JBC Papers in Press. Published on February 8, 2019 as Manuscript RA118.007182
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.RA118.007182

The N-glycan structures of the antigenic variants of chlorovirus PBCV-1 major capsid protein help to identify the virus-encoded glycosyltransferases

Immacolata Speciale[1], Garry A. Duncan[2], Luca Unione[3], Irina V. Agarkova[4,5], Domenico Garozzo[6], Jesus Jimenez-Barbero[3,7,8], Sicheng Lin[9], Todd L. Lowary[9], Antonio Molinaro[10], Eric Noel[4,11], Maria Elena Laugieri[12], Michela G. Tonetti[12], James L. Van Etten[4,5*], Cristina De Castro[1*]

From the [1]Department of Agricultural Sciences, University of Napoli Federico II, Via Università 100, 80055 Portici, NA, Italy; [2]Department of Biology, Nebraska Wesleyan University, Lincoln, NE 68504-2794 USA; [3] Chemical Glycobiology Lab, CIC bioGUNE, Bizkaia Technology Park, Bld 800, 48170 Derio, Spain; [4]Nebraska Center for Virology, University of Nebraska-Lincoln, Lincoln NE 68583-0900; [5]Department of Plant Pathology, University of Nebraska-Lincoln, Lincoln NE 68583-0722 USA; [6]CNR, Institute for Polymers, Composites and Biomaterials, Via P. Gaifami 18, 95126 Catania, Italy; [7] Basque Foundation for Science IKERBASQUE, 48009 Bilbao, Spain; [8] Dept. Organic Chemistry II, Faculty of Science and Technology, University of the Basque Country, EHU-UPV, Leioa, Spain; [9] Alberta Glycomics Centre and Department of Chemistry, University of Alberta, Gunning−Lemieux Chemistry Centre, 11227 Saskatchewan Drive, Edmonton, Alberta T6G 2G2, Canada; [10] Department of Chemical Sciences, Università of Napoli Federico II, Napoli, Italy; [11] School of Biological Sciences, University of Nebraska, Lincoln, NE 68588-0118, USA;[12] Department of Experimental Medicine and Center of Excellence for Biomedical Research, University of Genova, Viale Benedetto XV/1, 16132, Genova, Italy.

Running title: *Virus encoded glycosyltransferases*

* To whom correspondence should be addressed: prof. James L. Van Etten: Nebraska Center for Virology, University of Nebraska-Lincoln, Lincoln NE 68583-0900, email: jvanetten1@unl.edu. Prof. Cristina De Castro: Department of Agricultural Sciences, University of Napoli Federico II, Via Università 100, 80055 Portici, NA, Italy; decastro@unina.it; Tel. 0039 081674124.

Keywords: DNA viruses, glycosyltransferases; bioinformatics; antigenic variants; N-linked glycosylation; molecular modeling; Paramecium bursaria chlorella virus 1 (PBCV-1); posttranslational modification; capsid protein

ABSTRACT

The chlorovirus Paramecium bursaria chlorella virus 1 (PBCV-1) is a large dsDNA virus that infects the microalga *Chlorella variabilis* NC64A. Unlike most other viruses, PBCV-1 encodes most, if not all, of the machinery required to glycosylate its major capsid protein (MCP). The structures of the four N-linked glycans from the PBCV-1 MCP consist of nonasaccharides, and similar glycans are not found elsewhere in the three domains of life. Here, we identified the roles of three virus-encoded glycosyltransferases (GTs) that have four distinct GT activities in glycan synthesis. Two of the three GTs were previously annotated as GTs but the third GT was identified in this study. We determined the GT functions by comparing the wild-type glycan structures from PBCV-1 with those from a set of PBCV-1 spontaneous GT genes mutants resulting in antigenic variants having truncated glycan structures. According to our working model, the virus gene *a064r* encodes a GT with three domains: domain 1 has a β-L-rhamnosyltransferase activity, domain 2 has an α-L-rhamnosyltransferase activity and domain 3 is a methyltransferase that decorates two positions in the terminal α-L-rhamnose (Rha) unit. The *a075l* gene encodes a β-xylosyltransferase that attaches the distal D-xylose (Xyl) unit to the L-fucose (Fuc) that is part of the conserved N-glycan core region. Lastly, gene *a071r* encodes a GT that is involved in the attachment of a semiconserved element, α-D-Rha, to the same L-Fuc in the core region. Our results uncover GT activities that assemble four of the nine residues of the PBCV-1 MCP N-glycans.

Structural proteins of many viruses, such as rhabdoviruses, herpesviruses, poxviruses and paramyxoviruses, are glycosylated. Most virus glycoproteins are N-linked to Asn via N-acetylglucosamine (GlcNAc), while less frequent O-linked glycosylation also occurs (1). The majority of the viruses studied to date use host-encoded glycosyltransferases (GTs) and glycosidases located in the endoplasmic reticulum (ER) and Golgi apparatus to add and remove N-linked sugar residues from virus glycoproteins either co-translationally or shortly after translation of the protein (2-6). Post-translational glycosylation can aid in protein folding, protein stability, progression in the secretory pathway and host-virus interactions.

One group of viruses that differs from the above scenario is the plaque-forming Chloroviruses (family *Phycodnaviridae*) that infect certain isolates of chlorella-like green algae (7). The viruses are divided into four groups depending on the algal host infected: chloroviruses that infect *Chlorella variabilis* NC64A are referred to as NC64A viruses, those that infect *Chlorella variabilis* Syngen 2-3 are referred to as Osy viruses, those that infect *Chlorella heliozoae* SAG 3.83 are referred to as SAG viruses and those that infect *Micractinium conductrix* Pbi are referred to as Pbi viruses. The prototype chlorovirus Paramecium bursaria chlorella virus (PBCV-1) is an NC64A virus (8) and its 330 kb genome is predicted to encode 416 proteins (9). The major capsid protein (MCP or Vp54) is coded by the gene *a430l*, and its gene product has a predicted molecular weight of 48,165 Da, which increases to 53,790 Da due to N-glycosylation at four positions (10, 11).

The predominant glycoform of the four PBCV-1 N-glycans is a nonasaccharide (Fig. 1 WT) (11) with several unusual structural features including: i) it is not linked to the typical Asn-X-(Thr/Ser) sequon; ii) it is highly branched; iii) β-glucose (Glc) forms the N-linkage with Asn; iv) fucose (Fuc) is substituted at all available positions; and v) there are two rhamnose (Rha) residues with opposite configurations plus an L-Rha capped with two O-methyl groups. Two monosaccharides, arabinose (Ara) and mannose (Man), occur as non-stoichiometric substituents resulting in the existence of four glycoforms (Fig. 1). Currently, structures of this type are unknown elsewhere in the three domains of life.

The discovery of these glycans has prompted our interest in viral glyco-related genes. In this regard, 17 of the PBCV-1 genes encode enzymes that manipulate sugars at different levels (12) with six of them annotated as encoding putative GTs: *a064r* (638 amino acids), *a111/114r* (860 aa), *a219/222/226r* (677 aa), *a473l* (517 aa), *a546l* (396 aa), and *a075l* (280 aa) (Table S1). In addition, *a098r* encodes a functional hyaluronate synthase (13).

The genes encoding these GTs are scattered throughout the genome of the virus and PSORT predicted that most of these proteins are soluble and located in the cytoplasm; *a473l* and *a219/222/226r* are predicted to encode GTs with six and nine transmembrane domains, respectively, while the ngLOC program predicts that A075L is located in the chloroplast, but only at a 15% confidence level.

Thus, PBCV-1 encodes six GTs, a small number when compared to the complexity of its N-glycans. One solution to this paradox is that one or more of the six GTs might have multiple functional domains, making this restricted repertoire of enzymes sufficient to synthesize these structures. A second possibility is that some other virus genes encode enzymes that do not match GTs in the databases and so are overlooked during the searches. Finally, we cannot exclude the possibility that a host-encoded GT(s) could be involved in the glycosylation process.

The protein encoded by gene *a064r* is of special interest. This protein is coded by PBCV-1 but not by any other chlorovirus where the glycan structure is known (12), suggesting that the *a064r* gene product is involved in the synthesis of a structural motif of the N-glycan not shared with the other chloroviruses (14-16). In addition, several spontaneous mutants (or antigenic variants) in *a064r* have been isolated and their MCPs migrate faster on SDS-PAGE compared to the MCP of PBCV-1, suggesting that the glycan portion of the capsid protein is altered.

These PBCV-1 spontaneous mutants are divided into six antigenic classes, each denoted with a letter, based on their differential reaction to five different polyclonal antibodies (12, 17). Each class reacts specifically with one polyclonal antibody preparation, except Class E, which cross-reacts with Class A and B polyclonal antibodies, suggesting that the phenotype of Class E variants share some of the structural features of the other two variants. Notably, all six antigenic classes, except C, have a mutation in the a*064r* gene (Table 1, Fig. 1).

For the reasons cited above and elsewhere (12, 18), we hypothesize that the machinery for PBCV-1 glycosylation is encoded primarily, if not totally, by the virus. Furthermore, glycosylation of the PBCV-1 MCP occurs in the cytoplasm. The current study addresses two issues related to PBCV-1 glycosylation. First, the structures of the glycans attached to five PBCV-1 antigenic variant MCPs are described. Second, using genetic and glycan structure evidence, we identify four of the putative GT activities involved in the synthesis of the glycan(s).

