium xylopyricum} LX-11 (NCBI ID: CP002739), and \textit{Thermoanaerobacterium thermosaccharolyticum} DSM 571 (NCBI ID: CP002171). Functional comparisons of the \textit{clm}, \textit{thm}, and \textit{thb} gene clusters with the \textit{cym}, \textit{cpt}, and \textit{cyc} gene clusters involved in the CM-CD pathway from \textit{K. oxytoca} M5a1 (NCBI ID: CP020657; location 664764–678182), \textit{Thermococcus} sp. B1001 (NCBI ID: AB034969.2) and \textit{B. subtillis} (NCBI ID: CP011534; location 335514–3365346), respectively, were performed using the PATRIC genus-specific protein families (PLFams) method\textsuperscript{114}. Functional prediction of proteins encoded by the \textit{clm}, \textit{thm}, and \textit{thb} gene clusters (Table S4) was carried out using the CDD/SPARCLE\textsuperscript{106}, Pfam\textsuperscript{115}, and UniProt (https://www.uniprot.org/) databases. Protein subcellular localization and physicochemical property predictions were conducted using the CELLO\textsuperscript{116} and ProtParam (ExPASy) servers\textsuperscript{117}, respectively. The presence of a signal peptide was performed using the SignalP 5.0 server\textsuperscript{118}. Metabolic pathway analysis was conducted using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database\textsuperscript{119}.

Received: 13 August 2021; Accepted: 20 December 2021
Published online: 14 January 2022

References
1. Cruz, G. Review: A history of cyclodextrins. Chem. Rev. 114, 10940–10975 (2014).
2. Jambhekar, S. S. & Breen, P. Cyclodextrins in pharmaceutical formulations I: Structure and physicochemical properties, formation of complexes, and types of complex. Drug Discov. Today 21, 356–362 (2016).
3. Strompen, S., Miranda-Molina, A., López-Munguía, A., Castillo, E. & Saab-Rincón, G. Acceptor-induced modification of regioslectivity in CGTase-catalyzed glycosylations of β-nitrophenyl-glucopyranosides. Carbohydr. Res. 404, 46–54 (2015).
4. Jones, S. T. et al. Modified cyclodextrins as broad-spectrum antivirals. Sci. Adv. 6, eaaz9318 (2020).
5. Fan, J., Zhang, Y., Jiang, J., Lei, J. & Xue, H. Beta-cyclodextrin-functionalized Gd nanorods as building modules for ultrasensitive photoelectrochemical biosay of HIV DNA. Biosens. Bioelectron. 142, 1–6 (2019).
6. Mouset, E. et al. Influence of solubilizing agents (cyclodextrin or surfactant) on phenanthrene degradation by electro-fenton process—Study of soil washing recycling possibilities and environmental impact. Water Res. 48, 306–316 (2014).
7. Yu, G., Jie, K. & Huang, F. Supramolecular amphiphiles based on host-guest molecular recognition motifs. Chem. Rev. 115, 7240–7303 (2015).
8. Shishido, T. K. et al. Antifungal activity improved by coproduction of cyclodextrins and anaebenolyisins in Cyanobacteria. Proc. Natl. Acad. Sci. USA. 112, 13669–13674 (2015).
9. Wen, Y. et al. A supramolecular platform for controlling and optimizing molecular architectures of siRNA targeted delivery vehicles. Sci. Adv. 6, EABC2148 (2020).
10. Bardi, L., Mattei, A., Stefan, S. & Marzona, M. Hydrocarbon degradation by a soil microbial population with β-cyclodextrin as surfactant to enhance bioavailability. Enzyme Microb. Technol. 27, 709–713 (2000).
11. Trotta, F., Zanetti, M. & Camino, G. Thermal degradation of cyclodextrins. Polym. Degrad. Stab. 69, 373–379 (2000).
12. Kohata, S., Iyodai, K. & Ohyoshi, A. Thermal decomposition of cyclodextrins (α-, β-, γ-, and modified β-CyD) and of metal—(β-CyD) complexes in the solid phase. Thermochim. Acta 217, 197–198 (1992).
13. Uitdehaag, J. C., Kalk, K. H., van der Veen, B. A., Dijkstra, L. & Dijksta, B. W. The cyclization mechanism of cyclodextrin glycosyltransferase (CGTase) as revealed by a gamma-cyclodextrin-CGTase complex at 1.8-Å resolution. J. Biol. Chem. 274, 34868–34876 (1999).

14. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 42, 490–495 (2014).

15. Stam, M. R., Danchin, E. G. J., Rancurel, C., Coutinho, P. M. & Henrissat, B. Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of α-amylase-related proteins. Protein Eng. Des. Sel. 19, 555–562 (2006).

16. Janecek, S., Svensson, B. & MacGregor, E. A. α-Amylase: An enzyme specificity found in various families of glycoside hydrolases. Cell. Mol. Life Sci. 71, 1149–1170 (2014).

17. Meng, X. et al. Structure–function relationships of family GH70 glucanase and 4-α-glucanotransferase enzymes, and their evolutionary relationships with family GH13 enzymes. Cell. Mol. Life Sci. 73, 2681–2706 (2016).

18. Uitdehaag, J. C. et al. X-ray structures along the reaction pathway of cyclodextrin glycosyltransferase elucidate catalysis in the α-amylase family. Nat. Struct. Biol. 6, 432–436 (1999).

19. Janecek, S., MacGregor, E. A. & Svensson, B. Characteristic differences in the primary structure allow discrimination of cyclodextrin glycosyltransferases from α-amylases. Biochem. J. 305, 685–686 (1995).

20. van der Veen, B. A., van Alebeek, G. J., Uitdehaag, J. C., Dijkstra, B. W. & Dijksta, B. L. The three glycosylation reactions catalyzed by cyclodextrin glycosyltransferase from Bacillus circulans (strain 251) proceed via different kinetic mechanisms. Eur. J. Biochem. 267, 658–665 (2000).

21. Knegtel, R. M. A., Wind, D. & Dijksta, B. W. Crystal structure at 2.3 Å resolution and revised nucleotide sequence of the thermostable cyclodextrin glycosyltransferase from Thermanaerobacterium thermosulfurigenes EML. J. Mol. Biol. 256, 611–622 (1996).

22. Penninga, D. et al. The raw starch binding domain of cyclodextrin glycosyltransferase from Bacillus circulans strain 251. J. Biol. Chem. 271, 32777–32784 (1996).

23. Janeček, Š., Mareček, F., MacGregor, E. A. & Svensson, B. Starch-binding domains as CBM families-history, occurrence, structure, function and evolution. Biotechnol. Adv. 37, 107451 (2019).

24. Christiansen, C. et al. The carbohydrate-bridging module family 20—Diversity, structure, and function. FEBS J. 276, 5006–5029 (2009).

25. Rimpanitchayakit, V., Tonozuka, T. & Sakano, Y. Construction of chimeric cyclodextrin glucanotransferases from Bacillus circulans A11 and Paenibacillus macerans IAM1243 and analysis of their product specificity. Carbohydr. Res. 340, 2279–2289 (2005).

26. Ohdan, K., Kuriki, T., Takata, H., Kaneko, H. & Okada, S. Introduction of raw starch-binding domains into Bacillus subtilis α-amylase by fusion with the starch-binding domain of Bacillus cyclomaltodextrin glucanotransferase. Appl. Environ. Microbiol. 66, 3058–3064 (2000).

27. Strokopytov, B. et al. Structure of cyclodextrin glycosyltransferase complexed with a maltononaose inhibitor at 2.6 Å resolution. Implications for product specificity. Biochemistry 35, 4241–4249 (1996).

28. van der Veen, B. A., Uitdehaag, J. C. M., Dijkstra, B. W. & Dijkhuizen, L. Engineering of cyclodextrin glycosyltransferase reaction and product specificity. Biochem. Biophys. Acta. 1543, 336–360 (2000).

29. van der Veen, B. A. et al. Hydrophobic amino acid residues in the acceptor binding site are main determinants for reaction mechanism and specificity of cyclodextrin-glycosyltransferase. J. Biol. Chem. 276, 44557–44562 (2001).

30. Leemhuis, H., Dijkstra, B. & Dijkhuizen, L. Mutations converting cyclodextrin glycosyltransferase from a transglycosylase into a starch hydrolase. FEBS Lett. 514, 189–192 (2002).

