Structural characterization of a soil viral auxiliary metabolic gene encoding a functional chitosanase

Ruonan Wu
Pacific Northwest National Laboratory
Clyde A. Smith
Stanford Synchrotron Radiation Lightsource
Garry W. Buchko
Pacific Northwest National Laboratory
Ian K. Blaby
Lawrence Berkeley National Laboratory
David Paez-Espino
Mammoth Biosciences
Nikos C. Kyrpides
Lawrence Berkeley National Laboratory
Jason E. McDermott
Pacific Northwest National Laboratory
Kirsten S. Hofmockel
Pacific Northwest National Laboratory
Yasuo Yoshikuni
Lawrence Berkeley National Laboratory
John R. Cort
Pacific Northwest National Laboratory
Janet K. Jansson (janet.jansson@pnnl.gov)
Pacific Northwest National Laboratory

Research Article

Keywords: viral auxiliary metabolic gene, AMG, viral AMG, chitosanase, crystal structure, alpha fold

Posted Date: March 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1485844/v1

Abstract

Metagenomics is unearthing the previously hidden world of soil viruses[1,2]. Many soil viral sequences in metagenomes contain putative auxiliary metabolic genes (AMGs) that are not associated with viral replication. To date only one soil viral AMG has been expressed[3] and none has a solved structure. Here, we aimed to establish that AMGs on soil viruses actually produce functional, active proteins. We focused on AMGs that potentially encode chitosanase enzymes that metabolize chitin – a common carbon polymer. Glycoside hydrolase family 75 viral chitosanase genes were identified from environmental metagenomes. Several of these were expressed and functionally screened. One expressed protein showing endo-chitosanase activity (V-Csn) was crystalized and structurally characterized at ultra-high resolution, thus representing the first structure of a soil viral AMG product. A structure for an inactive mutant with a bound substrate (chitosan oligomer) was also determined. Conserved active site residues in V-Csn resided in a cleft between two domains. These structures provided details about the active site, and together with structure models determined using AlphaFold2, facilitated understanding of substrate specificity and enzyme mechanism. These findings support the hypothesis that soil viruses contribute auxiliary functions to their hosts. Soil viral chitosanase AMG products that assist chitin decomposition may therefore play a previously unrecognized essential role in soil carbon cycling.

Full Text

Recent metagenomic surveys have revealed a high diversity of DNA viruses across a range of soil habitats, including permafrost[4,5], thawed permafrost[3] and grasslands[1]. The majority of these viruses are bacteriophages[1,3], although several eukaryotic viruses have also been detected[4,5]. A fundamental and largely unanswered question concerns the functional roles of these viruses in the soil habitat. It is recognized that soil bacteriophages play a major role in the regulation of host dynamics[6]. Intriguingly, soil viruses may also contribute directly towards biogeochemical processes in soil through expression of genes that potentially encode functions not directly required for viral reproduction. Genes that correspond to non-essential viral functions are referred to as auxiliary metabolic genes (AMGs). Potential functions encoded by AMGs include carbon metabolism, sporulation and energy generation[2,7]. However, only one soil viral AMG has been expressed and functionally characterized to date[3] and there are no existing crystal structures for soil viral AMGs.

The study of AMGs in soil viruses lags behind that of marine environments, due to the high diversity and complexity of soil habitats that has confounded viral discovery. In marine ecosystems, viral AMGs that encode photosynthetic proteins have been extensively studied[8,9]. For example, the structure of a plastocyanin protein encoded by a cyanobacterial phage (cyanophage) has been modeled based on a related reference structure from *Synechococcus* sp. PCC7942[8]. By comparison to the similarly modeled structure from the host plastocyanin, it was possible to predict cyanophage-specific modifications to the structure and electrostatic potential of the cyanophage-encoded plastocyanin. In addition, the structure of a viral rhodopsin has recently been characterized at 1.4 Å resolution. The structure revealed that the viral

rhodopsins are unique light-gated channels that have a predicted role in supporting photosynthesis of algae[10]. These recent discoveries in marine viruses highlight the ecological importance of AMGs that potentially maximize the fitness of phages and hosts in the environment.

The first AMGs that were described in soil viruses were genes encoding enzymes for degradation of various organic compounds. For example, 14 glycoside hydrolase genes were detected in metagenomes from thawed permafrost. One of these, a viral gene encoding a glycosyl hydrolase group 5 (GH5) enzyme, was cloned, expressed and found to represent a functional endomannanase[3]. The vast majority of predicted soil viral AMGs have been assigned potential functions solely based on their sequence similarities to annotated genes in microbial genomic databases[1,6]. This approach is however limited in its ability to determine if the AMG is actually expressed and if the protein is functional.

Here we aimed to characterize and validate the function and structure of a soil viral AMG product. We focused on chitosanase AMGs because of the potentially important role that they play in organic matter decomposition in soil. Chitosanase enzymes are involved in the decomposition of chitin, the second most abundant carbon polymer on the planet after cellulose. Chitin is abundant in many soils because it is a component of fungal cell walls and insect exoskeletons. Following deacetylation of the chitin polymer into chitosan by chitin deacetylases, chitosanases cleave chitosan into smaller subunits that can be further degraded, thereby providing carbon and nitrogen sources for soil biota. Chitosanase genes have previously been annotated in the genome of a giant virus, *Chlorovirus,* that infects green microalgae in terrestrial waters[11] and the detected viral chitosanases were characterized as belonging to the glycosyl hydrolase group 46 (GH46). Here we aimed to study chitosanase-like AMGs carried on soil viruses. Interestingly, they primarily fall into another group previously categorized as GH75 fungal chitosanases (pfam07335) that cleave beta-1,4-chitosans with endo-splitting activity (http://pfam.xfam.org/family/PF07335).

Our approach began with computational screening of publicly available soil metagenomes to identify potential chitosanase sequences from soil viruses. Subsequently, we confirmed that one of the predicted chitosanases was indeed functional by cloning and expressing the gene and conducting activity assays. Based upon sequence comparisons with other chitosanases from the same GH75 family, potential active site residues were identified, and site specific mutation of two of these residues resulted in severely diminished enzyme activity. Finally, we obtained an ultra-high resolution crystal structure of the enzyme, along with a substrate complex of one of the inactive mutants, thus providing details of the potential active site for the GH75 family of chitosanases.

Phylogeny of viral chitosanases

Viral contigs that carried GH75 chitosanase AMGs were retrieved from the Integrated Microbial Genomes and Virome (IMG/VR) database (v3.0). A total of 142 qualified GH75 chitosanase-like AMGs were identified from viral contigs with lengths ranging from 8 to 202 kb. The majority of the sequences were

from bacteriophages with unclassified taxonomy. Two of the viral contigs were high-quality complete and circularized genomes (Supplementary Information Table 1).

A protein tree was constructed from the sequence data to delineate the relatedness of viral chitosanases to other microbial chitosanases deposited in public databases. The viral chitosanases were phylogenetically distinct from archaeal, fungal and bacterial GH75 chitosanases (Fig. 1) which clustered into separate clades according to their taxonomy assignments. The majority of the detected viral chitosanases formed tight and deep clades within bacterial chitosanases (Fig. 1). The phylogenetic placement of the viral chitosanases suggests that they originated from bacteria via genomic exchange and were modified into virus-specific versions during genetic drift and diversification processes[12]. This hypothesis is further supported by the finding that the viral contigs that contained GH75 chitosanases were classified as bacteriophages (Supplementary Information Table 1).

A multiple sequence alignment was constructed to determine whether the viral chitosanases contained four conserved residues residing in a presumed active site for bacterial and fungal GH75 chitosanase sequences[13,14](Supplementary Information Figure 1). The GH75 viral chitosanases generally retain the same four key residues for the predicted active sites, D-D-D-E, although some of the sequences had substitutions. The majority of the GH75 chitosanases have aspartate at the first of the four positions, with a few instances of substitutions of cysteine or asparagine (innermost ring in Fig. 1). As substitutions in the key residues may affect the predicted function, we further divided these viral chitosanases into sub-groups, as it appears that incidences of active site residue variants described above tend to cluster together. The identified viral chitosanases were clustered into three major clades (‘Clade 1’, ‘Clade 2’ and ‘Clade 3’ in Fig. 1). The substitutions of cysteine and asparagine at the first position were only observed in Clade 1 and 3 viral chitosanases, respectively. Clade 2 and Group 2 viral chitosanases contained the same residues as the majority of the bacterial and fungal. Viral chitosanases in Clade 1 with and without cysteine substitutions were named Group 1.1 and Group 1.2, respectively, and those in Clade 3 with and without asparagine substitutions were named Group 3.1 and Group 3.2, respectively. Some of the viral chitosanases grouped according to sample origin with soil viruses grouping separately from aquatic viruses. The soil specific sequences were primarily in Group 1.2 and Clade 3.