Results

Glycan structures from representatives of the antigenic variants

The structures of the glycans from five different classes (Class E variants were not included in this study) were determined by combining chemical (via GC-MS of the appropriate derivatives) and NMR analyses, using the consolidated approach developed for PBCV-1 and the other chloroviruses (11, 14-16). The ^{1}H NMR spectrum of each variant (Fig. 2) differed from that of wild-type virus PBCV-1, supporting our hypothesis that the glycans were different. Accordingly, the features of each antigenic class will be described in increasing order of the structural complexity of the glycans. By convention, the monosaccharide residues identified in each oligosaccharide are labeled with the same (or a similar) letter used for the analog unit of the glycan of the wild-type (WT) virus PBCV-1 (Fig. 1 and Figs in the supporting material). Unless mentioned otherwise, we have focused on the major glycan structure of each of the variants. In some cases, minor structures made up less than 5-10% of the total sample.

Class D: structure of the glycan of the antigenic variant P1L6

This virus is the most difficult to handle in laboratory conditions because it is unstable and tends to degrade during normal purification procedures. This intrinsic fragility of the viral particles hindered the obtaining of large amounts of biomass. Thus, to avoid further losses during chromatographic purification, the glycopeptide mixture was studied directly omitting this purification step.

Inspection of the ^{1}H NMR spectrum of the P1L6 glycan (Fig. 2, Fig. S1) showed two anomeric signals at 5.61 and 5.28 ppm along with a set of small signals at ca. 5.0 ppm, while the other regions were crowded with signals from peptides resulting from capsid protein digestion.

The structure of the glycan was determined by combining the chemical data with the analysis of the Heteronuclear Single Quantum Correlation (HSQC) spectrum. GC-MS analysis of the partially methylated and acetylated alditol (PMAA) derivatives (not shown) indicated that the sample contained a 2-linked Fuc, a terminal Gal, a terminal Xyl, and a 3,4-linked Glc. The HSQC spectrum (Fig. S1, Table S2) had four anomeric resonances, and the overlap with the spectrum of the PBCV-1 glycan, used as a reference, permitted the identification of three of the four residues. As for **H** (3,4-substitued N-linked β-Glc), **E** (terminal α-Gal) and **N** (terminal β-Xyl), the ^{13}C NMR chemical shifts were almost identical to that of the reference glycan, with only minor variations in the ^{1}H NMR chemical shifts. The remaining carbon resonances were assigned to Fuc **A**, whose values differed from those of the reference glycoside (19) due to glycosylation at C-2. Thus, the N-linked glycan of P1L6 is a tetrasaccharide, significantly truncated compared to that of WT PBCV-1. It contained all the units of the chloroviruses N-glycans present in the conserved core region except the Xyl attached to Fuc, referred to as the distal Xyl (14).

Class C: structure of the antigenic variants E11 and E1L3

The identity of the E1L3 and E11 glycans was deduced from the similarities between the anomeric region of their proton spectra (Fig. S2A) and by the coincidence of their HSCQ spectra (Figs. S2B, S3). In addition, when compared to P1L6 HSQC (Fig. S1), both E11 and E1L3 had one additional anomeric signal at 5.03 ppm.

As for E1L3, comparison of its HSQC spectrum with those of PBCV-1 glycopeptides, permitted the ready identification of several monosaccharide units (Table S3): the N-linked Glc **H**, the proximal Xyl **N**, and the Gal **E.** The anomeric proton at 5.03 ppm was identified by analyzing the other 2D NMR spectra including correlation spectroscopy (COSY), total correlation spectroscopy (TOCSY) and transverse rotating frame Overhauser effect spectroscopy (T-ROESY). These data revealed that this resonance arose from a D-Rha unit linked at O-3 of the Fuc unit **A.** Unlike the wild-type glycan, this D-Rha was not further substituted with the terminal Man **G**; therefore, it was labelled **F'** (and not **F**). Thus, E1L3 and E11 glycans are extended derivatives of P1L6 with the D-Rha unit. Accordingly, the Fuc residue was substituted at both O-2 and O-3 positions, in agreement with the GC-MS linkage analysis of this fraction (Fig. S2C).

Class B: structure of the glycan of antigenic variants EPA1 and EPA2

The ^{1}H NMR (Fig. S4A) and HSQC (EPA2 reported as an example in Fig S4B) spectra of EPA1 and EPA2 glycopeptides were almost identical, and both displayed two anomeric signals at 4.5-4.4 ppm, the region that in the PBCV-1 wild-type glycan contains the two β-Xyl units **M** and **N**.

The units with anomeric signals at 5.63, 5.23, 5.04, 5.01 and 4.42 ppm were readily attributed to the Fuc **A**, Gal **E**, D-Rha **F'**, Glc **H** and to the proximal Xyl **N**, respectively, by comparison with PBCV-1 glycan data. Notably, the proton and carbon chemical shifts of Fuc **A** (Table S4) indicated that this Fuc was substituted at all the positions available, while D-Rha **F'** was not glycosylated.

Hence, NMR analysis focused on the additional anomeric signal at 4.46 ppm. Studying the connectivities from all the 2D NMR spectra revealed that this signal belonged to a Xyl unit, connected to O-4 of **A** as in the PBCV-1 WT glycan, but not further substituted. Accordingly, this Xyl was labeled **M'**. Overall, the NMR data showed that the structure of the glycans of the antigenic class B, EPA1 and EPA2, consisted of the conserved core region characteristic of all the chloroviruses N-glycans described to date (14-16) plus the semi-conserved element, the D-Rha unit **F'** (Fig. 1).

Class A: structure of the glycans of antigenic variants P91, P9L1 and P9L10.

The glycopeptides from the three variants of Class A shared a similar pattern of signals in the anomeric region of the ^{1}H NMR spectra, although some marginal variations of their relative intensities were noted (Fig. S5A). Compared to Class B, the anomeric region of the ^{1}H NMR spectra of class A contained several additional signals, and the structural elucidation of these glycopeptides was performed by the combined analysis of the 2D NMR spectra (Figs. S5-7 and Tables S5-6) as discussed for the other variants, and detailed in the Supporting Information.

Indeed, P9L10 produces a family of four glycopeptides, named P9L10$_{A-D}$ (Fig. S6), with the distal Xyl **M** (or **M"**, see discussion in Supporting Information) further elongated with the β-L-Rha unit. These four glycans differed in the non-stoichiometric presence of two residues, Man and Ara, as shown in Fig. S6.

As for the P9L1 glycopeptides, the HSQC spectrum was almost superimposable on that of P9L10. Some variations occurred for the intensities of some signals as deduced by comparing their proton spectra (Fig S5). In P9L1, the intensities of **E** and **E'** were comparable suggesting that D-Rha was much less substituted with Man compared to P9L10 glycopeptides. Densities reporting the Ara (**D** unit) were not detected, indicating that this unit was absent or below the detection limit. Accordingly, P9L1 produced two main glycopeptides, analogues to P9L10$_A$ and P9L10$_B$ (Fig. S6).

As for the glycopeptides of P91 (Fig. S7, Table S6), most of the densities of its HSQC spectrum (Fig. S7A, B) overlapped with those of the P9L1 and P9L10 variants. Like P9L1, P91 variant lacks the Ara substitution to β-L-Rha, and produced two oligosaccharides equivalent to P9L10$_A$ and P9L10$_B$, as confirmed by analyzing the homonuclear spectra (T-ROESY in Fig. S7C and TOCSY in Fig. S7D). Importantly, these spectra enabled the assignment of densities that escaped the HSQC detection due to the low amount of the sample, including C-4 or C-5 of **M''** (the putative position of these densities is enclosed in a dotted circle in Fig. S7B).

The P91 HSQC spectrum contained some additional signals, including a methyl group at $^1H/^{13}C$ 3.56/62.4 ppm (Fig. S7A) and an anomeric proton at $^1H/^{13}C$ 4.75/103.4 ppm (Fig. S7B). NMR analysis of this new residue determined that it was a β-L-Rha, labelled as **I"**, that differed from **I** and **I'** because it was methylated at O-2 (Table S6). Methylation at O-2 did not induce relevant variations in the chemical shift values of all the other monosaccharides of the glycan, except for H-4/C-4 values ($^1H/^{13}C$ 3.71/80.5 ppm) of the Xyl unit to which it was attached, named **M'"**. For this new oligosaccharide (named P91, Fig. S7) it was not possible to confirm if the D-Rha unit **F** was further elongated with Man **G**, although this situation is the most plausible given the abundance of this residue, similar to that of Gal or Glc, as visible in the GC-MS profile of the monosaccharide analysis (Fig. S5C). GC-MS analysis also confirmed the presence of the rhamnose unit methylated at O-2, and detected a tiny amount of arabinose, that escaped the NMR analysis due to its low abundance.

Class F: structure of the glycan of antigenic variant CME6

Virus CME6 is the only variant isolated for antigenic class F. Its proton and HSQC spectra (Figs. 2 and S8) contained an additional anomeric signal at $^1H/^{13}C$ 5.04/102.7 ppm (Fig. S8) when compared to those of the P9L10 variant. Comparison of the CME6 HSQC spectrum with that of PBCV-1 readily identified most of the residues, namely **A**, **D**-**H**, **M** and **N**, for which the positions of all the proton and carbon resonances were almost identical. Of note, units **E'** or **F'** described in class A variants were absent or below the NMR detection limit indicating that D-Rha was fully substituted with Man. Minor differences were found for the chemical shifts of the residues **I** and **L**, the β-L-Rha units 2- and 2,3-linked, respectively.

Analysis focused on the signal **C** at $^1H/^{13}C$ 5.04/102.7 ppm, which had a weak H-1/H-2 correlation in the COSY spectrum (not shown), while TOCSY spectrum read from H-2 (4.09 ppm) position identified the position of all the other protons of the unit, with H-6 at 1.26 ppm. Thus, **C** was a Rha, and carbon chemical shift values (Table S7) determined that it was α-configured at the anomeric center and not further substituted. HMBC and T-ROESY spectra showed that it was attached to O-2 of **I**, as in the wild type virus, with the only difference that none of its positions were further decorated with a methyl group (Fig. 1).

