Characterization of a UDP-\textit{N}\hyp{}acetylglucosamine biosynthetic pathway encoded by the giant DNA virus Mimivirus

Francesco Piacente\textsuperscript{2}, Cinzia Bernardi\textsuperscript{2}, Margherita Marin\textsuperscript{2}, Guillaume Blanc\textsuperscript{3}, Chantal Abergel\textsuperscript{3,1}, and Michela G Tonetti\textsuperscript{2,1}

\textsuperscript{2}Department of Experimental Medicine, Center of Excellence for Biomedical Research, University of Genova, Viale Benedetto XV, 1 Genova 16132, Italy and \textsuperscript{3}Structural and Genomic Information laboratory, Centre National de la Recherche Scientifique, Aix-Marseille Université UMR7256, IMM, Parc Scientifique de Luminy, Marseille, France

Received on September 2, 2013; revised on October 1, 2013; accepted on October 2, 2013

Mimivirus is a giant DNA virus belonging to the Megaviridae family and infecting unicellular Eukaryotes of the genus \textit{Acanthamoeba}. The viral particles are characterized by heavily glycosylated surface fibers. Several experiments suggest that Mimivirus and other related viruses encode an autonomous glycosylation system, forming viral glycoproteins independently of their host. In this study, we have characterized three Mimivirus proteins involved in the de novo uridine diphosphate-\textit{N}\hyp{}acetylglucosamine (UDP-\textit{N}\hyp{}acetylglucosamine) production: a glutamine-fructose-6-phosphate transaminase (CDS L619), a glucosamine-6-phosphate \textit{N}\hyp{}acetyltransferase (CDS L316) and a UDP-\textit{N}\hyp{}acetylglucosamine pyrophosphorylase (CDS R689). Sequence and enzymatic analyses have revealed some unique features of the viral pathway. While it follows the eukaryotic-like strategy, it also shares some properties of the prokaryotic pathway. Phylogenetic analyses revealed that the Megaviridae enzymes cluster in monophyletic groups, indicating that they share common ancestors, but did not support the hypothesis of recent acquisitions from one of the known hosts. Rather, viral clades branched at deep nodes in phylogenetic trees, forming independent clades outside sequenced cellular organisms. The intermediate properties between the eukaryotic and prokaryotic pathways, the phylogenetic analyses and the fact that these enzymes are shared between most of the known members of the Megaviridae family altogether suggest that the viral pathway has an ancient origin, resulting from lateral transfers of cellular genes early in the Megaviridae evolution, or from vertical inheritance from a more complex cellular ancestor (reductive evolution hypothesis). The identification of a virus-encoded UDP-\textit{N}\hyp{}acetylglucosamine pathway reinforces the concept that GlcNAc is a ubiquitous sugar representing a universal and fundamental process in all organisms.

Keywords: glutamine-fructose-6P transaminase / glucosamine-6P acetyltransferase / Mimivirus / UDP-\textit{N}\hyp{}acetylglucosamine / UDP-\textit{N}\hyp{}acetylglucosamine pyrophosphorylase

Introduction

Mimivirus is a member of the nucleocytoplasmic large DNA viruses (NCLDV) group and belongs to the rapidly expanding Megaviridae family (Raoult et al. 2004; Arslan et al. 2011; Yutin et al. 2013). Metagenomic surveys have indicated that members of the Megaviridae family are widespread in the marine ecosystems and probably represent a significant component of eukaryotic DNA viruses (Monier et al. 2008; Hingamp et al. 2013). Mimivirus is one of the largest viruses ever described, with a 1.2 Mbp genome encoding >1000 genes and a diameter of 700 nm (Raoult et al. 2004; Legendre et al. 2011). Its genome encodes genes involved in several functions that were considered trademarks of cellular organisms, including elements of the translation apparatus, aminoacyl-tRNA synthetases and a number of enzymes involved in protein modification or production of metabolic intermediates (Raoult et al. 2004; Abergel et al. 2007; Legendre et al. 2011). Mimivirus, like other large DNA viruses such as Poxviruses, replicates in large cytoplasmic virion factories, around which the virions are assembled. Infectious virions are covered by a dense layer of long fibers which are heavily glycosylated. Glycans are probably essential for the interaction of the viral particles with the host, in particular to promote phagocytosis, but they can also provide protection against the harsh environments facing Mimivirus, including the host vacuole.