31. Yamamoto, T. et al. Alteration of product specificity of cyclodextrin glucanotransferase from Thermococcus sp. BBO1 by site-directed mutagenesis. J. Biochem. Bioeng. 89, 206–209 (2000).

32. Kelly, R. M., Leemhuis, L. H. & Dijkhuizen, L. Conversion of a cyclodextrin glucanotransferase into an α-amylase: Assessment of directed evolution strategies. Biochemistry 46, 11216–11222 (2007).

33. Kelly, R. M. et al. Elimination of competing hydrolysis and coupling side reactions of a cyclodextrin glucanotransferase by directed evolution. Biochem. J. 413, 517–525 (2008).

34. Wang, L., Duan, X. & Wu, J. Enhancing the α-cyclodextrin specificity of cyclodextrin glycosyltransferase from Paenibacillus macerans by mutagenesis masking subsite -7. Appl. Environ. Microbiol. 82, 2247–2255 (2016).

35. Janeček, Š., Svensson, B. & MacGregor, E. A. Relation between domain evolution, specificity, and taxonomy of the α-amylase family members containing a C-terminal starch-binding domain. Eur. J. Biochem. 270, 635–645 (2003).

36. Rashid, N. et al. Characterization of an archaeal cyclodextrin glucanotransferase with a novel C-terminal domain. J. Bacteriol. 184, 777–784 (2002).

37. Fiedler, G., Pajatsch, M. & Böck, A. Genetics of a novel starch utilisation pathway present in Klebsiella oxytoca. J. Mol. Biol. 256, 279–291 (1996).

38. Labes, A. & Schönheit, P. Unusual starch degradation pathway via cyclodextrins in the hyperthermophilic sulfate-reducing archaeon Archaeoglobus fulgidus strain 7324. J. Bacteriol. 189, 8981–8913 (2007).

39. Olsowski, D. M., Jung, J. H., Seo, D. H., Park, C. S. & Holden, J. F. Production of hydrogen from α-1,4- and β-1,4-linked saccharides by marine hyperthermophilic Archaea. Appl. Environ. Microbiol. 77, 3169–3173 (2011).

40. Pajatsch, M. et al. The periplasmic cyclodextrin binding protein GymE from Klebsiella oxytoca and its role in maltodextrin and cyclodextrin transport. J. Bacteriol. 180, 2630–2635 (1998).

41. Kamionka, A. & Dahl, M. K. Bacillus subtilis contains a cyclodextrin-binding protein which is part of a putative ABC-transporter. FEMS Microbiol. Lett. 204, 55–60 (2001).

42. Shim, J. H. et al. Role of maltogenic amylase and pullulanase in maltodextrin and glycosin metabolism of Bacillus subtilis 168. J. Bacteriol. 191, 4835–4844 (2009).

43. Van Den Berg, B., Bhamidimarri, S. P., Prajapati, J. D., Kleinekathöfer, U. & Winterhalter, M. Outer-membrane translocation of bulky small molecules by passive diffusion. Proc. Natl. Acad. Sci. USA. 112, E2991–E2999 (2015).

44. Thomas, C. & Tampé, R. Structural and mechanistic principles of ABC transporters. Annu. Rev. Biochem. 89, 605–636 (2020).

45. Hashimoto, Y., Yamamoto, T., Fujiwara, S., Takagi, M. & Inanaka, T. Extracellular synthesis, specific recognition, and intracellular degradation of cyclomaltodextrins by the hyperthermophilic archaeon Thermococcus sp. strain B1001. J. Bacteriol. 183, 5050–5057 (2001).

46. Hulsmann, A., Luz, R., Scheffel, F. & Schneider, E. Maltose and maltodextrin transport in the thermococcal phagosome of the hyperthermophilic archaean Acidoligococcus acidocaldarius is mediated by a high-affinity transport system that includes a maltose binding protein tolerant to low pH. J. Bacteriol. 182, 6292–6301 (2000).

47. Tonozuka, T. et al. Structural basis for cyclodextrin recognition by Thermatoactinomyces vulgaris cyclo/maltodextrin-binding protein. FEBS J. 274, 2109–2120 (2007).

48. Leisco, E. et al. Multitask ATases (NRDs) of bacterial ABC importers type I and their interspecies exchangeability. Sci. Rep. 10, 1–17 (2020).