AlphaFold predictions

Representatives from the different Clades (Fig. 1) were selected for structural prediction using the recently introduced artificial intelligence-based protein structure prediction software AlphaFold2[15]. The structures of several bacterial and fungal GH75 chitosanases that had been previously reported and characterized in the literature were also predicted. Two characteristic regions of sequence shared by all GH75 members formed a glycosyl hydrolase domain fold seen in other families, for example GH45. These shared regions bracket a novel domain, a variable region containing dissimilar insertions of varying lengths (Supplementary Information Figure 2a). In several predicted structures of viral AMG products this

insertion folds as a novel domain of about 70 amino acids that forms one side of a prominent cleft in the middle of the entire structure. Some bacterial members of GH75 appear to contain a homologous insertion (Supplementary Information Figure 2b). Other viral chitosanase-like AMG sequences and all of the predicted bacterial and fungal sequences lack this longer insertion and do not appear to form a substantial domain. Insofar as these sequences have non-homologous regions of sequence, these structure predictions may be helpful in comparative analysis of the different groups.

Identification of a chitosanase-like AMG with chitosanase activity

The DNA coding sequences for 10 of the 142 GH75 chitosanase-like AMGs were codon-optimized for recombinant expression in *Escherichia coli* (selected sequences are indicated with asterisks in Fig. 1). Expression was observed for nine out of the 10 proteins, but protein was observed primarily in the insoluble fractions in all but two of the initial targets. For the two other targets, enough protein was expressed and purified to assay for endo-chitosanase activity. Only one expressed protein, hereafter called V-Csn (for Viral Chitosanase), showed activity and this activity was maximum near pH 5 (Fig. 2a). The corresponding sequence originated from Group 3.1 and is indicated with two asterisks in Figure 1. This specific sequence originated from a forest soil metagenome (Supplementary Information Table 1). Endo-chitosanase activity for V-Csn was further corroborated by two single-residue substitutions at positions proposed to be part of the chitosanase active site. It was previously postulated, based on biochemical, kinetic and mutational studies on the fungal GH75 chitosanases from *Aspergillus fumigatus*[16] and *Fusarium solani*[14], that two residues (D160 and E169 in *A. fumigatus*, and D175 and E188 in *F. solani*) were essential catalytic residues. Sequence analysis of a bacterial GH75 chitosanase from *Streptomyces avermitilis* showed that these residues were also conserved in this enzyme[13]. Based upon a partial alignment of the V-Csn, *A. fumigatus*, *F. solani*, and *S. avermitilis* sequences (Fig. 2b), the corresponding residues in V-Csn are D148 and E157. Two V-Csn constructs harboring either a D148N substitution or a E157Q substitution were generated and the activities of the respective mutant enzymes measured. Activity was reduced five-fold for D148N and almost completely eliminated for E157Q (Fig. 2c). Both constructs eluted with gel-filtration chromatography retention times identical to native V-Csn, and circular dichroism spectra indicated both constructs were folded (data not shown). Crystallization trials on V-Csn and the two mutant enzymes were subsequently undertaken.

High resolution X-ray structure of V-Csn

The V-Csn structure was solved in two crystal forms; *apo1* containing a single molecule in the asymmetric unit, and *apo2* containing a dimer in the asymmetric unit. The *apo1* structure was solved by single anomalous diffraction (SAD) methods using the signal from bromide ions soaked into crystals of the *apo1* form. The structure was automatically built using *phenix.autobuild*[17] and completed with COOT[18]. The bromide anomalous signal extended to approximately 1.5 Å resolution and the structure

was initially refined into this data. Refinement was completed with *phenix.refine*[19] against a high resolution *apo1* data set extending to 0.89 Å resolution. The final model comprised 1811 protein atoms in a single chain, 398 water molecules, three glycerol and a sulfate anion. The final R_{work} and R_{free} were 0.1198 and 0.1307 for 174427 total reflections. The *apo2* crystal form was solved by molecular replacement using the refined *apo1* structure as the search model, and refined with *phenix.refine*.

A single V-Csn molecule was located in the *apo1* asymmetric unit. The structure consisted of two non-contiguous structural domains (Fig. 2d). The N-terminal part of Domain-1 (residues 1-36) is folded first and the polypeptide then folds the entire Domain-2 (residues 37-108) before crossing back to complete Domain-1 (residues 109-224). During the early stages of the refinement residual density was observed at the N-terminus equivalent to at least three additional residues. Inspection of the sequence of the expression vector suggested that three residues (G, H and S) from the linker were attached to the N-terminal methionine and these residues were added to the model. The *apo2* structure has two independent V-Csn molecules in the asymmetric unit and the two molecules form a non-crystallographic dimer (Extended Data Fig. 1a). Superposition of the two molecules of the *apo2* dimer onto the *apo1* structure gives a root-mean-squared deviation (RMSD) of 0.45 Å for both molecules. In the *apo1* crystal form, the same dimer is observed albeit generated by the crystallographic symmetry of the C2 space group (Extended Data Fig. 1a). This is consistent with the observation that V-Csn exists as a dimer in solution based on size exclusion chromatography, although there is no evidence that dimerization is required for enzyme activity. Formation of the dimer buries 1830 $Å^2$ (~9%) of the surface per monomer. The regions of contact involve the loop between β-strands β4 and β5 (in Domain-2) in one molecule slotting between helices α3 and α4 in the second molecule (Extended Data Fig. 1b), linked via hydrogen bonding and hydrophobic interactions with residues from the two helices and strand β8.

Domain-1 is composed of a central six-stranded antiparallel twisted β-sheet made up of strands β1, β2, β3, β7, β8 and β10 (Fig. 2e and Fig. 2f). Two short strands (β6 and β9) pack against the concave face of the central β-sheet, and two helices (α4 and α5) wrap across the convex face of the sheet. A Dali search[20] using the isolated Domain-1 gives over 1000 hits with a Z-score greater than 5. An initial analysis of the top hits shows that Domain-1 has structural similarity with a diverse range of proteins that all have a common core domain comprising a double-*psi* β-barrel (DPBB) made up of strands β3, β6, β7, β8, β9 and β10. These proteins include the plant defense proteins kiwellin[21,22], barwin[23] and carwin[24], the fungal phytotoxin cerato-platanin[25], the *Streptomyces* papain inhibitor (SPI)[26], domain 1 of the expansins, proteins which loosen plant cell walls[27,28], the human ubiquitin regulatory domain of ASPL[29], and the carbohydrate hydrolyzing endoglucanases[30-32]. Barwin, carwin and the endoglucanases are classified as members of the glycosyl hydrolase GH45 family (https://http://pfam.xfam.org/family/PF02015), and the comparison with V-Csn shows that both the GH45 and GH75 enzymes bear a strong structural similarity. Both families, however, have no structural similarity with the GH46 enzymes, most of which are annotated as chitosanases but which have a two-domain α-helical architecture reminiscent of T4 lysozyme [33].

The DPBB domain, comprising two interlocking *psi*-motifs, was first described for aspartate-α-decarboxylase, endoglucanase V, DMSO reductase and barwin[34]. In V-Csn each *psi*-motif is composed of two long antiparallel strands, one bent almost 90º such that the N-terminal half is almost orthogonal to the C-terminal half, and one single short strand running parallel with the C-terminal part of the long strand (Extended Data Fig. 2a). The two *psi*-motifs (*psi1*; β3, β9 and β10, and *psi2*; β6, β7 and β8) are oriented relative to each other such that the two short strands (β6 and β9) form an antiparallel pair with a pseudo-twofold axis between them mapping *psi1* onto *psi2* (Extended Data Fig. 2b). Superposition of several of the top DPBB-containing hits from Dali demonstrates the conserved topology of the two *psi*-motifs in these unrelated proteins (Extended Data Figs. 2c, 2d and 2e). Several loop extensions decorate the DPBB domains of these proteins. With respect to V-Csn strand numbering, these are: (i) Loop1, between strand β3 of the *psi1* motif and the strand β6 of the *psi2* motif, (ii) Loop2 between β8 of psi2 and β9 of psi1, and (iii) Loop3 between strands β6 and β10 of the *psi1* motif. In V-Csn, the Loop1 extension encapsulates all of Domain-2, and in other proteins this loop varies in length and structure (Extended Data Figs. 2c, 2d and 2e).