In the last years, it has become increasingly evident that, in contrast to other viruses that use the host ER/Golgi system for their glycoprotein production, some members of the NCLDV, such as Chlorella viruses, encode at least in part, if not all, the machinery required for the glycosylation of their structural proteins (Van Etten et al. 2010). Indeed, recent evidences have indicated that the glycans associated with \textit{Paramecium bursaria} chlorella virus 1 (PBCV-1), a representative chlorovirus, have very unusual and complex structures, never described before (De Castro et al. 2013). The presence of an independent glycosylation machinery has been considered again a distinctive feature of cellular organisms; however, it is now evident that viruses also encode a complex glycosylation system, which comprises enzymes involved in the formation of nucleotide-sugar, glycosyltransferases and glycosidases. However, the...
mechanisms leading to glycan formation in large DNA viruses still need to be elucidated.

In previous studies, we have demonstrated that Mimivirus encodes functional enzymes responsible for the complete UDP-α-threonose and UDP-α-viosamine biosynthetic pathways (Parakkottil Chothi et al. 2010; Piacente et al. 2012). Indeed, the analysis of the monosaccharide composition of the virions has revealed that the major sugar components of the Mimivirus fibers-associated glycans are N-acetylgalactosamine (GlcNAc), glucose, rhamnose and the 6-deoxyaminosugar viosamine (Piacente et al. 2012). These findings indicate that Mimivirus not only encodes specific glycosyltransferases (Luther et al. 2011), but it is also able to produce the precursors for glycan biosynthesis independently of the host, reinforcing the notion of a viral autonomous production of glycoconjugates. In this work, we extended our previous study to the identification and characterization of the Mimivirus UDP-GlcNAc biosynthetic pathway.

GlcNAc is a monosaccharide widely used in most cellular organisms, e.g., for polymer formation, in cell walls and in glycoproteins and glycolipids; the essential role of GlcNAc in Bacteria and Eukaryota is well known. Recent evidence has also unveiled the importance of this monosaccharide in archael glycoconjugates (Namboori and Graham 2008). Moreover, reversible O-GlcNAc modifications of soluble eukaryotic proteins represent a fundamental system of post-translational regulation and contribute to the fine regulation of cellular functions (Hart et al. 2011). Finally, GlcNAc is the precursor of modified acetamido sugars, such as N-acetylgalactosamine (Guo et al. 2006), sialic acid (Reinke et al. 2009), N-acetylmuramic acid (Zhu et al. 2012), N-acetylmannosaminuronate (Namboori and Graham 2008) and several 6-deoxy-2-acetamido sugars, including bacillosamine and pseudaminic acid, which are core components of bacterial N-linked and O-linked glycans, respectively (Schoenhofen et al. 2006). Thus, GlcNAc represents one of the most important and abundant sugars in the biosphere.

The universal distribution of GlcNAc and its derivatives suggests that the pathway leading to its formation emerged very early. Both eukaryotes and prokaryotes produce this 2-acetamido sugar as an UDP-activated compound, but using different biosynthetic pathways (Figure 1). They both start with a common step, the conversion of fructose-6-P (Fru-6P), a glycolysis intermediate, into glucosamine-6-P (GlcN-6P), through a complex reaction catalyzed by a well-conserved enzyme, glucosamine-6P synthase (GFAT) (GlmS in Bacteria; known as glutamine-fructose-6-phosphate transaminase (isomerizing), GFAT, in Eukarya; EC 2.36.1.16), which uses glutamine as ammonia donor (Badet et al. 1987; Milewski 2002; Raczyńska et al. 2007). In Bacteria and in some Euryarchaea, the pathway proceeds with the isomerization of GlcN-6P to GlcN-1P, catalyzed by glucosamine-6P mutase (GlmM; EC 5.4.2.10) (Mengin-Lecreulx and van Heijenoort 1996; Namboori and Graham 2008); GlcN-1P is then converted into UDP-GlcNAc by a bifunctional enzyme capable of acetylatyng the sugar phosphate and catalyzing UDP-GlcNAc formation (GlmU; EC 2.3.1.157 and EC 2.7.7.23) (Brown et al. 1999; Olsen and Roderick 2001). Alternatively, in eukaryotic cells, GlcN-6P is first acetylated by glucosamine-6-phosphate N-acetyltransferase (GNAT; EC 2.3.1.4), and then GlcNAc-6P is converted into GlcNAc-1P by a phosphoaectylglucosamine mutase (PGM3/AGM1; EC 5.4.2.3) (Mio et al. 1999; Milewski et al. 2006); finally, the UDP-GlcNAc pyrophosphorylase (UAP; EC 2.7.7.23) transfers a uridine monophosphate (UMP) to the phosphorylated sugar using UTP as donor (Maruyama et al. 2007). Two possible salvage routes are also used by different organisms. Specifically, free glucosamine or GlcNAc, derived from exogenous sources or from recycling of glycoconjugates, are phosphorylated and, if necessary, acetylated (Yamada-Okabe et al. 2001). Alternatively, UDP-GlcNAc can be formed by epimerization of UDP-N-acetylgalactosamine (Guo et al. 2006).