49. Zeldes, B. M. et al. Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals. Front. Microbiol. 6, 1–17 (2015).
50. Roth, C. et al. Amylase recognition and ring-size determination of amylomaltase. Sci. Adv. 3, e1601386 (2017).
51. Hemme, C. L. et al. Correlation of genomic and physiological traits of Thermoaerobacter species with biofuel yields. Appl. Environ. Microbiol. 77, 7998–8008 (2011).
52. Jørgensen, S. T., Tangney, M., Starnes, R. L., Amemiya, K. & Jørgensen, P. L. Cloning and nucleotide sequence of a thermostable cyclodextrin glycosyltransferase gene from Thermoaerobacter sp. ATCC 53627 and its expression in Escherichia coli. Biotechnol. Lett. 19, 1027–1031 (1997).
53. Ara, K. Z. G. Cloning and nucleotide sequence of a thermostable cyclodextrin glycosyltransferase with high coupling activity using γ-cyclodextrin isolated from a novel strain clustering under the genus Carboxydovella. Glyobiology 25, 514–523 (2015).
54. Wind, L. E. D. et al. Cyclodextrin formation by the thermostable α-amylase of Thermoanaerobacter tengcongensis MB4. J. Biotechnol. 274, 1698–1709 (1999).
55. Yan, S.-J. & Wu, G. Bottleneck in secretion of α-amylase in Bacillus subtilis sp. strain 1523vc, a thermophilic bacteria, isolated from a hot spring of Uzon Caldera, (Kamchatka, Russia J. Biotechnol. 13, 509–513 (2000).
56. Lee, S.-J. Data on draft genome sequence of Carboxybacter yonseiensis, isolated from a geothermal hot stream. Genome Announc. 1, 1–2 (2013).
57. Martinez-Alonso, E. et al. Taxonomic and functional characterization of a microbial community from a volcanic englacial ecosystem in Deception Island, Antarctica. Sci. Rep. 9, 1–14 (2019).
58. Vemula, H., Ayon, N. J. & Gutheil, W. G. Cytoplasmic peptidoglycan intermediate levels in Bacillus subtilis. Curr. Opin. Chem. Biol. 22, 1992–2003 (2014).
59. Geno, K. A., Hauser, J. R., Gupta, K. & Yother, J. Cystathionine β-synthase (CBS) domain-containing pyrophosphatase as a target for diadenosine tetraphosphate. Biochim. Biophys. Acta 1843, 1687–1697 (2015).
60. Zhao, C.-Q., Wang, J.-Q., Qian, Z., Ma, Y.-H. & Liu, S.-Q. Cloning, expression, and characterization of the 6-phosphofructokinase from Thermoaerobacter tengcongensis. Wei Sheng Wu Xue Bao 46, 249–254 (2006).
61. Navdaeva, V. et al. Phosphoenolpyruvate: Sugar phosphotransferase system from the hyperthermophilic Thermoaerobacter tengcongensis. Biochemistry 50, 1184–1193 (2011).
62. Lin, L. et al. The Thermoaerobacter glycoamylase reveals mechanisms of pentose and hexose co-utilization in bacteria. PLoS Genet. 7, e1002318 (2011).
63. Zhu, M., Lu, Y., Wang, J., Li, S. & Wang, X. Carbon catabolite repression and the related genes of cvpA, ptsH and kprK in Thermoaerobacter aestuariense. PLoS ONE 7, 1–18 (2015).
64. McHenry, C. S. Bacterial replicates and related polymers. Curr. Opin. Biotechnol. 15, 587–594 (2011).
65. Anakhan, V. A. et al. Cystathionine beta Synthase (CBS) domain-containing pyrophosphatase as a target for diadenosine polyphosphates in bacteria. J. Biol. Chem. 290, 27954–27963 (2015).
66. Baykov, A. A., Tuominen, H. K. & Lahti, R. The CBS Domain : A Protein module with an emerging prominent role in regulation. Structure 15, 1156–1163 (2011).
67. Vemula, H., Ayon, N. J. & Gutheil, W. G. Cytoplasmic peptidoglycan intermediate levels in Staphylococcus aureus. Biochimie 121, 72–78 (2015).
68. Geno, K. A., Hauser, J. R., Gupta, K. & Yother, J. Cystathionine β-synthase (CBS) domain-containing pyrophosphatase. J. Bacteriol. 196, 992–2003 (2014).
69. Gonzalez, G. M. et al. Structural insights into RapZ-mediated regulation of bacterial amino-sugar metabolism. Nucleic Acids Res. 45, 10845–10860 (2017).
70. Muchová, K., Chromková, Z., Valencíková, R. & Barák, I. Interaction of the morphogenic protein RodZ with the Bacillus subtilis Mini System. Front. Microbiol. 8, 1–10 (2018).
71. Stamsas, G. A. et al. Identification of ELoR (Spr1851) as a regulator of cell elongation in Streptococcus pneumoniae. Mol. Microbiol. 105, 954–967 (2017).
72. Kaiser, B. K., Clifton, M. C., Chen, B. W. & Stoddard, B. L. The structure of a bacterial DUF199/WhiA protein: Domestication of an invasive endonuclease. Cell Struct. 17, 1366–1376 (2009).
73. Ainsa, J. A. et al. WhiA, a protein of unknown function conserved among Gram-positive bacteria, is essential for sporulation in Streptomyces coelicolor A3 (2). J. Bacteriol. 182, 5470–5478 (2000).
74. Ackert, K., Wu, M., Bergeron, H., Grosse, S. & Lau, P. C. K. Thermostable furuloyl esterase for the bioproduction of ferulic acid from triticale bran. Appl. Microbiol. Biotechnol. 87, 195–203 (2010).
75. Ouygi, M. A., Bashiri, G., Baker, E. N. & Johnson-winters, K. L. Investigating the reaction mechanism of F420-dependent glucose-6-phosphate dehydratase from Mycobacterium tuberculosis: Kinetic analysis of the wild-type and mutant enzymes. Biochemistry 55, 5566–5577 (2016).
76. Greening, C. et al. Physiology, biochemistry, and applications of F420- and F430-dependent redox reactions. Microb. Mol. Biol. Rev. 80, 451–493 (2016).
77. Hans, M. et al. 2-Hydroxyglutaryl-CoA dehydratase from Clostridium symbiosum. Eur. J. Biochem. FEBS 265, 404–414 (1999).
78. Tillander, V., Alexson, S. E. H. & Cohen, D. E. Deactivating fatty acids: Acyl-CoA thioesterase-mediated control of lipid metabolism. Trends Endocrinol. Metab. 28, 473–484 (2018).
90. Arcondeguy, T., Jack, R. & Merrick, M. P II Signal transduction proteins, pivotal players in microbial nitrogen control. *Microbiol. Mol. Biol. Rev.* 65, 80–105 (2001).