The V-Csn Domain-2 is very unusual in that it displays a distinct lack of secondary structure, and at first glance appears to be essentially unstructured (Fig. 2d). A Ramachandran plot analysis of the Domain-2 phi/psi angles (Extended Data Fig. 3) shows clustering of the main chain torsion angles into the favored α- and β- regions as would be expected for a well-folded protein. However, unlike a typical protein structure, Domain-2 seems to lack long continuous stretches of α-helical or β-strand structure. Calculation of the secondary structure characteristics using multiple algorithms including DSSP (as implemented in PROCHECK and PyMOL) and the STRIDE server[35] all identify two single turn 3_{10} helices (α1 and α2) and two short strands (β4 and β5), along with four individual residues annotated as β-bridges (residues G56, W63, V68 and P74). The two short β-strands run anti-parallel to each other and are connected via three interstrand hydrogen bonds (Extended Data Fig. 4a). The four β-bridge residues are localized to a piece of polypeptide between helix α1 and strand β4 which is folded into two hairpin turns and held together by hydrogen bonding interactions between main chain atoms of these β-bridge residues, along with several side chain/main chain hydrogen bonds. (Extended Data Fig. 4b). The AlphaFold2 prediction of V-Csn, made after the crystal structure had been completed, was remarkably close across the entire sequence (0.6 Å RMSD for 222 matching Cα atoms), including the novel Domain-2 (Extended Data Fig. 4c).

The active site

Inspection of the *apo1* structure shows that the two residues identified as putative active site residues (D148 and E157) are located in a cleft between the two structural domains (Fig. 3a). Two additional acidic residues (D34 and D36) are also located in this cleft adjacent to D148, and these residues were also conserved in the other GH75 chitosanases (Fig. 2b). In V-Csn, the side chain of D148 makes hydrogen bonding interactions with the main chain amide nitrogen of A92 and the side chains of both D34 and D36 (Fig. 3b). Although the clustering of acidic residues like this is unusual, it is not

unprecedented and occurs either in metalloproteins where acidic side chains are brought close together by their roles in metal binding, or in enzyme active sites where they share protons[36]. At the pH of crystallization (4.6) it would be expected that most of these acidic residues would be protonated and thus essentially neutral, and calculation of the electrostatic surface within the active site cleft at this pH shows very little negative charge (Extended Data Fig. 5a). At the pH optimum of the enzyme (5.1 - 5.5), however, the aspartate residues would be somewhat less protonated and there is significant negative charge within the cleft (Fig. 3c and Extended Data Fig. 5b), which may be important for attracting the chitosan substrate into the pocket.

Structures of two site-directed chitosanase mutants

Site-directed mutants D148 and E157 were constructed. In both cases the carboxylate was converted to the corresponding amide, generating the mutant proteins D148N and E157Q. The two mutant V-Csn proteins were crystallized under the same conditions as the wild-type protein, and their structures were determined by molecular replacement to high resolution. Superposition of the two mutant structures onto the wild-type *apo1* structure gave RMSDs of 0.11 and 0.07 Å, respectively, for all Cα positions, suggesting very little conformational differences between the mutant and the wild-type structures. Co-crystallization of both mutant proteins with chitohexaose (a β-(1-4)-linked polymer of six D-glucosamine (GlcN) residues) gave a complex with the E157Q mutant only. The substrate was located in the interdomain cleft (Extended Data Figs. 5c and 5d), with three of the six GlcN residues (hereinafter named GlcN-1, GlcN-2 and GlcN-3) visible in F_o-F_c electron density (Fig. 3d), oriented such that GlcN-1 is the reducing end. The GlcN-2 and GlcN-3 residues are in a standard chair conformation, however the density for the first observed residue (GlcN-1) suggested a distorted boat conformation.

The GlcN-1 residue is anchored by a single hydrogen bond between the O6 atom and the side chain of Q157 (Fig. 3e). The central GlcN residue (GlcN-2) makes hydrogen bonding interactions with the side chains of D148 and D36 via its free amine, along with a third to the backbone carbonyl oxygen of A90 from Domain-2. The GlcN-3 residue makes a hydrogen bonding interaction with the carbonyl oxygen of T91 via the O6 atom, and another to a water molecule. Although some additional F_o-F_c density was observed at the non-reducing end of GlcN-3 (the O4 atom), the sparsity of the density did not allow for a fourth GlcN to be modelled. The location of the trisaccharide fragment and the interactions it makes with the protein suggests that residues D36 and D148 form the -2 subsite[37] and play a key role in binding and orienting the substrate (in this case via the GlcN-2 residue). The E157 residue represents the -1 subsite and may serve as the nucleophilic group responsible for bond cleavage, assuming that hydrolysis occurred at the reducing end of GlcN-1.

Proposed mechanism

As noted earlier, biochemical and enzymological studies on some known GH75 chitosanases implicated two acidic residues (equivalent to D148 and E157 in V-Csn) as being critically involved in catalysis[13,14,16]. It was established that the GH75 enzymes are endoglucanases[13] that invert the stereochemistry at the anomeric carbon, producing the α anomer of the oligosaccharide products[14,16], so it is likely that V-Csn is also an inverting enzyme. It is notable that in DPBB enzymes annotated as carbohydrate binding and/or hydrolyzing enzymes, the acidic residue at a position equivalent to E157 in V-Csn is universally conserved (either a glutamate or an aspartate), based upon the superposition of V-Csn with these DPBBs and the subsequent generation of a structure-based partial sequence alignment (Extended Data Fig. 5e). Structural studies on endoglucanase V (Cel45) from *Humicola insolens* (PDB code 3ENG)[30,31] and the endo-β-1,4-glucanase (CaCel45) from *Cryptopygus antarcticus* (PDB code 5H4U)[32] suggest that this acidic residue is the catalytic proton donor in the GH45 enzymes, and given the structural similarity of the DPBB domains in the GH45 and GH75 enzymes, it is highly likely that E157 is the catalytic proton donor in V-Csn. It should be noted that the GH45 enzymes are classified as endoglucanases which also lead to inversion of configuration at the anomeric carbon of the cleaved glycosidic bond[38].

The identity of the catalytic base is less clear, although based upon the same superposition and sequence alignment (Extended Data Fig. 5e), members of the GH45 family (Cel45 and CaCel) have acidic residues near the N-terminus of the respective enzymes (D10 in Cel45 and D13 in CaCel) which have been identified as the general base accepting the proton during hydrolysis of the β(1,4) glycosidic bond[30,32]. These residues are structurally equivalent to D36 in V-Csn (Extended Data Fig. 5f), which suggests that this residue may be acting as the catalytic base in the viral enzyme also. As previously noted, the fungal and bacterial GH75 enzymes have an aspartate residue at this same location (Fig. 2b). Conversely, other DPBB enzymes tentatively annotated as carbohydrate binding and/or hydrolytic enzymes lack an acidic residue equivalent to D34 or D36 (Extended Data Fig. 5e) with the exception of the *Streptomyces* papain inhibitor protein (PDB code 5NTB)[26] and kiwellin (PDB code 4PMK)[21], yet they do have an aspartate structurally equivalent to D148 which may be acting as the catalytic base in these enzymes. Although the function of each of the four acidic residues in the V-Csn active site are not yet fully understood, their clustering within the cleft, and the corroborating evidence from the GH45 and GH75 enzymes suggest that it is highly likely that they will have roles in substrate binding (D34 and D148) and the catalytic mechanism (D36 and D157). Validation of the assignment of function must await further mutational and substrate binding studies.

Ecological implications

To summarize, there are several ecological implications of this study. We conclusively demonstrate that at least some AMGs carried on soil viruses are functional. Our rigorous analyses not only resulted in the first crystal structure of a soil viral AMG product, but also enabled us to propose the mode of action of this novel chitosanase enzyme in the GH75 family of glycosyl hydrolases. The chitosanase sequences that were included and compared revealed a phylogenetic distinction between viral chitosanases and

those previously described in bacteria and fungi. However, because the viral chitosanases were subgroups within bacterial clades and the viruses detected with the chitosanase AMGs were bacteriophages, this suggests that they originated from bacteria. The soil viral chitosanases also formed subgroups that were distinct from their counterparts in aquatic systems. The V-Csn enzyme that we functional and structurally characterized originated from a sequence from a forest soil (DOI 10.46936/10.25585/60000627, Supplementary Information Table 1). Forest soils are often characterized as having more fungi than other soil types[39]. This may be a reason for selection of viruses that carry the capacity to help to decompose chitin - a major component of fungal cell walls and an important source of both carbon and nitrogen. The reason for selection of a virus that carries this capacity independently of its host is currently unknown. By analogy to marine systems where viruses carry AMGs that help to support energy generation via photosynthesis in their respective hosts[8,9,40], soil viruses may also help their hosts to decompose available carbon resources in soil as they become available.