The studies reported here have led to the identification of three Mimivirus genes, encoding enzymes of the de novo UDP-GlcNAc biosynthetic pathway. The order of the catalyzed reactions indicates that this pathway follows the eukaryotic-like strategy. However, some properties of the viral enzymes make them more similar to the bacterial ones, highlighting a viral pathway mixing eukaryotic and prokaryotic features. The fact that these genes are shared among most known members of the Megaviridae family suggests that GlcNAc is widely used for the glycans synthesis in these giant viruses. Finally, phylogenetic studies indicated that the origin of the enzymes is complex, but all proteins appear to be anciently encoded by Megaviridae, branching at deep nodes in the phylogenetic trees. These findings bring support to the hypothesis that the viral pathway is ancestral and argue against a recent acquisition by Megaviridae through horizontal gene transfer.

Figure 1. De novo UDP-GlcNAc biosynthetic pathways in Eukaryota and Bacteria. Glc-6P, glucose-6-phosphate; Fru-6P, fructose-6-phosphate; GlcN-6P, glucosamine-6-P; GlcN-1P, glucosamine-1-phosphate; GlcNAc-6P, N-acetylgalactosamine-6-phosphate; GlcNAc-1P, N-acetylgalactosamine-1-phosphate; UDP-GlcNAc, UDP-N-acetylgalactosamine. (1) glucosamine-6-P synthase (GFAT in Eukaryota; GlmS in Bacteria); (2) glucosamine-6-P N-acetyltransferase, GNAT; (3) PGM3/AGM1; (4) UDP-N-acetylgalactosamine pyrophosphorylase, UAP; (5) phosphoglucosamine mutase, GlmM; (6) bifunctional glucosaminie-1-P acetytransferase/uridylyltransferase, GlmU. Mimivirus gene products involved are indicated in correspondence of the catalyzed enzymatic step.
Characterization of a virus-encoded hexosamine pathway

Results
Identification and expression of Mimivirus genes involved in the UDP-GlcNAc biosynthetic pathway

Sequences of well-characterized enzymes involved in the de novo UDP-GlcNAc pathway (Figure 1) in Eukarya, Bacteria and Archaea were used in search of orthologs in the Mimivirus genome. Three candidates were identified: L619, a putative GlcN-6P synthase (GFAT, YP_003987136); L316, a GlcN-6P acetyltransferase (GNAT, YP_003986819) and R689, a UAP (YP_0039867216). For clarity, abbreviations of the eukaryotic enzymes were used to refer to the viral enzymatic activity. In contrast to UDP-voisamine biosynthetic genes (Piacente et al. 2012), the three genes are not clustered in Mimivirus genome, as evidenced by their gene numbers. No homologs of the plant enzymes. Interestingly, although previous reports have indicated that this insertion is a hallmark of eukaryotic GFAT, it is also absent in sequences from a group of unicellular eukaryotic organisms, which include Amoeboa, Alveolata and Stramenopiles/Heterokonta. However, even if these protista do not have the insertion between the two domains, as in bacterial and viral GFATs, they significantly differ from the viral enzymes by the presence of an N-terminal extension of variable length (Supplementary data, Figure S1).

Mimivirus GFAT presents all the residues involved in substrate binding and catalysis previously identified in bacterial and eukaryotic enzymes. In particular, it displays the N-terminal cysteine (Cys) residue obtained after methionine cleavage, a key residue for glutaminase activity (Milewski 2002). Indeed, GFAT belongs to the Gn-AT class II or Ntn subfamily of amidotransferases; in members of this subfamily, activation of the catalytic Cys1 thiol is mediated by the presence of the N-terminal amino group of the amino acid. For this reason, the presence of extra amino acids at the N-terminus has been shown to dramatically impair the activity of the enzyme (Richez et al. 2007; Mouilleron et al. 2008; Czarnecka et al. 2012). Other critical residues are highlighted in Supplementary data, Figure S1.