As for other minor anomeric signals, it was possible to assign only that at $^1H/^{13}C$ 5.15/101.8 ppm (Fig. S8), labeled **B**. It was an α-Rha, that differed from the most intense **C** because it was attached to O-2 of **L**, the β-Rha unit with **D** at O-3.

Collectively, these data indicated that CME6 produces a mixture of two N-glycans (Figs. 1 and S8), identical to those described for WT PBCV-1 with the only difference being that the α-L-Rha unit was not methylated.

Predicted PBCV-1 encoded genes involved in the synthesis of its glycan

Based on the discovery that the structures of the glycans from the five antigenic classes were truncated versions of the glycan from the WT PBCV-1 MCP and knowing that PBCV-1 already had several genes annotated as GTs (12, Table S1), allowed us to predict the function of some of them. These predictions were aided by the fact that the complete genomes of a few representatives of each of the four groups of chloroviruses mentioned in the introduction (NC64A, Osy, SAG, and Pbi viruses) have been sequenced (7, 15) and the structures of the glycan(s) attached to their MCPs have been determined (14-16). The basic glycan core structure is conserved in all of these viruses (outlined in Fig. 1). Additional sugar decorations occur on their respective core N-glycan structures and present a molecular signature for each chlorovirus. Therefore, knowledge of the glycan structures and the presence/absence of genes from the various groups of chloroviruses contributed to our prediction of the role for two of the PBCV-1 GTs. The features of each putative GT are hereafter discussed in light of the structural information of the glycans determined in this study.

Class D mutants and the role of A111/114R in the core glycan assembly/attachment

The four antigenic mutants in Class D (P1L2, P1L3, P1L6 and P1L10) contain large genomic deletions in similar locations that range in size from 27.2 kb to 38.6 kb (genome nucleotide locations: 12,407 to 39,576; 5,238 to 42,695; 2,190 to 40,797 and 7,698 to 39,722, respectively), or collectively from genes *a014r* through *a078r* (20). All of the chloroviruses (14-16), as well as four of the five PBCV-1 antigenic classes, have the conserved glycan core structure composed of five monosaccharides (one unit of D-Glc, D-Gal and L-Fuc, and two units of D-Xyl) attached to the MCP (Fig. 1), plus D-Rha, the semiconserved unit attached to Fuc.

Class D mutants are the only class with a reduced core glycan structure (Fig. 1) and this finding supported two conclusions: i) the core structure is coded by a gene (or genes) located outside of the collective region of the four large deletion mutants; ii) the genes encoding the attachment of the distal D-Xyl, the fifth unit of the conserved core region, and the D-Rha, the semiconserved unit of the core glycan, probably lie in the region between *a014r* and *a078r*.

Accordingly, these two observations suggest that the protein A111/114R (NP_048459.2) is involved in the initiation and/or attachment of four of the monosaccharides in the conserved core glycan. Further evidence to support this inference is that this gene and the corresponding tetrasaccharide are present in all the chloroviruses analyzed to date. Because the chlorovirus core glycan is unique in nature, we assume it must be at least partially, if not fully, assembled/attached to the MCP by either a single large viral protein or several smaller viral proteins that are common to all the chloroviruses. In PBCV-1, gene *a111/114r* is the only annotated orthologous GT gene found outside of the regions of the large deletion mutants described above and which is present in all the other chloroviruses. In addition, no MCP glycans smaller than the four-sugar units have been found in any of the chloroviruses (14-16) or in the PBCV-1 antigenic variants (this manuscript).

The *a111/114r* gene, not including the promoter region, consists of 2583 nts, which is longer than any of the other annotated PBCV-1 GT genes. This gene encodes protein A111/114R, which consists of 860 amino acids organized into at least three domains; the second domain has been annotated as a GT (9).

Because of the size of the *a111/114r* gene, it is highly likely that mutations have occurred in the gene over time, but such mutations are probably lethal because no variant has been isolated. The working model is that a mutation in the *a111/114r* gene, and in the orthologous genes from all of the chloroviruses, would result in the absence of a glycan core attached to the MCP, leading to the loss of a viable virus. This leads to the conclusion that the unique four-sugar structure is the minimal glycan required for chlorovirus survival, although the way it contributes to the viability of the virus is still unclear. Certainly, the hydrophilic nature of the glycans would help the virus to remain soluble, and thus infective, in an aquatic environment. Furthermore, as noted above, Class D variants are unstable in the laboratory, suggesting that the glycans may be important for MCP folding and/or for the correct assemblage of the capsid.

Thus, all available evidence supports the conclusion that A111/114R plays a key role in the assemblage of part or all of the conserved core oligosaccharide.

Class C mutants and the role of A071R

The glycans of the Class C mutants (Fig. 1) differ from those of Class D because they have a D-Rha attached to L-Fuc; however, like the Class D mutants, the Class C mutants lack the distal D-Xyl and its attached sugars. The gene encoding the enzyme involved in the attachment of D-Rha to L-Fuc is located somewhere in the region between genes *a005r* through *a077l* because the Class D antigenic mutant P1L6, which lacks the D-Rha, has no mutations in its genome other than the large deletion that extends from genes *a005r* through *a077l*. Therefore, the gene that encodes the protein that performs this task must: i) lie between genes *a005r* and *a077l*, ii) be found in other NC64A viruses and OsyNE5, all of which have a D-Rha attached to the L-Fuc, and iii) must be absent in SAG and Pbi viruses because they lack a D-Rha in that position. The only PBCV-1 gene that fits all of these requirements is *a071r*. The protein encoded by this gene, A071R (NP_048419.1), is of sufficient size (354 aa) to be a GT although it has not been annotated as such.

However, A071R does have sequence similar to A064R domain 2, which also has not been annotated as a GT but for which there is strong evidence that it is a rhamnosyl GT (discussed below).

Class C mutants and the role of A075L

Class C antigenic variants also have mutations that prevent the addition of the distal D-Xyl, which is necessary before the addition of the two L-Rha units. As noted above in the discussion of the Class D large deletion mutants, the protein that attaches the distal D-Xyl to the L-Fuc must be encoded by a gene that lies between the genes *a014r* and *a078l*. The genetic evidence supports the hypothesis that the protein A075L (NP_048423.1) attaches D-Xyl to L-Fuc. This protein was previously annotated as an exostosin (Pfam database code PF03016.8), classified within the GT47 family. Six Class C antigenic mutants were genomically sequenced (E1L1, E1L2, E1L3, E1L4, P41 and E11), and all six have mutations in the *a075l* gene, but none of them shared mutations in any of the other genes that putatively encode GTs. The mutations in the gene *a075l* were of three types: i) frameshift mutations that cause premature termination of the synthesis of the A075L protein; ii) nonsense mutations that lead to premature termination of A075L protein synthesis; and iii) in one case a missense mutation in the initiator codon that results in a failure to initiate A075L protein synthesis. One of the Class C mutants, E11, is of particular interest. E11 is derived from the Class B antigenic mutant EPA-1, but the mutation common with EPA-1 became masked when E11 acquired a second mutation, located in gene *a075l*, which prevents the addition of the distal D-Xyl, the feature that distinguishes the Class C mutants from those in Class B.

Additional evidence that supports the role of A075L comes from seven other chlorovirus species whose MCP glycan structures have been determined (14-16). Even though these seven chloroviruses infect four different algal hosts, all of them have the same glycan core structure and the same distal D-Xyl attached to the L-Fuc, as does the wild type PBCV-1. These seven chloroviruses all encode orthologs of the PBCV-1 A075L protein.

Class B mutants and the role of A064R domain 1 (A064R-D1)

Class B antigenic variants lack the terminal two L-Rha moieties (Fig. 1) but have the terminal D-Xyl because they all have a functional *a075l* gene; however, they have mutations that prevent further elongation of the oligosaccharide chain. Three Class B antigenic variants have been genomically sequenced (EPA-1, EPA-2, and P31), and all have mutations in the *a064r* gene, which encodes the protein A064R (NP_048412.1). This protein consists of 638 amino acids that are organized into at least three functional domains of approximately 200 amino acids each (Fig. 1). More specifically, all three Class B antigenic variants have mutations in the region of the *a064r* gene, that encodes domain 1 of A064R.

Previous studies annotated this domain as a "fringe class" GT and hypothesized that the DXD motif coordinated the phosphate of the nucleotide sugar, presumably UDP-Glc, and a divalent cation, Mn^{2+} (21, 22). The specificity of this domain could not be inferred at that time because the structure(s) of the N-glycans were unknown, preventing any hypothesis about the nature of the possible acceptor and donor substrates.

Understanding the function of domain 1 was inferred by comparing the N-glycans of Class B with those of Class A variants. These two sets of N-glycans (Fig. 1) differ in the presence of the β-L-Rha unit **I** (or **L**), supporting the hypothesis that the first domain of A064R encodes a β-L-rhamnosyltransferase that adds this monosaccharide to O-4 of a Xyl unit **M**.

As a possible nucleotide donor, we hypothesize that A064R-D1 uses UDP-L-rhamnose as substrate, while most prokaryotic enzymes recognize the dTDP-bound sugar. This difference can be explained by the fact that, in contrast to GDP-L-fucose or GDP-D-Rha that can be produced by virus derived enzymes (23), a UDP-L-rhamnose biosynthetic pathway is not encoded by the PBCV-1 genome. Thus, the nucleotide activated L-rhamnose is presumably synthesized by the chlorella host, using a typical plant UDP-L-rhamnose biosynthetic pathway (24).