References

1. Wu, R. *et al.* DNA viral diversity, abundance, and functional potential vary across grassland soils with a range of historical moisture regimes. *mBio* **12**, e0259521 (2021).

2. Trubl, G. *et al.* Soil viruses are underexplored players in ecosystem carbon processing. *MSystems* **3**, e00076-00018 (2018).

3. Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat. Microbiol.* **3**, 870-880 (2018).

4. Christo-Foroux, E. *et al.* Characterization of mollivirus kamchatka, the first modern representative of the proposed molliviridae family of giant viruses. *J. Virol.* **94**, e01997-01919 (2020).

5. Legendre, M. *et al.* In-depth study of *Mollivirus sibericum*, a new 30,000-y-old giant virus infecting *Acanthamoeba*. *Proc. Natl. Acad. Sci. USA* **112**, E5327-5335 (2015).

6. Trubl, G. *et al.* Active virus-host interactions at sub-freezing temperatures in Arctic peat soil. *Microbiome* **9**, 208 (2021).

7. Breitbart, M. Marine viruses: Truth or dare. *Ann. Rev. Marine Sci.* **4**, 425-448 (2012).

8. Puxty, R. J., Millard, A. D., Evans, D. J. & Scanlan, D. J. Shedding new light on viral photosynthesis. *Photosynth. Res.* **126**, 71–97 (2015).

9. Crummett, L. T., Puxty, R. J., Weihe, C., Marston, M. F. & Martiny, J. B. H. The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. *Virology* **499**, 219-229 (2016).

10. Zabelskii, D. *et al.* Viral rhodopsins 1 are an unique family of light-gated cation channels. *Nature Comm.* **11**, 5707 (2020).

11. Jeanniard, A. *et al.* Towards defining the chloroviruses: A genomic journey through a genus of large DNA viruses. *BMC Genomics* **14**, 158 (2013).

12. Sanjuán, R. & Domingo-Calap, P. in *Encyclopedia of Virology* Vol. 1 (eds D. Bamford & M. Zuckerman), pp. 53-61 (2021).

13. Heggset, E. B. *et al.* Mode of action of a family 75 chitosanase from *Streptomyces avermitilis*. *Biomacromol.* **13**, 1733–1741 (2012).

14. Shimosaka, M., Sato, K., Nishiwaki, N., Miyazawa, T. & Okazaki, M. Analysis of essential carboxylic amino acid residues for catalytic activity of fungal chitosanases by site-directed mutagenesis. *J. Biosci. Bioeng.* **100**, 545-550 (2005).

15. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

16. Cheng, C.-Y., Chang, C.-H., Wu, Y.-J. & Li, Y.-K. Exploration of Glycosyl Hydrolase Family 75, a Chitosanase from *Aspergillus fumigatus*. *J. Biol. Chem.* **281**, 3137–3144 (2005).

17. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr.* **D66**, 213-221 (2010).

18. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr.* **D66**, 486–501 (2010).

19. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallogr.* **D68**, 352-367 (2012).

20. Holm, L. Using Dali for protein structure comparison. *Methods Mol. Biol.* **2112**, 29-42 (2020).

21. Hamiaux, C. *et al.* Crystal structure of kiwellin, a major cell-wall protein from kiwifruit. *J. Struct. Biol* **187**, 276–281 (2014).

22. Offermann, L. R. *et al.* Elusive structural, functional, and immunological features of Act d 5, the green kiwifruit kiwellin. *J. Agric. Food Chem.* **63**, 6567–6576 (2015).

23. Ludvigsen, S. & Poulsen, F. M. Three-dimensional structure in solution of barwin, a protein from barley seed. *Biochemistry* **31**, 8783–8789 (1992).

24. Huet, J. *et al.* High-resolution structure of a papaya plant-defence barwin-like protein solved by in-house sulfur-SAD phasing. *Acta Crystallogr.* **D69**, 2017-2026 (2013).

25. de Oliveira, A. L. *et al.* The structure of the elicitor cerato-platanin (CP), the first member of the CP fungal protein family, reveals a double psi beta-barrel fold and carbohydrate binding. *J. Biol. Chem.* **286**, 17560-17568 (2011).

26. Juettner, N. E. *et al.* Illuminating structure and acyl donor sites of a physiological transglutaminase substrate from *Streptomyces mobaraensis*. *Protein Sci.* **27**, 910-922 (2018).

27. Georgelis, N., Yennawar, N. H. & Cosgrove, D. J. Structural basis for entropy-driven cellulose binding by a type-A cellulose-binding module (CBM) and bacterial expansin. *Proc. Natl. Acad. Sci.* **109**, 14830-14835 (2012).

28. Yennawar, N. H., Li, L. C., Dudzinski, D. M., Tabuchi, A. & Cosgrove, D. J. Crystal structure and activities of EXPB1 (Zea m 1), a β-expansin and group-1 pollen allergen from maize. *Proc. Natl. Acad. Sci.* **103**, 14664-14671 (2006).

29. Arumughan, A. *et al.* Quantitative interaction mapping reveals an extended UBX domain in ASPL that disrupts functional p97 hexamers. *Nat. Commun.* **7**, 13047-13047 (2016).

30. Davies, G. J. *et al.* Structure determination and refinement of the *Humicola insolens* endoglucanase V at 1.5 A resolution. *Acta Crystallogr.* **D52**, 7-17 (1996).

31. Davies, G. J. *et al.* Structure and function of endoglucanase V. *Nature* **365**, 362-364 (1996).

32. Song, J. M. *et al.* Genetic and structural characterization of a thermo-tolerant, cold-active, and acidic endo-β-1,4-glucanase from Antarctic springtail, *Cryptopygus antarcticus*. *J. Agric. Food Chem.* **65**, 1630-1640 (2017).

33. Marcotte, E. M., Monzingo, A. F., Ernst, S. R., Brzezinski, R. & Robertas, J. D. X-ray structure of an anti-fungal chitosanase from *Streptomyces* N174. *Nature Struct. Biol.* **3**, 155-162 (1996).

34. Castillo, R. M. *et al.* A six-stranded double-psi beta barrel is shared by several protein superfamilies. *Structure* **7**, 227–236 (1999).

35. Heinig, M. & Frishman, D. STRIDE: A web server for secondary structure assignment from known atomic coordinates of proteins. *Nucl. Acids Res.* **32**, W500-502 (2004).

36. Flocco, M. M. & Mowbray, S. L. Strange bedfellows: interactions between acidic side-chains in proteins. *J. Mol. Biol.* **254**, 96-105 (1995).

37. Davies, G. J., Wilson, K. S. & Henrissat, B. Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochem. J.* **321**, 557-559 (1997).

38. Schou, C., Rasmussen, G., Kaltoft, M., Henrissat, B. & Schülein, M. Stereochemistry, specificity and kinetics of the hydrolysis of reduced cellodextrins by nine cellulases. *Eur. J. Biochem.* **217**, 947-953 (1993).

39. Pollierer, M. M., Dyckmans, J., Scheu, S. & Haubert, D. Carbon flux through fungi and bacteria into the forest soil animal food web as indicated by compound-specific 13C fatty acid analysis. *Funct. Ecol.* **26**,

978-990 (2012).

40. Wang, M. *et al.* Genomic analysis of *Synechococcus* phage S-B43 and its adaption to the coastal environment. *Virus Res.* **289**, 198155 (2020).