To test the enzymatic activity, we expressed the viral GFAT as a fusion protein, to allow a reliable purification procedure and prevent contaminations by the Escherichia coli enzyme, using either an N-terminal GST-tag or a C-terminal 6 × His-tag, in order to compare their catalytic activities and assess the effect of the tags at the two extremities of the protein. In our expression system, the proteolytic cleavage of the N-terminal GST resulted in four extra amino acids, thus masking the α-amino group of the catalytic Cys. Conversely, the presence of a C-terminal tag can disturb the so-called C-loop, which is fundamental for closure of the isomerase active site upon Fru-6P binding and for the activation of the glutaminase activity (Mouilleron et al. 2008).

GFAT catalyzes the sequential isomerization of Fru-6P to GlcN-6P, followed by the transfer of an amino group from glutamine to form GlcN-6P, in an ordered bi–bi kinetic (Milewski 2002). However, in the absence of glutamine, the C-terminal isomerase domain behaves like a phosphoglucose isomerase and this activity can be assayed independently. The isomerase activity was determined by following the Glc-6P formation from this activity can be assayed independently. The isomerase activity was determined by following the Glc-6P formation from Fru-6P with a continuous fluorimetric assay, by monitoring the glucose-6P dehydrogenase-catalyzed formation of NADPH. As a result, the C-terminal His-tag clearly affected the isomerase activity (Figure 2A). Specifically, the isomerization rate of the protein with the C-terminal tag was ~30% of the one observed for the enzyme produced with an N-terminal GST (Figure 2A).

In a similar way, we independently tested the glutaminase activity, by following glutamate formation with a coupled assay with glutamate dehydrogenase (GDH) (Figure 2B). In agreement with previous observations, these results indicated that the presence of extra amino acids at the N-terminus severely perturbed the enzyme, with an activity <10% of the enzyme with the C-terminal His-tag (Figure 2B). Even if both types of recombinant proteins can catalyze the two independent reactions with different efficiencies, the formation of GlcN-6P determined by the modified Morgan–Elson assay could be observed only with the L619 produced with the C-terminal His-tag. This finding indicates that the presence of extra amino acids at the N-terminal end not only severely decreases the glutaminase activity, but...
probably also hinders ammonia channeling to the second substrate. Qualitative glucosamine formation was confirmed by gas-chromatography - mass spectrometry (GC-MS) analysis (not shown). Specific activity for the total reaction catalyzed by the L619 recombinant protein with a C-terminal His-tag and determined using the Morgan–Elson assay was found to be 3.8 ± 0.7 \( \mu \text{mol min}^{-1} \text{mg}^{-1} \) of the L619 recombinant protein. This value is \(~10\) times higher than the one previously reported for the \( E. \ coli \) GlmS containing a C-terminal His-tag (Richez et al. 2007). Finally, L619 protein is not feedback inhibited by the final product of the pathway, the UDP-GlcNAc, using concentrations up to 1 mM. This mechanism of regulation of the catalytic activity is a hallmark of all eukaryotic GFAs studied so far, while it is absent in prokaryotes. Thus, this property interestingly makes L619’s reagent, to detect the appearance of coenzyme A (CoA) free thiol group; using this assay, the specific activity was 3.4 ± 1.5 \( \mu \text{mol min}^{-1} \text{mg}^{-1} \). Thus, both assays indicated that the L316 gene product is active and catalyzes the transfer of an acetyl group from Acetyl-CoA to GlcN-6P. The activity using GlcN-1P as substrate, determined with the Ellman’s reagent, was about five times lower than the one observed for GlcN-6P (0.7 ± 0.3 \( \mu \text{mol min}^{-1} \text{mg}^{-1} \)), confirming that the enzyme behaves as its eukaryotic counterparts (Peneff et al. 2001).

**Figure 2.** Isomerase and glutaminase activity of the L619 gene product. Two concentrations of the recombinant proteins, produced with either an N-terminal GST or a C-terminal 6 × His-tag, were used for each assay. GST-fusion protein was proteolytically removed after purification; however, four extra residues remained at the N-terminus. (A) Isomerase activity. (B) Glutaminase activity. Results are obtained from one representative experiment and are expressed as fluorescence units.