This hypothesis is corroborated by both docking studies (see below) and by biochemical evidence that A064R domain 1 has this predicted activity.[1]

Class A mutants and the role of A064R domain 2

Class A antigenic variants only lack the terminal α-L-Rha. Three variants of this antigenic class have been genomically sequenced (P91, EPA-11 and P9L1), and all of them lack mutations in domain 1 of *a064r*, but all three have mutations in domain 2 of this gene (Fig. 1), which encodes a bacterial-derived protein of unknown function. Furthermore, none of these three variants has mutations in the gene that encodes A075L. Thus, these three members of class A are able to add the terminal D-Xyl and the β-L-Rha to the core glycan (Fig. 1). Two of the three mutants, P9L1 and EPA-11, have nonsense mutations (point mutations that become stop codons) that lead to premature termination of translation. The third mutant, P91, has a 27-nucleotide insertion in the region of *a064r* that encodes domain 2. In addition, the *a064r* gene of another class A mutant, P9L10, was amplified by PCR and sequenced. P9L10 acquired a nonsense mutation in *a064r* that results in premature termination of A064R near the N-terminus of domain 2; consequently, the functions of both domain 2 and 3 are lost.

In summary, all of the class A variants encode a functional domain 1 of A064R but are unable to encode a functional domain 2. These results are consistent with the glycan structure results, which support the hypothesis that domain 2 is involved in the attachment of the second Rha unit (**C** or **B**), the one with an α glycosidic linkage and placed at O-2 of the first L-Rha unit (**I** or **L**). For the reasons mentioned for A064R-D1, our working hypothesis is that the second domain of A064R accepts UDP-L-Rha as donor.

Class F mutant and the role of A064R domain 3

The antigenic variant CME6, whose genome has been sequenced, is the only member of antigenic class F (Table 1). The CME6 *a064r* gene has a TC dinucleotide insertion at genome position 36,276, which leads to premature chain termination during translation of domain 3, while the first two domains are unaffected. Accordingly, the N-glycans of this variant contain all the sugars that comprise the wild-type glycan but lack the methyl groups at the O-2 and O-3 positions of the terminal Rha units **C** or **B** (Fig. 1). This finding confirms the role of the first two domains of A064R, namely the first domain attaches β-L-Rha (**I** or **L**) to β-Xyl **M** and domain 2 attaches the last Rha unit (**C** or **B**) to the β-L-Rha.

While PBCV-1 encodes numerous DNA methyltransferase genes, the only mutated methyltransferase gene in CME6 is in the region of gene *a064r* that encodes domain 3 (Fig. 1). The best hit in a BLAST search of the wild type amino acid sequence of domain 3 to the nucleotide database at NCBI was to 2'-O-rhamnosylmethyltransferases. Indeed, when the rhamnosyl-methyltransferase from *Mycobacterium smegmatis* (25), was BLASTED against the PBCV-1 protein database, the only hit was to domain 3 of A064R (33% identity and 49% positive). Consequently, when *a064r* is mutated in the third domain, the corresponding N-glycan is not methylated at either the O-2 or O-3 positions, suggesting that this methyltransferase can perform a double methylation of the monosaccharide. However, we cannot exclude the possibility that domain 3 attaches only one methyl group and that a second methyltransferase, yet to be identified, is required to complete the full decoration of this Rha.

The mutant P91, a Class A variant, provides additional support that domain 3 is a rhamnosyl-methyltransferase. In the case of P91, domain 3 of A064R is translated correctly because the 27-nucleotide insertion in the second domain of *a064r* maintains the correct reading frame during its translation. Analysis of the P91 glycopeptide mixture detected a minor form with β-L-Rha methylated at O-2 (**I"**, Fig. S7). Thus, domain 3 appears to be functional in P91, but it lacks its natural substrate because the α-L-Rha unit **C** (or **B**) is not assembled at the end of the glycan (see PBCV-1 structure in Fig. 1 and P91 glycan structure in Fig. S7) but there is a β-L-Rha for which it may have a reduced affinity.

Modeling of UDP-Rha into the catalytic domain 1 of A064R

[1] De Castro et al., work in progress

Based on the evidence collected in this study, A064R-D1 is predicted to use UDP-L-Rha as the activated donor to synthetize the β-L-Rha-(1,4)-β-D-Xyl linkage. In a previous study (22), this domain was cloned and crystallized and the authors tested several nucleotide sugars, but not UDP-L-Rha. The best affinity was found for UDP-Glc. Therefore, a molecular modeling approach was used to determine whether UDP-L-Rha is an acceptable substrate for this GT.

Docking methods are widely used to define or predict the binding mode of a small ligand to a receptor. Although these methods are efficient in discriminating between binders and non-binders, they are not particularly accurate and typically cannot discriminate between substrates that differ by less than one order of magnitude in affinity (26). This is the case for the possible Rha/Glc relative selectivity for A064R-D1. Subtle differences originate from contributions of hydrogen bonds and CH/π staking interactions. Thus, a molecular dynamics (MD) approach was used (27, 28). The structure deposited in the protein database (PDB accession number 2P72) contains UDP-Glc as substrate. Therefore, its behavior was compared to that computed for the UDP-L-Rha analogue.

In the X-ray structure, the donor UDP-Glc is stabilized in the binding site by several hydrogen bonds and Mn^{2+}-oxygen interactions. Specifically, upon ligand binding, the protein undergoes a major conformational change, which shifts Phe13 towards the uracil ring of UDP. Two hydrogen bonds between the Phe13 backbone atoms and the uracil ring are established. These same hydrogen bond interactions are established (compare Figs. 3A,D and 3B,E) in the model of A064R-D1 bound to UDP-L-Rha and kept throughout the entire simulation time (Fig. S9). Similarly, the hydrogen bond network described for the ribose moiety in the X-ray structure is also maintained during the whole MD simulation (Figs. 3B,E and S9). Specifically, the carbonyl oxygen of Gly11 interacts with the OH2 of ribose, while the NH of Ser79 is a hydrogen bond donor to O3 of ribose. The hydrogen bond between the carboxyl side chain of Asp78 and the OH3 of ribose is established only in the last 25 ns of the MD simulation (Fig. S9).

It has previously been described that His54 experiences a significant conformational change upon UDP-Glc binding (Fig. 3C), suggesting a key role for this residue as the required catalytic base (22). Along this line, the MD simulation of A064R-D1 with UDP-Rha clearly identifies a stacking interaction among the histidine electron cloud and the C-H moiety of the less polar face of Rha (Fig. 3F). This CH/π interaction is stable in all the MD simulations (Fig. S9). These findings are in full agreement with the previous hypothesis (22) about the role of Nε in His54, which stabilizes the partial positive charge that develops during the course of the reaction at the C1 atom. In fact, the C1-N distance is only 4.5 Å on average (Figs. 3F and S9). Finally, UDP-L-Rha is kept in the catalytic site by a polar network involving Arg57 and Pro149, which makes hydrogen bonds with OH4 of the sugar ring (Figs. 3F and S9).

Next, the free energy of binding for both UDP-L-Rha and UDP-Glc to A064R-D1 was estimated by MM/GBSA methods (25, 26). Interestingly, the computations predict that the binding free energy for the A064R-D1/UDP-L-Rha complex is ca. 6 kcal/mol lower than that for the A064R-D1/UDP-Glc analogue (Table 2), in agreement with the prediction of the specificity of this GT, and our preliminary experimental results.[2] Among the key components of the binding free energy, UDP-Glc displayed the largest electrostatic energy changes upon binding (ΔE_{ELE}) and also showed the largest polar solvation energy (ΔG_{Sol}). Thus, the computational analysis strongly suggests that the major factor favoring the binding of A064R-D1 for UDP-L-Rha over UDP-Glc is the lower polar character of Rha with respect to Glc. These results are also in agreement with the low polarity of the A064R-D1 binding site defined by Pro149, Phe148 and Tyr150, which perfectly fit with the methyl substituent of Rha through hydrophobic interactions. Furthermore, we investigated why the antigenic variant EPA-1 lacks the terminal L-Rha moieties, by docking the A064R_S79L mutant with UDP-Rha substrate. The results of

[2] De Castro, work in progress

the analysis show that the bulky side chain of the Leu79 entails the reorganization of the DXD motif, with a consequent loss of Mn^{2+} coordination and most of the essential interactions with the substrate. Specifically, although the hydrogen bonds between the Phe13 and the uracil ring are preserved (Fig. 3G), the hydrogen bond network for the ribose moiety is completely abrogated (Fig. 3H). Even more, His54 is now ~90° flipped precluding CH/π interactions (Fig. 3I). These structural data provide an explanation for the observed truncated oligosaccharide chain in EPA-1 antigenic variant.