91. Badger, J. *et al.* Structural analysis of a set of proteins resulting from a bacterial genomics project. *Proteins* 60, 787–796 (2005).

92. Johansson, M. J. O. & Byström, A. S. Dual function of the tRNA (m5U54) methyltransferase in tRNA maturation. *RNA* 8, 324–335 (2002).

93. He, X. *et al.* Structure of a cation-bound multidrug and toxic compound extrusion transporter. *Nature* 467, 991–994 (2011).

94. Hvorup, R. N. *et al.* The multidrug/oligosaccharidyl-lipid/poly saccharide (MOP) exporter superfamily. *Eur. J. Biochem.* 313, 799–813 (2003).

95. Del-Rio, G., Morett, E. & Soberon, X. Did cyclodextrin glycosyltransferases evolve from α-amylases? *FEBS Lett.* 416, 221–224 (1997).

96. Kuchtovea, A., Gentry, M. S. & Janecek, S. The unique evolution of the carbohyde-binding module CRB20 in lactofin. *FEBS Lett.* 592, 586–598 (2018).

97. Ngo, S. T. *et al.* Interaction of carbohydrate binding module 20 with starch substrates. *RSC Adv.* 2, 24833–24842 (2019).

98. Ludwig, W., Schleifer, K.-H. & Whitman, W. *Bergey’s Manual of Systematic Bacteriology* Vol. 3 (Springer, 2009).

99. Ferreira, M. J. & De Sá-Nogueira, I. A multiskat ATPase serving different ABC-type sugar importers in *Bacillus subtilis*. *J. Bacteriol.* 192, 5312–5318 (2010).