**L316—glucosamine 6P acetyltransferase.** The L316 gene encodes a 148 amino acid long protein. Sequence analysis indicated that it has a GCN5-related acetyl transferase fold, which is characterized by a CoA-binding pocket (Peneff et al. 2001). BLAST search revealed that this enzyme is shared by some members of the Megaviridae family. The L316 protein shares 57% identical residues with the Moumouvirus putative acetyltransferase (YP_007354484) and 56% with the Megavirus mg590 protein (YP_004894641). Its closest homologs in cellular organisms correspond to putative GNAT-family proteins from *Trichomonas vaginalis* G3 (41% identity, XP_001303745) and from the *Candidatus Nitrosopumilus* Archea (50% identity, ZP_1039775). The multiple alignment of the viral sequences with Thaumarcheota and Eukaryota enzymes indicated that all the residues known to be involved in substrate recognition and catalysis are conserved (Supplementary data, Figure S2).

The GNAT enzymatic activity was measured using discontinuous assays by following the formation of the acetylated amino sugar using the Morgan–Elson assay. However, in this case, the acetylation step with acetic anhydride, which is used to detect acetylated 2-amino sugars, was omitted in order to specifically visualize the product of the acetyltransferase activity. The enzyme specific activity of the L316 recombinant protein using this assay was found to be 5.1 ± 0.9 \( \mu \text{mol min}^{-1} \text{mg}^{-1} \). This enzymatic activity was also assayed using the UDP-GlcNAc pyrophosphorylase. The R689 gene encodes a 255 amino acid long protein belonging to the GTA-type glycosyltransferase superfamily and corresponds to the N-terminal domain of the bacterial GlmU (Brown et al. 1999; Olsen and Roderick 2001). Using the BLAST search, the closest homolog corresponds to the Megavirus mg240 protein (66% identity YP_004894291). The Moumouvirus genome appears to encode a homolog of the R689 protein with which it share 68% identical residues but only over 172 amino acids, thus making the *Acanthamoeba polyphaga* Moumouvirus sequence (YP_007354151) surprisingly shorter than the Mimivirus and Megavirus ones. The best hits in cellular organisms correspond to the N-terminal domains of a Cyanobacteria and a Proteo-bacteria putative bifunctional GlmU (36 and 38% identical residues over 242 amino acids). The sequence identity with the well-characterized N-terminal domain of *E. coli* GlmU is 32% over 244 amino acids. The viral protein displays a slightly modified pyrophosphorylase sequence, from Leu15 to Lys29, corresponding to Leu11 to Lys25 in the *E. coli* enzyme. The only difference in the accepted consensus sequence (L(X2)GXGTXM (X4)PK is the substitution of the Thr17 of the *E. coli* protein by a Lys21 conserved in the Megaviridae enzymes. The multiple alignment of the R689 sequence with Megavirus and *E. coli* enzymes is reported in Supplementary data, Figure S3.

UAP activity of R689 protein was assayed both in the forward (transferase, Figure 3A) and in the reverse (pyrophosphorylase, Figure 3B) reactions using high-pressure (or high performance) liquid chromatography (HPLC) analysis. Transfer of the nucleotide triphosphate donors and hexose-1P acceptors, while the pyrophosphorylase was tested using several nucleotide sugars. High levels of activity were found
Glc-1P (Figure 3A and B), while GlcN-1P was not recognized. Activity in both directions was also detected using UDP-Glc or UDP-GlcNAc. Dotted line corresponds to an identical protein concentration incubated in the presence of Glc-1P; the corresponding formation of UDP-Glc is observed. (B) The reverse pyrophosphorylase reaction was tested using UDP-GlcNAc or UDP-Glc and inorganic pyrophosphate. Dotted and continuous chromatograms were obtained using UDP-Glc or UDP-GlcNAc as substrates, respectively.

only when uridine triphosphate (UTP) and GlcNAc-1P or UDP-GlcNAc were used as substrates (Figure 3A and B). Specific activity of the recombinant R689 protein for the uridylyltransferase reaction was 2.4 ± 0.4 µmol min⁻¹ mg⁻¹. A low activity in both directions was also detected using UDP-Glc or Glc-1P (Figure 3A and B), while GlcN-1P was not recognized (not shown). Regarding the nucleotide triphosphate donors, deoxythymidine triphosphate (dTTP) and guanosine triphosphate (GTP) could also be used with very low efficiency, while adenosine triphosphate (ATP) and cytidine triphosphate (CTP) were ineffective. As commonly observed for nucleotidyltransferases, the viral UAP to be active required divalent cations, with a preference for Mg²⁺ over Ca²⁺ and Mn²⁺.