Methods

Viral contig acquisition and chitosanase AMG detection. The Integrated Microbial Genomes and Virome (IMG/VR) database (v3.0)[41]was screened for sequences corresponding to predicted chitosanase genes. Viral contigs with genes annotated by a chitosanase HMM (pfam07335) were first identified by applying a JGI viral detection pipeline[42]. For a more conservative functional assignment, the viral chitosanase sequences were further checked against annotation databases including EggNOG[43], the carbohydrate-active enzyme database (CAZY)[44] and the functional ontology assignments for metagenomes database (FOAM)[45] using hmmsearch (Hmmer v3.1b2)[46] as described previously[1] and searching for sequence similarities to NCBI chitosanases using blastp[47]. The putative viral chitosanases were then screened against a profile of lysozyme HMMs to remove the mis-annotated lysozymes (PF13702, PF00959, PF04965, PF18013, PF00062 and a self-curated lysozyme HMM[1] using the lysozyme sequences deposited at NCBI viruses (accessed on 16 November 2020).

For a confident assignment of the chitosanase genes as viral AMGs, the genomic content of the viral contigs carrying chitosanase genes screened from the above steps were inspected. Genes from viral contigs were predicted and translated using Prodigal[48]. The protein sequences were annotated by EggNOG bacterial and archaeal databases and three viral databases as previously described[1,49], in addition to the 7185 microbial-specific and 8773 viral-specific HMMs implemented in checkV (v0.7.0)[50]. The chitosanase AMG candidates were classified into five categories according to their gene positions on viral contigs and presence or absence of viral hallmark genes as described previously[1]. Only viral contigs with high confidence scores (categories 0-2) for chitosanase AMGs were retained for subsequent analyses (Supplementary Information Table 1).

Viral contig clustering and host prediction. The viral contigs with chitosanase AMGs were clustered with Viral RefSeq genomes (v201) based on a scored protein sharing matrix. A clustering network including pairwise interactions was generated by applying vConTACT using default parameters (v2.0.9.10)[51]. The soil viral contigs did not share sufficient genes with previously deposited reference viruses to enable a confident taxonomic assignment (data not shown).

The putative hosts of the viral contigs that carried chitosanase AMGs were predicted using three published bioinformatic tools: 1) WIsH[52] (best-hit), 2) VirHostMatcher[53] (best-hit) and 3) Prokaryotic virus Host Predictor (PHP)[54] (‘consensus’). The final host taxonomy of a viral contig was assigned when results from at least two of the three tools reached consensus.

Phylogenetic analysis of chitosanases. To delineate the phylogenetic relatedness of the detected viral chitosanases to GH75 chitosanases in other taxa, a phylogenetic tree was constructed based on multiple sequence alignments of protein sequences of archaeal, bacterial, fungal and viral chitosanases. The tree was re-rooted using a bacteriophage lysozyme (YP_006987285.1). In order to cover the diverse genetic space across all domains of life, we first queried 'chitosanase' from NCBI protein database (https://www.ncbi.nlm.nih.gov/protein, accessed on Oct 11th, 2021) and further screened by the GH75 chitosanase pfam (PF07335). Sequences of the bacterial and fungal GH75 chitosanases used to identify key residues in the active sites were also included as part of the reference[13,14]. The reference sequences were then clustered at 70% amino acid identity to remove redundancy using CD-HIT (v4.8.1)[55] and the representative sequence of each cluster with length longer than 150 amino acids was included in the final reference set, resulting in two sequences from archaea, 230 from bacteria and 180 from fungi. The viral chitosanases and the reference sequences were aligned using MAFFT with default parameters (v7)[56]. The multiple sequence alignments were manually inspected and adjusted based on positions of the four key residues of the predicted active site across the viral and reference sequences (Supplementary Information Figure 1). The phylogenetic tree was built using FastTree (v2.1)[57] with default parameters.

Protein Expression and Purification. The gene encoding for a putative soil viral chitosanase sequence (Ga0126380_1000012531: noted with a double asterisk in Fig. 1) was chemically synthesized and inserted into the NdeI site of pET28a inclusive of a 20-residue extension at the N-terminus (MGSS**HHHHHH**SSGLVPRGSH-) containing a poly-histidine metal affinity tag (bold) and thrombin protease cleavage site (underlined) in the primary amino acid sequence of the expressed protein. The recombinant plasmid was used to transform chemically competent *Escherichia coli* BL21(DE3) (Invitrogen, Carlsbad, CA) from which ~1 mL ~15% glycerol stocks (LB media, OD_{600nm} = ~0.8) were prepared from a single colony and frozen (-80 °C) for future use. This glycerol stock was used to seed 25 mL of LB medium that was grown to an OD_{600nm} of ~ 0.8 and then transferred to 750 mL of autoinduction LB medium[58] (2 L flasks, 200 rpm shaker, 0.34 ug/uL kanamycin, 37 °C). Upon reaching an OD_{600nm} of approximately 1, the temperature was lowered to 30 °C. The cells were harvested ~16 h later (next day) by gentle centrifugation and then frozen (-80 °C). Cells were lysed by thawing the frozen pellet followed by sonication (~ 1 min) before and after three passes through a French Press (SLM Aminco, Rochester, NY). Following centrifugation, the protein in the soluble fraction was purified using a conventional two-step purification protocol: metal chelate affinity chromatography on a 20 mL Ni-Agarose 6 FastFlow column (GE Healthcare, Piscataway, NJ) followed by gel-filtration chromatography on a Superdex HiLoad 26/60 column (GE Healthcare, Piscataway, NJ)[59]. Fractions containing the target protein after the last column step were concentrated to 2 – 5 mg/mL (Protein Buffer: 100 mM NaCl, 20 mM Tris, 1 mM DTT, pH 7) and stored at 4 °C until used for crystallization or enzyme assays. Yields of 2 – 4 mg purified protein were obtained per liter LB medium. The same protocol was applied to prepare two

modified proteins each containing the point substitution D148N or E157Q. Mutagenesis was performed as previously described [60].

Chitosanase activity assays. Wildtype V-Csn and the two modified proteins were tested for *endo*-chitosanase activity using an azurine cross-linked (AZCL) chitosan substrate (AZCL-chitosan; Megazyme, Wicklow, Ireland)[61]. Stock solutions (1200 µg/mL) of each protein were prepared in Protein Buffer along with AZCL-chitosan suspensions (250 µg/mL) at pH 4.3, 5.1, and 6.5 in 40 mM sodium acetate, 100 mM NaCl, 1 mM DTT. The reactions were performed in triplicate, at room temperature, by adding 17 µL of protein (20 µg) to 100 µL of AZCL-chitosan in a 500 µL Eppendorf tube. The tubes were agitated by rotation (40 rpm) in a Multi-Purpose Tube Rotator (Fisher Scientific). Activity was monitored by pelleting the substrate with brief centrifugation and measuring the absorbance of released azurine-linked product at 590 nm (NanoDrop 2000c; Thermo Scientific) using a 2 µL aliquot. Blank reactions showed no release of azurine-linked product in the absence of protein and pH measurements before and after the reaction varied less than 0.1 pH unit.

Crystallization, X-ray data collection and processing. Initial crystallization conditions for V-Csn were obtained using the hanging drop method employing the Top96 screen (Anatrace). Crystals were observed in multiple conditions. Crystals from several conditions were harvested and flash-cooled in liquid nitrogen in their respective crystallization conditions augmented with 20% ethylene glycol. The crystals were sent to SSRL for diffraction screening on beamline BL9-2. Three conditions gave crystals which diffracted to high resolution; condition #45 (0.2 M ammonium sulfate, 0.1 M sodium acetate pH 4.6, 30% MMePEG2000) in space group C2 with unit cell dimensions a=108.84 Å, b=47.63 Å, c=45.55 Å, β=97.8º, with one monomer in the asymmetric unit (AU); condition #38 (0.1 M citrate pH 5.5, 20% PEG3000) in space group C2 with unit cell dimensions a=163.30 Å, b=46.00 Å, c=73.56 Å, β=92.3º, with two monomers in the AU; and condition #20 (0.2 M ammonium sulfate, 0.1 M bis-tris pH 5.5, 25% PEG3350) in space group C2 with unit cell dimensions a=80.47 Å, b=35.76 Å, c=80.66 Å, β=118.5º, with one monomer in the AU.

Data sets were collected from single crystals in conditions #45 and #38. For the condition #45 crystal (designated *apo1*), 1800 0.2º images were collected on BL12-2 using X-rays at 17000 eV (0.72929 Å) and a Pilatus 6M PAD detector running in shutterless mode. The images were processed with XDS[62] and scaled using AIMLESS[63]. The final data set comprised 174574 unique reflections to 0.89 Å resolution. For the condition #38 crystal (*apo2*), 1800 0.2º images were collected on BL9-2 using X-rays at 12658 eV (0.97946 Å) and a Pilatus 6M PAD detector running in shutterless mode. The images were processed with XDS[62] and scaled using AIMLESS[63], and the final data set comprised 117982 unique reflections to 1.35 Å resolution. Additional data collection and processing statistics for both crystal forms are given in Extended Data Table 1.