Discussion

Unlike the great majority of viruses, the chlorovirus PBCV-1 encodes most, if not all, of the machinery to glycosylate its MCP. In this study, the truncated glycan structures from a set of PBCV-1 antigenic variants were determined. These structures plus genetic evidence led to the identification of four GT activities encoded by three PBCV-1 genes: *a064r*, *a071r* and *a075l*, and enabled us to predict their functions. The *a064r* gene encodes a 638 amino acid protein with three domains: domain 1 is a putative β-L-rhamnosyltransferase, domain 2 is a previously unknown GT, here defined as probably being an α-L-rhamnosyltransferase and domain 3 is a methyltransferase that decorates one or both positions of the terminal α-L-Rha unit. Therefore, we predict that the A064R protein produces the disaccharide 2,3-diOMe-α-L-Rha-(1,2)-β-L-Rha attached to O-4 of the distal Xyl unit (Fig. 4). Gene *a071r* is predicted to encode a GT that attaches a semiconserved α-D-Rha to O-3 of the L-Fuc located in the N-glycan core region (Fig. 4). Gene *a075r* is predicted to encode the β-xylosyltransferase that attaches a D-Xyl unit to O-4 of the same L-Fuc unit used as a substrate by A071R (Fig. 4). Finally, the putative GT(s) encoded by the *a111/114r* gene is predicted to be involved in the synthesis of four of the five residues composing the conserved glycan core structure (Fig. 4), because it is one of the few annotated GT encoding genes that is present in all the chloroviruses. However, it is not known if A111/114R transfers one or more of the sugars of the tetrasaccharide described in variant P1L6 (Fig. 1). We cannot exclude the possibility that at least one or more other unknown PBCV-1 encoded proteins besides A111/114R is involved in the assembly/attachment of the core glycan structure to the MCP. Alternatively, a host-encoded enzyme(s) could also be involved in the synthesis/attachment of the glycan core structure.

Viral encoded glycosylation is an emerging concept in virology, and it is not confined to virus PBCV-1. Other chloroviruses possess the genetic tools to accomplish this task and the same scenario probably occurs in other taxonomically unrelated viruses. Indeed, bovine herpesvirus and myxoma virus encode a β-(1,6)-GlcNAc- and a α-(2,3)-sialyl-transferase, respectively (29). In addition, we predict that many of the giant viruses, characterized by large particle size and very large genomes (30), will encode at least some, if not all, of the enzymes needed to glycosylate their glycoproteins.

The list of giant viruses includes all the known members of the *Mimiviridae* and *Pandoraviridae* families, along with *Pithovirus* and *Mollivirus sibericum* (30), and many others that are rapidly being discovered.

The giant viruses mentioned above differ in morphology, genome size and physical size, but they share a common trait in that their genomes have several glycogenes, i.e., genes encoding enzymes able to manipulate sugar-nucleotides and GTs. Past work has demonstrated that Mimivirus encodes three of the enzymes needed to synthetize UDP-GlcNAc (31), along with a functional pathway for UDP-Rha (32) and for UDP-N-Acetyl-viosamine production (33, 34), and a functional glycogenin paralogue (35). For many of the giant viruses, except for the glycogenin of Mimivirus (35), the activities of the putative glycogenes are yet to be determined.

If our prediction about giant viruses is correct, investigations of these organisms will provide the field of glycobiology with new discoveries and may ultimately answer the central question about what benefits these viruses have by maintaining their own glycosylation machinery through evolution.

Finally, interest in the new GTs described in this report goes beyond the viral world, as they have similarities with prokaryotic and eukaryotic enzymes as exemplified by A064R-D1 for which the best matches are with LgtA, a prokaryotic protein, and with a bovine α-(1,3)-galactosyltransferase (22).

Experimental procedures

Virus growth and glycopeptide isolation from major capsid proteins

Antigenic variants of PBCV-1 (P9L10, P9L1, P91, EPA-1, EPA-2, E11, E1L3, P1L6 and CME6) were grown by infection of 4 L of *Chlorella variabilis* NC64A cells and purified as previously described (36) with some modifications: lysates were exposed to 1% Triton X-100 before two successive rounds of gradient centrifugation in 10-40% sucrose. The virus was treated with proteinase K (0.02 mg/ml) between the sucrose gradients to remove external contaminating proteins. The isolation of their MCPs followed an established protocol (16). Isolated amounts of MCP ranged from 0.2 mg to 1.5 mg. To isolate the glycopeptides, the MCP from each variant was dissolved in water and digested with proteinase K (Sigma code P6556). The digestion process was conducted three times over a time interval of 36 hr using proteinase K at ~20% by weight of the starting amount of the MCP. The glycopeptide mixture was purified via size exclusion chromatography as reported (11). Each antigenic variant produced one main glycopeptide fraction, eluting at approximately 50% of the column volume, with the exception of virus E1L3, which produced two glycoforms. The less abundant form was about 5-10% of the total mixture; its proton spectrum was similar to that of the WT virus and it was not investigated further in this study. For the P1L6 virus variant, due to the low amount of the virus available, the purification step was omitted to avoid further losses. This glycopeptide was analysed directly via NMR spectroscopy; signals arising from the non-glycosylated peptides did not hinder the analysis of the glycan portion.

Glycopeptide chemical analysis

Partially methylated acetylated alditols, and fully acetylated alditols were generated from the purified glycoprotein (ca. 0.2 mg) or on the glycopeptides according to published protocols (37). Analyses of these derivatives was carried out via GC-MS on an Agilent Technologies Gas Chromatograph 6850A interfaced with a selective Mass Detector 5973N equipped with a SPB-5 capillary column (Supelco, 30 m x 0.25 mm i.d., flow rate 0.8 mL min-1 using He as the carrier gas). Electron impact mass spectra were recorded with an ionization energy of 70 eV and an ionizing current of 0.2 mA. The following temperature program was used: 150°C for 5 min, 3°C/min up to 300°C, 300°C for 5 min.

NMR acquisition conditions

All NMR experiments were carried out on a Bruker DRX-600 spectrometer equipped with a cryo-probe, calibrated with acetone used as an internal standard (δ_H = 2.225 ppm; δ_C = 31.45 ppm) and recorded in D_2O at 310 K (CME6, E11, EPA1, EPA2, P9L1, P9L10 and E1L6), or at 316 K (P91) or at 323 K (E1L3) to shift the signal of the residual water peak in order to minimise its overlap with some of the glycopeptide signals. The complete set of two-dimensional spectra (DQ-COSY, TOCSY, T-ROESY, gHSQC and gHMBC) were recorded for all glycopeptides, except for PIL6, P9L1 and E11 for which the gHSQC experiment was enough to establish the structure of the glycan. The spectral width was set to 10 ppm and the frequency carrier placed at the residual HOD peak, suppressed by pre-saturation. Two-dimensional spectra were measured using standard Bruker software; for all the experiments, 512 FIDs of 2048 complex data points were collected, 32 scans per FID were acquired for homonuclear spectra and mixing times of 100 and 300 ms were used for TOCSY and T-ROESY spectra acquisition, respectively. Heteronuclear ^{1}H-^{13}C spectra were measured in the ^{1}H-detection mode, gHSQC spectrum was acquired with 30-60 scans per FID (depending on the abundance of the sample), the GARP sequence was used for ^{13}C decoupling during acquisition; gHMBC scans doubled those of gHSQC spectrum.

Bioinformatics

Geneious 11.0.5 software (https://www.geneious.com, 38) was used to deepen the genetic analysis of the variants to identify other genes involved in glycosylation or to define the nature of the mutation. The software was used to assemble and map antigenic mutant sequencing reads (total genome) to the PBCV-1 WT reference genome. The software allowed for quick and easy identification of genetic variants in the antigenic mutant viruses.

MD simulation of UDP-Rha and UDP-Glc with domain 1 of A064R.

The starting structure for MD simulation of the GT A064R D1 in complex with the substrate UDP-L-Rha was generated by replacing the Glc unit of the UDP-Glc with L-Rha in the complex with the enzyme (PDB code 2P72). The structures of the UDP-Rha and UDP-Glc were relaxed by energy minimization with a low gradient convergence threshold (0.05) in 1000 steps with MM3* force field integrated in the MacroModel program run under Maestro (Schrödinger Release 2016-4, LLC, New York, NY, 2016). The nucleotide moiety of the minimized UDP-L-Rha and UDP-Glc were superimposed on the UDP molecule in the structure of A064R in complex with UDP (PDB code 2P73). Partial charges of the ligands were calculated using the Gaussian 09 package with HF/6-31G(d) basis set. All molecular parameters were converted using the Antechamber program in the Amber16 package. The resulting structures were used as starting points for molecular dynamics simulation of 100 ns in explicit water solvent, using the Amber16 program with the protein.ff14SB force field parameters for protein and the generalized amber force field gaff2 for the sugar-nucleotide, whereas water.tip3p force field was used for water and ions. The atomic parameters for Mn^{2+} ion were defined according to published data (39). The starting structures were solvated in a 10 Å octahedral box of explicit TIP3P waters, and counterions were added to maintain electroneutrality. In order to fill all of the protein cavities with water molecules, a previous minimization for only solvent and ions was made. To reach a low energy starting structure, the entire system was minimized with a higher number of cycles, using the accurate steepest descent algorithm. The system was subjected to two rapid molecular dynamic simulations (heating and equilibration) before starting the real dynamic simulation: i) 20 ps of MD heating the whole system from 0 to 300 K, using NVT (fixed number of atoms, N, a fixed volume, V, and a fixed temperature, T) ensemble and a cutoff of 10 Å, followed by ii) equilibration of the entire system over 100 ps at 300 K using NPT (fixed number of atoms, N, a fixed pressure, P, and a fixed temperature, T) ensemble and a cutoff of 10 Å. A relaxation time of 2 ps was used in order to equilibrate the entire system in each step. The equilibrated structure was the starting point for the final MD simulations at constant temperature (300 K) and pressure (1 atm). Molecular dynamics simulation without constraints was recorded, using an NPT ensemble with periodic boundary conditions, a cutoff of 10 Å, and the particle mesh Ewald method. Coordinates and energy values were recorded every 10000 steps (20 ps) for a total simulation time of 100 ns and an ensemble of 5000 MD frames. A detailed analysis of the MD trajectory (for example, RMSD evaluation, kinetic and potential energies, etc.) was accomplished using the cpptraj module included in the Amber-Tools 16 package, and it is gathered in the Supporting Information (Fig. S9). The structure of the GT from antigenic variant EPA-1, GT A064R D1_S79L mutant, in complex with the substrate UDP-L-Rha was generated by mutagenesis tool integrated in the PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC) program. The generated structure was minimized using the Amber16 program as previously described in this section.