Phylogenetic study of the viral UDP-GlcNAc pathway

Phylogenetic analysis of the three Mimivirus enzyme families revealed three different evolutionary histories. Mimivirus GFAT homologs were found in all the three domains of cellular organisms as well as in Megaviridae and Phycodnaviruses. The three cellular domains (Bacteria, Archaea and Eukarya) were broadly recovered into separate groups in the phylogenetic tree (Figure 4A). The grouping of bacteria and their separation from Archaea were not fully resolved and characterized by moderate statistical support, most likely because of weakening in the phylogenetic signal for these very ancient evolutionary events. The distribution of viral GFAT enzymes is remarkable. First, proteins from Chlorera virus (Phycodnavirus) branched within the bacteria clade, suggesting that the corresponding genes were acquired by a chlorovirus ancestor from bacteria through lateral gene transfer (Landstein et al. 1998). In contrast, the Megaviridae grouped in a monophyletic clade with an apparent phylogenetic affinity for eukaryotes. At first approximation, the emergence of Megaviridae within eukaryotes could be interpreted as a capture of an eukaryotic GFAT gene by a Megaviridae ancestor. However, the relative phylogenetic affinities between the different clades of eukaryotic proteins do not follow the accepted phylogeny of eukaryotes (Cavalier-Smith 2004). In the GFAT phylogenetic tree, eukaryotes formed two broad, highly supported clades: One clade contained organisms that have a central insertion responsible for the feedback inhibition mechanism (e.g., plants, Fungi, Metazoa, Trichomonas and Trypanosoma), while the other clade contained small eukaryotes that lack the insertion domain. This topology may reflect a case of hidden paralogy resulting from the duplication of an ancestral GFAT gene before the radiation of eukaryotes and subsequent differential losses of the duplicated copies in eukaryotic lineages. Under this scenario, the ancestral Megaviridae gene could have emerged from the shorter duplicate at some time possibly close to the beginning of eukaryogenesis.

The most similar homologs to the Mimivirus GNAT were found exclusively in Megaviridae, Cryptococcus neoformans virus (CroV), Archaea and eukaryotes. The resulting phylogenetic tree shown in Figure 4B is characterized by low bootstrap supports for deep internal branches and, therefore, any conclusion drawn shall be taken with some degree of caution. Archaea and eukaryotes were separated into distinct groups. A Megaviridae clade emerged at an intermediate position between Archaea and eukaryotes. CroV also emerged as a separated branch in between Archaea and eukaryotes, but without grouping directly with the other Megaviridae. The statistical support for the CroV branch that separated it from the Megaviridae clade is particularly low (78% approximate likelihood ratio test (aLRT) and 9% bootstrap); we cannot rule out that the CroV protein actually shares a common ancestor with Megaviridae and the observed placement of CroV in the phylogenetic tree is artifactual. Thus, based on the current state of knowledge and sequence data, the phylogenetic criterion cannot rule out the hypothesis of an ancient origin of the viral GNAT, but does not support the hypothesis of a close relationship with cellular homologs. However, the intermediate position of the viral branches in between Archaea and eukaryotes needs to be confirmed when more viral and eukaryotic sequences will be available for analysis.