For experimental phasing, a KBr soaking solution was prepared by dissolving solid KBr in condition #45 crystallization buffer augmented with 25% glycerol until a saturated solution was obtained (as determined visually under a microscope). This solution was diluted with fresh buffer to form a 1/8 saturated crystal soaking solution. Several *apo1* crystals were swished quickly in this solution and flash-cooled in liquid nitrogen. Diffraction data sets were collected from KBr-soaked *apo1* crystals on beamline BL12-2 at the bromide edge (13481 eV, 0.91967 Å). A total of 3600 images were collected with a rotation angle of 0.2º/image, using the inverse beam method and 20º wedges. The images were processed with XDS[62] and scaled using AIMLESS[63]. Additional statistics are given in Extended Data Table 1. Initial analysis of the data indicated a strong anomalous signal from the bromide extending to approximately 1.7 Å resolution.

Structure determination and refinement. The V-Csn structure was solved by Br-SAD (bromide single anomalous diffraction) methods implemented in PHENIX[17]. Following solvent flattening and density modification, the overall figure of merit (FOM) was 0.363 for 16 bromide sites. Autobuilding in PHENIX generated a model comprising 221 out of 224 expected residues. Initial refinement with *phenix.refine*[19] gave an R_{work} and R_{free} of 0.158 and 0.187, respectively. The model was completed using COOT[18] and refined further with *phenix.refine* using the *apo1* data to 0.89 Å resolution. Water molecules were added at structurally and chemically relevant positions, and the atomic displacement parameters for all atoms in the structure were refined isotropically. The *apo2* structure was solved by molecular replacement using the program MOLREP[64] from the CCP4 suite[65], using the refined *apo1* structure as the search model. Final refinement statistics for the two apo-V-Csn structures are given in Extended Data Table 2.

Chitosanase mutant and substrate structures. V-Csn mutants D148N and E157Q were screened for crystallization using conditions #20, #38 and #45, and crystals were observed in all three. Diffraction data sets were collected from single D148N and E157Q crystals from condition #45. For the D148N crystals, 1800 images (0.2º rotation/image) were collected on BL12-2, and the data processed and scaled with XDS[62] and AIMLESS[63]. For the E157Q crystal, 1850 images were collected on BL12-2, and the data processed and scaled with XDS[62] and AIMLESS[63]. Data collection statistics are given in Extended Data Table 3. Both structures were solved by molecular replacement with MOLREP[64] using the refined wild-type V-Csn structure as the starting model, with all water molecules removed. The D148N and E157Q structures were refined with *phenix.refine*[19], and final statistics are also given in Extended Data Table 4.

The E157Q-substrate complex was prepared by dissolving 0.06 mg of chitohexaose (Biosynth) in 10 uL of E157Q at 3.3 mg/ml, giving a final chitohexaose concentration of around 5 mM. The complex was incubated at 4 ºC for 1 h prior to setting up sitting drops against crystallization condition #45. The crystallization drops were streak-seeded several hours after setup and crystals of the complex were observed in all drops overnight. The crystals were morphologically similar to wild-type and mutant

crystals grown under the same conditions. The crystals were transferred into crystallization buffer augmented with 25% glycerol, and flash-cooled in liquid nitrogen. Diffraction data were collected at BL12-2. A total of 1800 images were collected, and the data processed and scaled with XDS[62] and AIMLESS[63]. The E157Q-substrate complex structure was solved by molecular replacement with MOLREP[64] using the refined wild-type V-Csn structure with all water molecules removed as the starting model, and refined with *phenix.refine*[19]. Data collection and refinement statistics are given in Extended Data Tables 3 and 4.

Structure modeling by AlphaFold2. The AlphaFold2 structure predictions were run using either a locally-installed version of the software retrieved from the official GitHub repository (https://github.com/deepmind/alphafold) or the Google collaborative AlphaFold2 notebook (https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb). Solvent accessible surfaces were calculated with PyMOL (v2.5.2) (Schrodinger) and ICM-Pro (v3.8-6a) (Molsoft), using a probe radius of 1.4 Å (equivalent to the radius of a single water molecule). The electrostatic surfaces were generated with the Adaptive Poisson-Boltzmann Solver (APBS) plugin for PyMOL (v2.5.2).

Data availability. The atomic coordinates and structure factors for the protein structures have been submitted to the Protein Data Bank as follows: V-Csn *apo1*, PDB code 7TVL; V-Csn *apo2*, 7TVM; V-Csn-D148N, 7TVN; V-Csn-E157Q, 7TVO; V-Csn-E157Q chitohexaose complex, 7TVP. The wwPDB X-ray structure validation reports are included in Supplementary Information Figure 3.

Code availability. No custom code or custom mathematical algorithms were applied to this study.

41. Roux, S. *et al.* IMG/VR v3: An integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucl. Acids Res.* **49**, D764–D775 (2021).

42. Paez-Espino, D., Pavlopoulos, G. A., Ivanova, N. N. & Kyrpides, N. C. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nat. Protoc.* **12**, 1673-1682 (2017).

43. Huerta-Cepas, J. *et al.* EggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucl. Acids Res.* **44**, D286–D293 (2016).

44. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucl. Acids Res.* **37**, D233-D238 (2009).

45. Prestat, E. *et al.* FOAM (Functional Ontology Assignments for Metagenomes): A Hidden Markov Model (HMM) database with environmental focus. *Nucl. Acids Res.* **42**, e145 (2014).

46. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, 1002195 (2011).

47. Johnson, M. *et al.* NCBI BLAST: A better web interface. *Nucl. Acids Res.* **36**, W5–W9 (2008).

48. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).

49. Wu, R. *et al.* Moisture modulates soil reservoirs of active DNA and RNA viruses. *Commun. Biol.* **4**, 992 (2021).

50. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* **39**, 578-585 (2021).

51. Bolduc, B. *et al.* vConTACT: An iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243 (2017).

52. Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113-3114 (2017).

53. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d_2* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucl. Acids Res.* **45**, 39-53 (2017).

54. Lu, C. *et al.* Prokaryotic virus host predictor: A Gaussian model for host prediction of prokaryotic viruses in metagenomics. *BMC Biol.* **19**, 5 (2021).

55. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

56. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).

57. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PloS One* **5**, e9490 (2010).

58. Studier, F. W. Protein production by auto-induction in high-density shaking cultures. *Prot. Expr. Purif.* **41**, 207-234 (2005).

59. Buchko, G. W., Clifton, M. C., Wallace, E. G., Atkins, K. A. & Myler, P. J. Backbone chemical shift assignments and secondary structure analysis of the U1 protein from the Bas-Congo virus. *Biomol. NMR Assign.* **11**, 51–56 (2017).

60. Wrenbeck, E. E. *et al.* Plasmid-based one-pot saturation mutagenesis. *Nature Methods* **13**, 928–930 (2016).

61. Schönbichler, A., Díaz-Moreno, S. M., Srivastava, V. & McKee, L. S. Exploring the Potential for Fungal Antagonism and Cell Wall Attack by *Bacillus subtilis natto*. *Front. Microbiol.* **11**, 521 (2020).

62. Kabsch, W. XDS. *Acta Crystallogr.* **D66**, 125-132 (2010).

63. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr.* **D69**, 1204–1214 (2013).

64. Vagin, A. & Teplyakov, A. MOLREP: An automated program for molecular replacement. *J. Appl. Cryst.* **30**, 1022-1025 (1997).

65. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Cryst.* **D67**, 235-242 (2011).

66. Weiss MS. 2001. Glocal indicators of X-ray data quality. *J Appl Crystallogr* 34:130-135.

67. Karplus PA, Diederichs K. 2012. Linking crystallographic model and data quality. *Science* 336:1030-1033.

68. Evans PR. 2006. Scaling and assessment of data quality. *Acta Crystallogr* D62:72-82.

69. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC. 2010. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallogr* D66:12-21.

Declarations

Online Content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at XXXX (Note to reviewers: to add at publication stage).