MM/GBSA calculations.

Binding free energy analysis was performed by applying the MM/GBSA calculation. The molecular mechanic energies combined with the generalized Born and surface area continuum solvation method was applied to the structure of UDP-L-Rha and UDP-Glc, A064R-D1 and to the UDP-L-Rha/A064R-D1 and UDP-Glc/A064R-D1 complexes generated from the MD simulations, as described in the previous section. The receptor-ligand complexes were energetically minimized by the MM/GBSA method as implemented in the Sander program of the Amber package. The atomic radii (set default PBRadii mbondi2) were chosen for all GBSA calculations. For the ensemble-average MM/GBSA rescoring, a MD trajectory was obtained for both UDP-L-Rha/- and UDP-Glc/A064R-D1 complexes. Snapshots of the complex were extracted from the MD trajectory at every 10 ps. Thus, the final binding free energy is an average of the binding free energies of 10000 snapshots.

Conflict of Interest: the authors declare that they have no conflict of interest with the content of this article.

Abbreviation list: COSY, correlation spectroscopy; GBSA, generalized Born surface area; gHMBC, gradient selected heteronuclear multiple bond correlation; gHSQC, gradient selected heteronuclear single quantum coherence; GT, glycosyltransferase; MCP, major capsid protein; MD, molecular dynamic; MM, molecular modeling; PBCV-1, Paramecium bursaria chlorella virus type 1; PMAA, partially methylated and acetylated alditols; SEC, size exclusion chromatography; TOCSY, total correlation spectroscopy; T-ROESY, transverse rotating frame Overhauser effect spectroscopy; WT, wild-type

References

1. Bagdonaite, I. I., and Wandall, H. H. (2018) Global aspect of viral glycosylation. *Glycobiology* **28**, 443-467.

2. Doms, R. W., Lamb, R. A., Rose, J. K., and Helenius, A. (1993) Folding and assembly of viral membrane proteins. *Virology* **193**, 545-562.

3. Olofsson, S., and Hansen, J. E. S. (1998) Host cell glycosylation of viral glycoproteins: a battlefield for host defense and viral resistance. *Scand. J. Infect. Dis*. **30**, 435-440.

4. Flint, S. J., Enquist, L. W., Racaniello, V. R., and Skalka, A. M. (2004) *Principles of Virology, Molecular Biology, Pathogenesis, and Control of Animal Viruses.* ASM Press, Washington.

5. Vigerust, D. J., and Shepherd, V. L. (2007) Virus glycosylation: role in virulence and immune interactions. *Trends Microbiol*. **15**, 211-218.

6. Hunter, E. (2007) *Virus assembly*. In Fields Virology, 5th ed., pp. 141-168. Edited by DM Knipe, PM Howley, DE Griffin, RA Lamb, SE Straus, MM Martin, B Roizman. Philadephia: Wolters Kluwer/Lippincott Williams & Wilkins, Publ.

7. Jeanniard, A., Dunigan, D. D., Gurnon, J. R., Agarkova, I. V., Kang, M., Vitek, J., Duncan, G., McClung, O. W., Larsen, M., Claverie, J. M., Van Etten, J. L., and Blanc, G. (2013) Towards defining the chloroviruses: a genomic journey through a genus of large DNAviruses. *BMC Genomics* **14**, 158.

8. Van Etten, J. L., Burbank, D. E., Xia, Y., and Meints, R. H. (1983b) Growth cycle of a virus, PBCV-1, that infects chlorella-like algae. *Virology* **126**, 117-125.

9. Dunigan, D. D., Cerny, R. L., Bauman, A. T., Roach, J. C., Lane, L. C., Agarkova, I. V., Wulser, K., Yanai-Balser, G. M., Gurnon, J. R., Vitek, J. C., Kronschnabel, B. J., Jeannard, A., Blanc, G., Upton, C., Duncan, G. A., McClung, O. W., Ma, F., and Van Etten J. L. (2012) Paramecium bursaria chlorella virus 1 proteome reveals novel architectural and regulatory features of a giant virus. *J. Virol*. **86**, 8821-8834.

10. De Castro, C., Klose, T., Speciale, I., Lanzetta, R., Molinaro, A., Van Etten, J. L., and Rossmann, M. G. (2018) Structure of the chlorovirus PBCV-1 major capsid glycoprotein determined by combining crystallographic and carbohydrate molecular modeling approaches. *Proc. Natl. Acad. Sci. USA* **115**, E44-E52.

11. De Castro, C., Molinaro, A., Piacente, F., Gurnon, J. R., Sturiale, L., Palmigiano, A., Lanzetta, R., Parrilli, M., Garozzo, D., Tonetti, M., and Van Etten, J. L. (2013) Structure of the N-linked oligosaccharides attached to virus PBCV-1 major capsid protein: an unusual class of complex N-glycans. *Proc. Natl. Acad. Sci. USA* **110**, 13956-13960.

12. Van Etten, J. L., Agarkova, I., Dunigan, D. D., Tonetti, M., De Castro, C., and Duncan, G. A. (2017) Chloroviruses have a sweet tooth. *Viruses* **9**, 88.

13. DeAngelis, P., Jing, W., Graves, M. V., Burbank, D. E., and Van Etten, J. L. (1997) Hyaluronan synthase of chlorella virus PBCV-1. *Science* **278**: 1800-1803.

14. De Castro, C., Speciale, I., Duncan, G., Dunigan, D. D., Agarkova, I., Lanzetta, R., Sturiale, L., Palmigiano, A., Garozzo, D., Molinaro, A., Tonetti, M., and Van Etten, J. L. (2016) N-linked glycans of chloroviruses sharing a core architecture without precedent. *Angew. Chem. Int. Ed*. **55**, 654-658.

15. Quispe, C. F, Esmael, A., Sonderman, O., McQuinn, M., Agarkova, I., Battah, M., Duncan, G. A., Dunigan, D. D., Smith, T. P. L., De Castro, C., Speciale, I., Ma, F., Van Etten, J. L. (2017) Characterization of a new chlorovirus type with permissive and non-permissive features on phylogenetically related strains. *Virology* **500**, 103-113.

16. Speciale, I., Agarkova, I., Duncan, G. A., Van Etten, J.L., and De Castro, C. (2017) Structure of the N-glycans from the chlorovirus NE-JV-1. *Antonie Leeuwenhoek*, **110**, 1391-1399.

17. Wang, I-N., Li, Y., Que, Q., Bhattacharya, M., Lane, L. C., Chaney, W. G., and Van Etten, J. L. (1993) Evidence for virus-encoded glycosylation specificity. *Proc. Natl. Acad. Sci. USA* **90**, 3840-3844.

18. Van Etten, J. L, Gurnon, J. R, Yanai-Balser, G. M, Duningan, D. D., and Graves, M. V. (2010) Chlorella viruses encode most, if not all, of the machinery to glycosylate their glycoproteins independent of the endoplasmic reticulum and Golgi. *Biochim. Biophys. Acta* **1800**, 152-159.

19. Bock, K., and Pedersen, C. (1983) Carbon-13 nuclear magnetic resonance spectroscopy of monosaccharides. *Adv. Carbohydr. Chem. Biochem*. **41**, 27-66.

20. Landstein, D., Burbank, D. E., Nietfeldt, J. W., and Van Etten, J. L. (1995) Large deletions in antigenic variants of the chrorella virus PBCV-1. *Virology* **214**, 413-420.

21. Graves, M.V., Bernadt, C.T., Cerny, R., and Van Etten, J.L. (2001) Molecular and genetic evidence for a virus-encoded glycosyltransferase involved in protein glycosylation. *Virology* **285**, 332-345.

22. Zhang, Y., Xiang, Y., Van Etten, J. L., and Rossmann, M. G. (2007) Structure and function of a chlorella virus encoded glycosyltransferase. *Structure* **15**, 1031–1039.

23. Tonetti, M., Zanardi, D., Gurnon, J. R., Fruscione, F., Armirotti, A., Damonte, G., Sturla, L., De Flora, A., and Van Etten, J. L. (2003) Paramecium bursaria Chlorella virus 1 encodes two enzymes involved in the biosynthesis of GDP-L-fucose and GDP-D-rhamnose. *J. Biol. Chem*. **278**:21559-65.

24. Oka, T., Nemoto, T., and Jigami, Y. (2007) Functional analysis of Arabidopsis thaliana RHM2/MUM4, a multidomain protein involved in UDP-D-glucose to UDP-L-rhamnose conversion. *J. Biol. Chem.* **282**:5389-403.

25. Jeevarajah, D., Patterson, J. H., Taig, E., Sargeant, T., McConville, M. J., and Billman-Jacobe, H. (2004) Methylation of GPLs in *Mycobacterium smegmatis* and *Mycobacterium avium*. *J. Bact*. **186**: 6792-6799.

26. Gohlke, H., and Klebe, G. (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angew Chem Int Ed*. **41**, 2644-2676.

27. Genheden, S., and Ryde, U. (2015) The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov*. **10**, 449–461.

28. Mulakala, C., and Viswanadhan V.N. (2013) Could MM-GBSA be accurate enough for calculation of absolute protein/ligand binding free energies? *J. Mol. Graph. Model*. **46**,41-51.