100. Schönert, S. *et al.* Maltose and maltooligosaccharide utilization by *Bacillus subtilis*. *J. Bacteriol.* 188, 3911–3922 (2006).

101. Sahm, K., Matuschek, M., Müller, H., Mitchell, W. J. & Bahl, H. Molecular analysis of the amy gene locus of *Thermoanaerobacterium thermosulfurigenes* EM1 encoding starch-degrading enzymes and a binding protein-dependent maltose transport system. *J. Bacteriol.* 178, 1039–1046 (1996).

102. Matuschek, M., Burchhardt, G., Sahm, K. & Bahl, H. Pullulanase of *Thermoanaerobacterium thermosulfurigenes* EM1 (*Clostridium thermosulfurogenes*): Molecular analysis of the gene, composite structure of the enzyme, and a common model for its attachment to the cell surface. *J. Bacteriol.* 176, 3295–3302 (1994).

103. Chen, I. A. *et al.* IMG/M v.5.0: An integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* 47, D666–D677 (2019).

104. Binder, F., Huberob, O. & Augustb, B. Cyclodextrin-glycosyltransferase from *Klebsiella pneumoniae* M5a: Cloning, nucleotide sequence and expression. *Gene* 47, 269–277 (1986).

105. Lee, M. H. *et al.* Characterization of a thermostable cyclodextrin glucanotransferase from *Pyrococcus furiosus* DSM6368. *Extremophiles* 11, 537–541 (2007).

106. Marchler-Bauer, A. *et al.* CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203 (2017).

107. Stevers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* 27, 135–145 (2018).

108. Crooks, G. E., Hon, G., Chandonia, J. & Brenner, S. E. WebLogo: A sequence logo generator. * Genome Res.* 14, 1188–1190 (2004).

109. Goel, A. & Nene, S. A novel cyclomaltooligosaccharide glucanotransferase from *Bacillus firmus* that degrades raw starch. *Biotechnol. Lett.* 17, 411–416 (1995).

110. Müller, G. L. Use of dinitrosalicyclic acid reagent for determination of reducing sugar. *Anal. Chem.* 31, 426–428 (1959).

111. Milligan, B. G. Maximum-likelihood estimation of relatedness. *Genetics* 1167, 1153–1167 (2003).

112. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729 (2013).

113. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: Recent updates and. *Nucleic Acids Res.* 47, 256–259 (2019).

114. Davis, J. J. *et al.* The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. *Nucleic Acids Res.* 48, D606–D612 (2020).

115. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Res.* 42, 222–230 (2014).

116. Yu, C.-S., Chen, Y.-C., Lu, C.-H. & Hwang, J.-K. Prediction of protein subcellular localization. *Proteins* 64, 643–651 (2006).

117. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* 31, 3784–3788 (2003).

118. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423 (2019).

119. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, 353–361 (2017).

Acknowledgements

SCL and HSP acknowledge financial support from CONACYT grants CB-2016–01-285001, CB-2018–A1-8-18011, and CF-2019–33-549477. SGM was supported by INP 031/2019 and the E022 program from National Institute of Pediatrics, Mexico City, Mexico (Recursos Fiscales para la Investigación). YCC, SCL, and HSP also acknowledge financial support from CONACYT grants INFR-2017–01-280608, and INFR-2021–17-316456. We thank the Centro de Investigación y Asistencia en Tecnología y Diseño del Estado de Jalisco (CIATEJ) for mass spectrometry analysis, in particular Dr. Jacobo Rodriguez-Campos and Dr. Antonio Escobedo-Reyes. LEB was supported by a Postdoctoral fellowship from CONACYT (No. 218862). An MEng fellowship from CONACYT supported RVO (No. 593790) and BVC (753081). We would also like to thank Maria Elena Rodriguez-Alegria and Fernando Gonzalez-Munoz for their technical support.

Author contributions

S.C.L. and H.S.P conceived and designed the project, analyzed the data, and wrote the manuscript. L.E.B., B.V.C., R.V.O., B.H.O., and G.V.C. performed gene cloning, protein production, protein purification, and functional studies. S.C.L., Y.C.C., R.A.S., Y.M.T., and L.E.B. conducted the bioinformatic analysis. S.G.M., L.M.R.R., R.Z.P., J.A.O.C., and A.L.M. contributed to data analysis and general discussion. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-04569-x.