Acknowledgements

This work was supported by the Department of Energy (DOE) Office of Biological and Environmental Research (BER) and is a contribution of the Scientific Focus Area “Phenotypic response of the soil microbiome to environmental perturbations”. A portion of this work was performed on a project award (https://doi.org/10.46936/cpcy.proj.2021.60161/60000437) under the FICUS program (PI-Jason McDermott, PNNL) and used resources at the DOE Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL), which are DOE Office of Science User Facilities. Both facilities are

sponsored by the Biological and Environmental Research program and operated under Contract Nos. DE-AC02-05CH11231 (JGI) and DE-AC05-76RL01830 (EMSL). The crystal structures were determined at the Stanford Synchrotron Radiation Lightsource (SSRL). SSRL is a National User Facility operated by Stanford University on behalf of the U.S. Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy (BES, BER) and by the National Institutes of Health (NCRR, BTP, NIGMS). The project described was also supported by Grant Number 5 P41 RR001209 from the NCRR, a component of the National Institutes of Health. Preliminary results of enzymatic function were provided by Gregor Tegl and Stephen Withers from the University of British Columbia.

Author contributions

RW carried out the majority of the bioinformatics analyses. CAS obtained the crystal structures of the chitosanase enzymes and elucidated their mechanisms. GWB expressed the chitosanase proteins, screened crystallization conditions and performed activity assays. IKB and YY carried out the synthesis and cloning of chitosanase genes and mutants. NCK provided the database retrieval and DOIs of publicly available chitosanase sequences. DP-E, JC and CAS performed the AlphaFold2 predictions of protein structure. KSH, JKJ, JEM and CAS funded the research. JKJ coordinated the study. All authors contributed to writing of the manuscript.
The authors declare no competing interests.

Figures

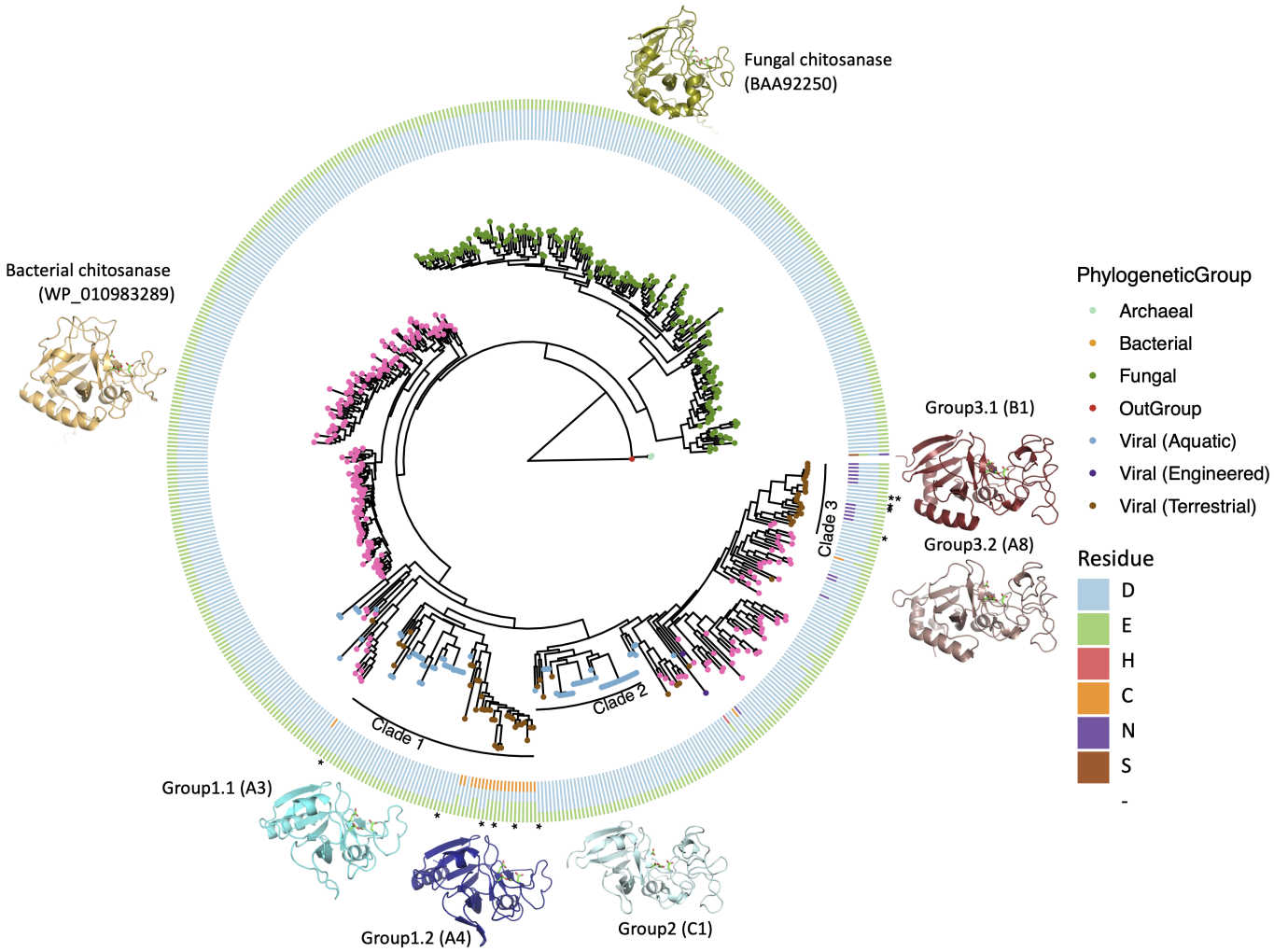

Figure 1

Phylogenetic tree of GH75 chitosanases detected across domains of life. The tree with viral, fungal (green nodes), bacterial (pink nodes) and archaeal (light blue nodes) chitosanases is rooted by a bacteriophage lysozyme (YP_006987285.1, red node). The tree tips representing viral chitosanases are colored by types of habitat. The four key residues are color coded and shown in the circular rings sequentially with the residue close to the N-terminal position in the innermost ring. The protein structures of the representative chitosanases were predicted by AlphaFold2[15]. The viral chitosanases selected for enzymatic function validation are highlighted with asterisks. The viral chitosanase used for crystallization is labeled with two asterisks.