29. Markine-Goriaynoff, N., Gillet, L., Van Etten, J. L., Korres, H., Verma, N., and Vanderplasschen, A. (2004) Glycosyltransferases encoded by viruses. *J. Gen. Virol.* **85**, 2741-2754

30. Colson, P., De Lamballerie. X., Yutin, N., Asgari, S., Bigot, Y., Bideshi, D. K., Cheng, X.-W., Federici, B. A., Van Etten, J. L., Koonin, E. V., Scola, B., and Raoult, D. (2013) "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. *Arch virol.* **158**:2517-2521

31. Piacente, F., Bernardi, C., Marin, M., Blanc, G., Abergel, C., and Tonetti, M. G. (2014) Characterization of a UDP-N-acetylglucosamine biosynthetic pathway encoded by the giant DNA virus Mimivirus. *Glycobiology*, **24**: 51-61.

32. Chothi, M. P., Duncan, G. A., Armirotti, A., Abergel, C., Gurnon, J. R., Van Etten, J. L., Bernardi, C., Damonte, G., and Michela Tonetti (2010) Identification of an L-Rhamnose Synthetic Pathway in Two Nucleocytoplasmic Large DNA Viruses. *J. Virol*. **84**: 8829-8838.

33. Piacente, F., Marin, M., Molinaro, A., De Castro, C., Seltzer, V., Salis, A., Damonte, G., Bernardi, C., Claverie, J. M., Abergel, C., and Tonetti, M. (2012) Giant DNA virus mimivirus encodes pathway for biosynthesis of unusual sugar 4-amino-4,6-dideoxy-D-glucose (Viosamine). *J. Biol. Chem*. **287**:3009-3018.

34. Piacente, F., De Castro, C., Jeudy, S., Gaglianone, M., Laugieri, M. E., Notaro, A., Salis, A., Damonte, G., Abergel, C., and Tonetti, M. G. (2017) The rare sugar N-acetylated viosamine is a major component of Mimivirus fibers. *J. Biol Chem*. **292**: 7385-7394.

35. Rommel, A, Hülsmeier, A. J., Jurt, S., and Hennet, T. (2016) Giant mimivirus R707 encodes a glycogenin paralogue polymerizing glucose through α- and β-glycosidic linkages. *Biochem. J.* **473**, 3451-3462.

36. Van Etten, J. L., Burbank, D. E., Kuczmarski, D., and Meints, R. H. (1983) Virus infection of culturable chlorella-like algae and development of a plaque assay. *Science* **219**, 994-996.

37. De Castro, C., Parrilli, M., Holst, O., and Molinaro, A. (2010) Microbe-associated molecular patterns in innate immunity: extraction and chemical analysis of Gram-negative bacterial lipopolysaccharides. *Methods Enzymol*. **480**, 89-115.

38. Kearse, M, Moir, R, Wilson, A, Stones-Havas, S, Cheung, M, Sturrock, S, Buxton, S, Cooper, A, Markowitz S., Duran, C, Thierer, T, Ashton, B, Mentjies, P, and Drummond, A. (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647-1649.

39. Neves, R. P., Sousa, S.F., Fernandes, P. A., and Ramos, M. J. (2013) Parameters for molecular dynamics simulations of manganese-containing metalloproteins. *J. Chem. Theory. Comput.* **9**, 2718–2732.

FOOTNOTES

This research was supported by grants from the Mizutani Foundation under grant no. 180047 (CDC), NIH-INBRE (GD), and the National Science Foundation under grant no. 1736030 (JVE). The group CIC at bioGUNE also thanks MINECO of Spain for Grant CTQ2015-64597-C2-1P and for the Severo Ochoa Excellence Accreditation (SEV-2016-0644).

Table 1: Genetic features of the 20 PBCV-1 antigenic variants. These variants are divided into six classes depending on their reaction with polyclonal antibodies. Class E is not included in this study and omitted in this table, because the members of this group cross-react with the antibodies selective for both Classes A and B, suggesting an intermediate phenotype. All the antigenic classes, but C, have a mutation in *a064r*, this gene encodes for a protein of 638 amino acids that includes three domains of ca. 200 amino acids each (Fig. 1). The N-terminal domain 1 is a GT, domain 2 only matches bacterial proteins of unknown function, domain 3 has a weak match with methyltransferases. The structure of the N-glycans has been determined in this work (Fig. 1) for the antigenic variants typed in bold.

Antigenic variant	**Antigenic Class A: mutation in the second domain of *a064r*** (NP_048412.1)
EPA-11	Nonsense mutation (G→A) at nt position 932 in the gene; truncated protein (307 amino acids)
P9L1	Nonsense mutation (C→A) at nt position 704 in the gene; truncated protein (234 amino acids)
P9L4	Missense mutation (G→T) at nt position 1102 in the gene; amino acid substitution (G368W)
P9L7	Missense mutation (G→T) at nt position 1102 in the gene; amino acid substitution (G368W)
P9L10	Nonsense mutation (T→G) at nt position 669 in the gene; truncated protein (222 amino acids)
P91	Insertion of 27 nts after nt position 952 in the gene; insertion of 9 additional amino acids
	Antigenic Class B: mutation in the first domain of *a064r* (NP_048412.1)
EPA-1	Missense mutation (C→T) at nt position 236 in the gene; amino acid substitution (S79L)
EPA-2	Frameshift mutation; single nt insertion of A after position 262; truncated protein (88 + 11 amino acids: frameshift results in 11 incorrect amino acids added before stop codon is encountered)
P31	Missense mutation (G→A) at nt position 362in the gene; amino acid substitution (G121E)
	Antigenic Class C: no mutation in *a064r* (except E11); mutation in *a075l* (NP_48423.1: 280 amino acids)
E11	Same mutation in *a064r* as EPA-1 (antigenic class B). Frameshift mutation in *a075l*; single nucleotide addition at position 670; truncated protein (223 + 16 amino acids; frameshift results in 16 incorrect amino acids added before premature stop codon is encountered).
E1L1	Frameshift mutation; single nt deletion at position 666 in the gene; truncated protein (222 amino acids)
E1L2	Nonsense mutation (G→A) at nt position 630 in the gene; truncated protein (209 amino acids)
E1L3	Missense mutation (A→T) in the initiator codon at nt position 1 in the gene; no protein synthesis
E1L4	Nonsense mutation (T→A) at nt position 659 in the gene; amino acid substitution (I220N)
P41	Frameshift mutation single nt insertion of C after position 194 in the gene; truncated protein (65 + 14 amino acids: frameshift results in 14 incorrect amino acids added before stop codon is encountered)
	Antigenic Class D: *a064r* fully deleted
P1L2	Large deletion of 27,160 nts between genomic positions 12,320 and 39,480; the deletion includes the 44 nts at the 3' region of gene *a014r* through and including approximately half of gene *a075l.*
P1L3	Large deletion of 37,450 nts between genomic positions 4,970 and 42,420; the deletion includes gene *a009r* through gene *a078r* (the 5' breakpoint is between genes *a007/008l* and *a009r* while the 3' breakpoint is between genes *a078r* and *a079r*).
P1L6	Large deletion of 38,642 nts between genomic positions 2,168 and 40,810; the deletion includes the terminal one-sixth of gene *a003r* through the middle of gene *a077l.*
P1L10	Large deletion of 32,060 nts between genomic positions 7,700 and 39,760; the deletion includes approximately 60% of gene *a011l* through two-thirds of gene *a075l.*

	Antigenic Class F: mutation in the third domain of *a064r* (NP_048412.1)
CME6	Frameshift mutation; insertion of 2 nts at position 1320 in the gene; truncated protein (440 + 10 amino acids: frameshift results in 10 incorrect amino acids added before stop codon encountered)

Table 2. Calculated binding free energies (kcal/mol) for UDP-Rha (UDP-Rha) and UDP-Glc (UDP-Glc) to A064R-D1. The binding free energy (ΔG_{bind}) is decomposed into different energy terms: the gas-phase interaction energy (ΔE_{gas}) between the receptor and the ligand is the sum of electrostatic (ΔE_{ELE}) and van der Waals (ΔE_{VDW}) interaction energies. The solvation free energy (ΔG_{sol}) is divided into the polar (ΔG_{GB}) and non-polar (ΔG_{Surf}) energy terms.

Substrate	$\mathbf{\Delta E_{ELE}}$	$\mathbf{\Delta E_{VDW}}$	$\mathbf{\Delta E_{gas}}$	$\mathbf{\Delta G_{Surf}}$	$\mathbf{\Delta G_{GB}}$	$\mathbf{\Delta G_{Sol}}$	$\mathbf{\Delta G_{bind}}$
UDP-Rha	-200.21	-26.13	-226.35	-6.17	152.90	146.73	**-79.63**
UDP-Glc	-230.84	-24.86	-255.71	-7.11	188.85	181.74	**-73.97**

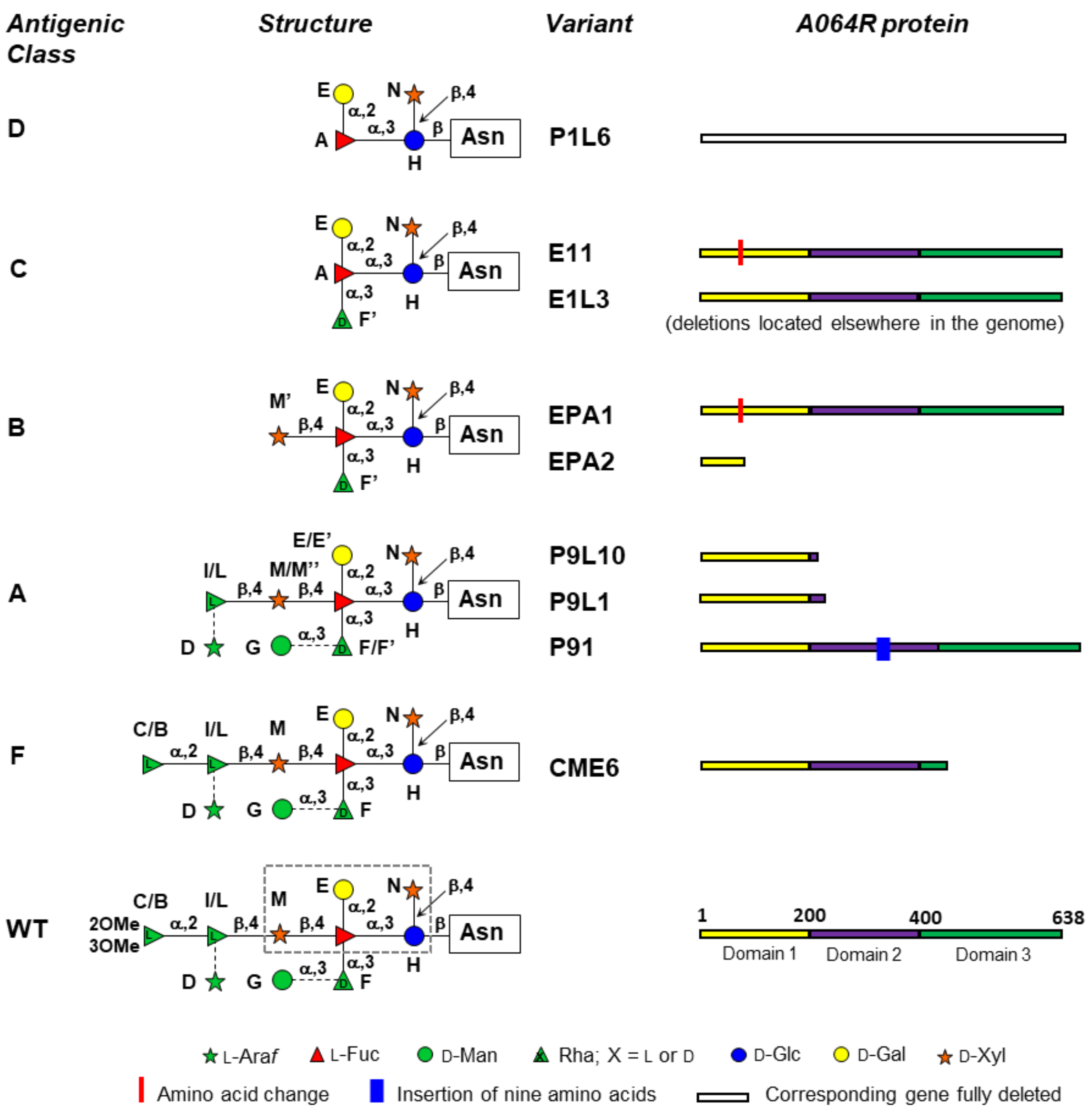

Fig. 1. Glycan structures of WT and selected antigenic mutants of PBCV-1. In the WT, each monosaccharide is denoted with the letter used during the NMR experiments, and residues inside the dotted box define the conserved core region. The glycosidic linkage of non-stoichiometric substituents is denoted with a broken line. Glyco1 and glyco2 are the most abundant glycoforms, both have Man. Glyco1 lacks Ara and the units of the terminal L-Rha disaccharide are labelled as **C** and **I**. Ara is present in glyco2 and the two L-Rha units are labelled **B** and **L**. Antigenic classes are labeled with a capital letter, (A-D and F) and the structure of their glycans depends on the kind of mutation found in the genome, which in all cases, except Classes C and D, occurs in the gene *a064r*. EPA1 (antigenic Class B) and E11 (antigenic Class C) have identical mutations in domain 1 of *a064r*, but E11 has an additional mutation in gene *a075l*. P91 has an additional glycoform with β-L-Rha methylated at O-2, the letters used to label this oligosaccharide are not reported due to crowding, and the structure is displayed in Fig. S7.

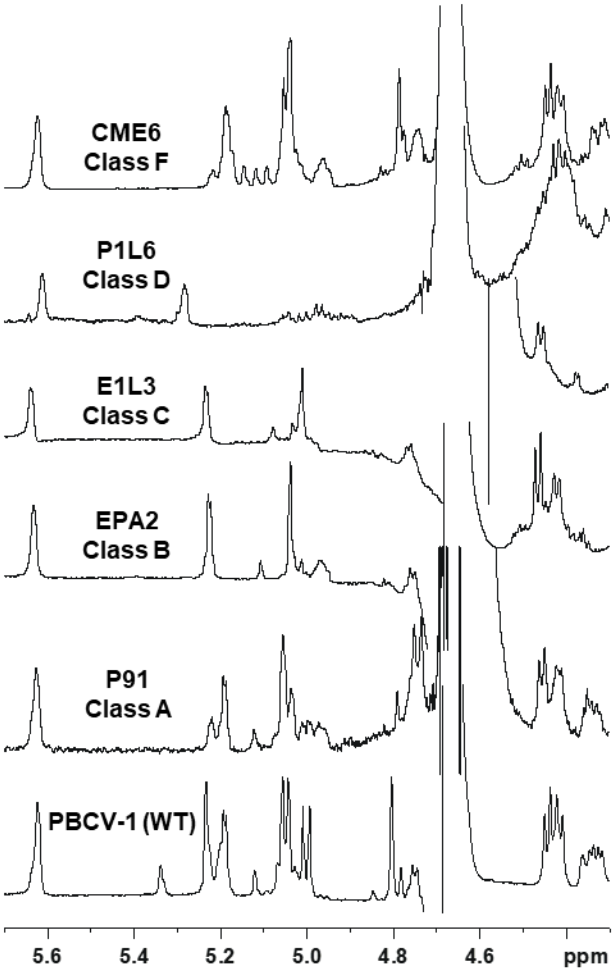

Fig. 2. Anomeric region of the proton spectra of different variants, each representative of its antigenic class (A-D and F). The proton spectrum of the WT virus PBCV-1 is given for comparison [with permission from De Castro et al. (11)]. NMR spectra were recorded at 310 K except for P91 (316 K) and E1L3 (323 K).

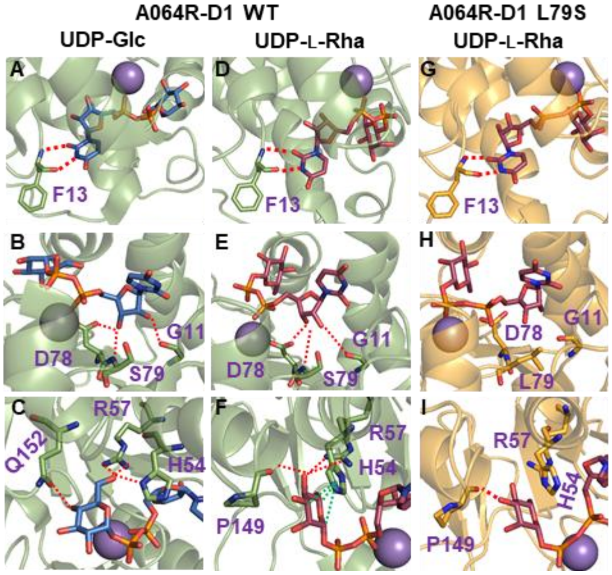

Fig. 3. Comparison of the active sites from the X-ray crystal structure of A064R-D1_WT:UDP-Glc (A-C), a representative frame from MD simulation of A064R-D1_WT:UDP-L-Rha (D-F) and a minimized structure of the mutant A064R-D1_S79L:UDP-L-Rha (G-I). (A,B,D,E) The binding mode of UDP is essentially the same in either UDP-Glc crystal structure or UDP-L-Rha MD simulated structure. (C) and (F): as expected, a different intermolecular interaction network is observed. For UDP-L-Rha, the His54 moves towards the Rha ring establishing CH-π interactions with the sugar (green dotted lines). (G-I) The mutation of the Ser79 by Leu dramatically alters substrate binding. Most of the intermolecular interactions are abrogated. Hydrogen bonds are highlighted with red dotted lines, π-stacking (in F) are highlighted with green dotted lines. UDP-Glc, UDP-L-Rha, and key residues of the enzyme are shown as blue, red-purple, and green sticks, respectively. Manganese is represented as purple sphere.

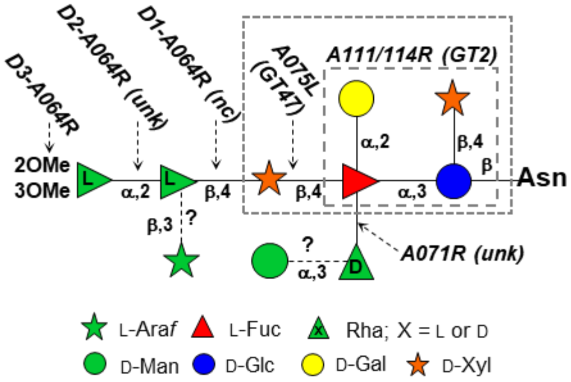

Fig. 4. PBCV-1 glycan structure with the predicted GTs involved in the glycosidic bond formation. The CAZY family fold is indicated in brackets next to the gene name (unk = unknown function, predicted as GT in this study; nc: unclassified GT fold). Based on the data available, *a111/114r* is involved in the formation of the tetrasaccharide enclosed in the inner grey box. Currently it is unknown if it catalyzes the formation of one or more of the four glycosidic linkages enclosed. The larger grey box indicates the monosaccharide units of the conserved N-glycan core region that is common to all other chloroviruses examined to date. D1-, D2- and D3- refer to the three domains of protein A064R. Linkages for which the corresponding GT is not identified yet are the non-stoichiometric substituents, Man and Ara, and have a "?" next to them.