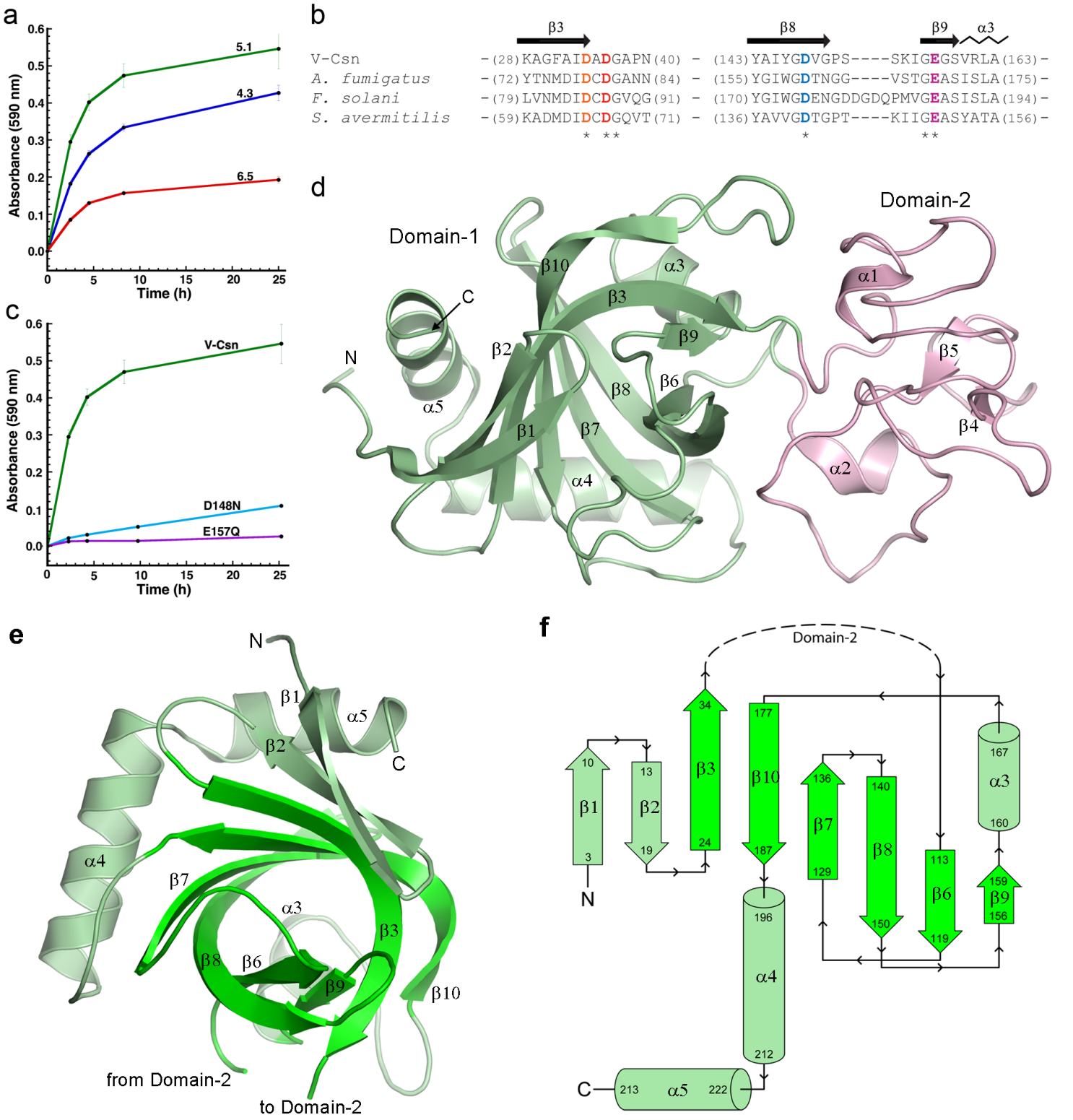

Figure 2

The activity and structure of the soil viral chitosanase enzyme V-Csn. a. Activity of V-Csn was assayed by monitoring azurine absorbance at 590 nm following its release as soluble azurine-sugar fragments from an insoluble azurine cross-linked chitosan substrate (AZCL-chitosan). Reactions were performed in triplicate at room temperature over a range of pHs in acetate buffer with the standard deviation indicated by the error bars. **b.** Partial sequence alignment of V-Csn with three chitosanase enzymes from

Aspergillus fumigatus, *Fusarium solani* and *Streptomyces avermitilis*. The secondary structure for V-Csn is indicated above the alignment. Four conserved acidic residues identified as being potentially involved in catalysis are colored orange, red, blue and purple. Alignment of available sequences of GH75 family enzymes show that only six positions are universally conserved, these four acidic residues and two glycine residues, as indicated by the asterisks beneath the alignment. **c.** Release of azurine by wild-type V-Csn and two constructs containing either a D148N or E157Q substitution, in acetate buffer at pH 5.1. **d.** Ribbon representation of the *apo1* form of the V-Csn enzyme with the two structural domains colored light green (Domain-1) and pink (Domain-2). The secondary structure nomenclature is given, along with the locations of the N- and C-termini. **e.** Domain-1 of the *apo1* structure oriented to look down the double-psi β-barrel (DPBB) structural motif. The strands making up the DPBB are highlighted in bright green. **f.** Topology diagram of Domain-1 highlighting the fold of the DPBB motif (bright green).

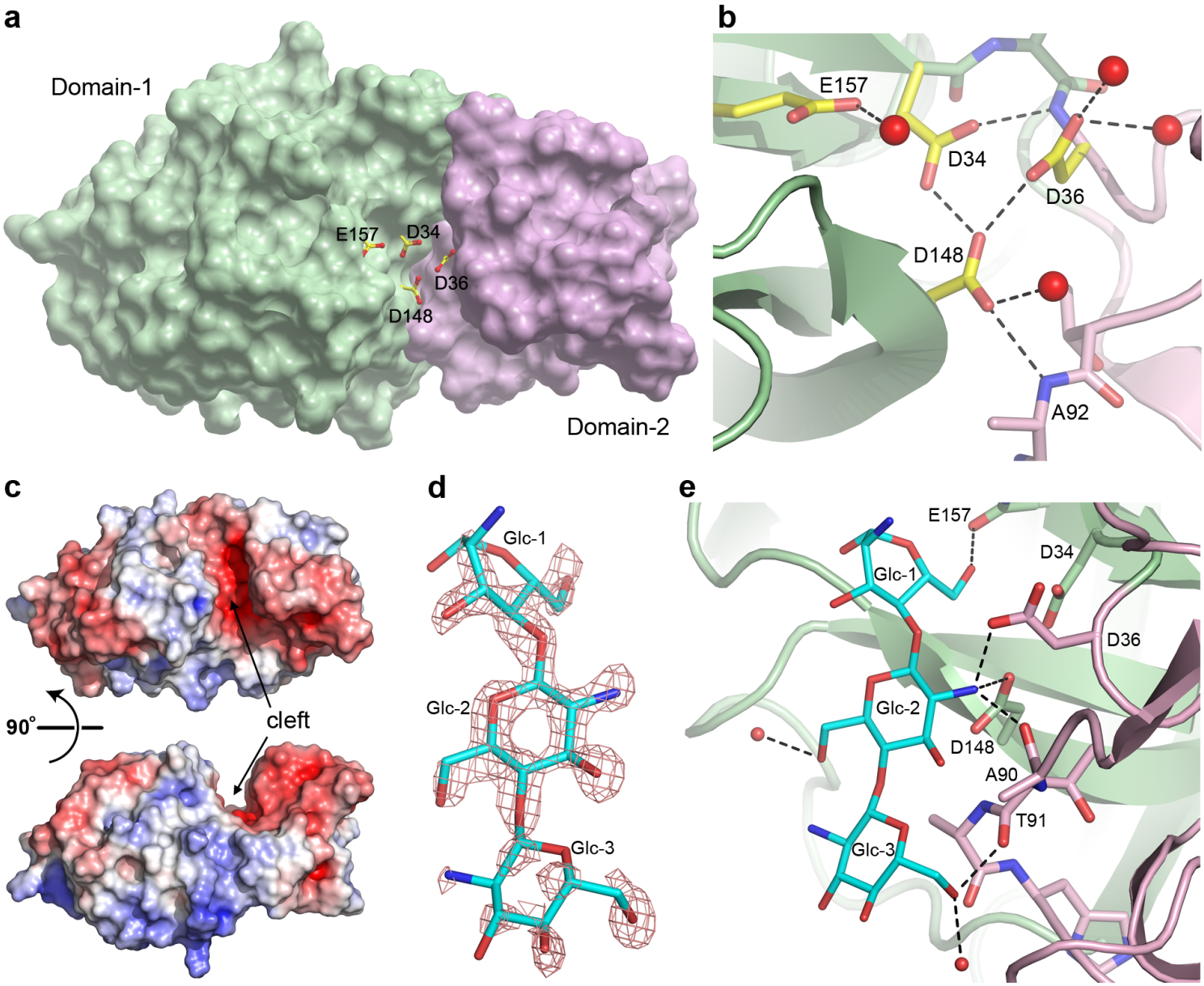

Figure 3

V-Csn active site and substrate complex. a. Solvent accessible surface representation of the V-Csn *apo1* structure showing the location of the four conserved acidic residues (yellow and red sticks) clustering within the inter-domain cleft. The surface is colored as for **Fig.2d** (Domain-1 in light green and Domain-2 in pink). **b.** Close-up view of the putative active site highlighting the four conserved acidic residues. Hydrogen bonds are shown as dashed black lines and water molecules as small red spheres. **c.** Electrostatic surface of V-Csn *apo1* calculated at pH 5.1. The orientation of the molecule in the top view is approximately the same as in panel **a**, and the button view is rotated 90° to show the active site cleft from the side. **d.** Residual F_o-F_c electron density (pink mesh) for the bound substrate contoured at 2.5 σ. The electron density map was calculated following molecular replacement and prior to the incorporation of the substrate into the structure. The final refined trisaccharide molecule (GlcN-1, GlcN-2 and GlcN-3) is shown as cyan sticks. **e.** Ribbon representation of chitohexaose-V-Csn complex showing the trisaccharide (cyan sticks) bound in the active site. Hydrogen bonds are indicated by dashed black lines and water molecules as small red spheres. The ribbon representation is colored as for **Fig.2d** (Domain-1 in light green and Domain-2 in pink).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformationFigure1MSAallTaxaFINAL.pdf
- SupplementalFigure2MSAvsstructureseditedRWCAS.docx
- SupplementaryInformationFig3ValidationReportFINAL.pdf
- SupplementaryInformationTable1.xlsx
- ExtendedDataFig15Table14FINAL.docx