In contrast, homologs to Mimivirus UAP were only found in Bacteria, Archaea and other Megaviridae. Two homologs were also sequenced from insects Bombus impatiens and Ceratitis capitata, but their close phylogenetic proximity to gamma δ-proteobacteria proteins (Figure 4C) suggests that these sequences are in fact bacterial contamination. Phylogenetic reconstruction readily grouped Archaea and bacteria enzymes.
Figure 4. Unrooted phylogenetic trees of the viral UDP-GlcNAc pathway. (A) ML tree constructed from the multiple alignment of GFATs containing 460 gap-free sites. (B) ML tree constructed from the multiple alignment of GNATs containing 123 gap-free sites. (C) ML tree constructed from the multiple alignment of UAPs containing 192 gap-free sites. Note that enzymes were also found in two Moumouvirus genomes (YP_007354151 and AEX63120) with close phylogenetic proximity to Megavirus chilensis. However, the two enzymes were not included in the present phylogenetic tree because they were truncated, which sensibly reduced the number of amino acid positions available for analysis. aLRT (%) and bootstrap support (out of 100 bootstrap replicates) for important interior branches are indicated. Scale bars represent the number of amino acid substitutions per site. All trees were midpoint rooted for the purposes of clarity. Original phylogenetic trees showing all sequences under study are provided as Supplementary data, Figures S4–S6.
into separate clades. Furthermore, the Megaviridae sequences formed a monophyletic group that emerged in between the archaeal and bacterial clades. Here, again statistical supports for deep internal branches were in general low, including the branch placing virus between Archaea and Bacteria (i.e., 31% aLRT and 6% bootstrap support). Therefore, the apparent deep branching of viruses as a separate clade from cellular organisms awaits confirmation when new sequences are available in databases. Nevertheless, the relative phylogenetic isolation of the virus proteins among cellular homologs does not support the hypothesis of a recent acquisition of a bacterial or archaeal gene by a Megaviridae ancestor.

**Discussion**

In this study, we have shown that Mimivirus encodes a multistep pathway involved in the de novo UDP-GlcNAc biosynthetic process. These data represent the first identification of a functional and almost complete hexosamine pathway encoded by a virus. GlcNAc is one major component of glycans associated with the fibers which cover Mimivirus capsids (Piacente et al. 2012) and it is the most abundant sugar found in *Megavirus chilensis* (Piacente et al. 2012). UDP-GlcNAc is also used as a precursor for the production of modified 6-deoxy-2-acetaminsugars by enzymes encoded by the Megavirus genome (F.P., unpublished results). Unlike the previously identified enzymes involved in the UDP-hexosamine and UDP-N-acetylhexosamine production (Parakkottil Chothi et al. 2010; Piacente et al. 2012), which are found in Mimivirus only, homologous enzymes for UDP-GlcNAc production are present in most known members of the Megaviridae family (Fischer et al. 2010; Arslan et al. 2011; Yoosuf et al. 2012; Santini et al. 2013; Yutin et al. 2013). This finding suggests that GlcNAc can represent a core component of giant virus glycans, in agreement with the widespread presence and the fundamental role played by this amino sugar in all kingdoms. The identification of enzymes involved in the UDP-GlcNAc production in viral genomes is quite surprising, since, unlike hrmhose and viosamine, which may be absent or produced in limited supply by the eukaryotic host cells, GlcNAc is produced by most organisms. However, due to its key role in several functions, UDP-GlcNAc production is tightly controlled, both in bacterial and eukaryotic cells, by feedback and transcriptional mechanisms and, specifically, the limiting step of the pathway is represented by the GlcN-6P production (Ferré-D’Amaré 2010). The use of its own pathway, in particular the presence of the first limiting enzyme GFAT, may allow the virus to escape all the cellular control systems. It may also represent a strategy to decrease the virus host-dependency, thus broadening the type of cells that can be infected. In this way, Mimivirus can be completely independent of its host for the supply of the nucleotide-sugar pools, which are then used by the viral glycosyltransferases for protein glycosylation. This finding suggests that surface glycans are highly important at some point of the virus replicative cycle. Since the Megaviridae infection occurs through phagocytosis, it is likely that the viruses use glycans as a “Trojan horse” to mimic bacterial preys. Moreover, the heavily glycosylated surface can provide protection against the external environment, as it is seen for bacterial capsular material and for some types of spores, and also it may play a fundamental role in the phagocytic vacuole.

Enzymatic characterization of Mimivirus recombinant proteins expressed in *E. coli* has revealed that Mimivirus pathway follows the eukaryotic-like strategy, in which the acetylation of the amino sugar precedes the intramolecular phosphate transfer (Figure 1). However, sequence and phylogenetic analyses revealed some peculiar features of the viral enzymes compared with their typical eukaryotic counterparts. In the case of the L619 gene product, GFAT, the protein is shorter and, like the bacterial and archaeal proteins, lacks an insertion of 40–70 amino acids, which connects the N-terminal with the C-terminal domains in fungi, animal and plant enzymes (Milewski 2002). This insertion was linked to the susceptibility of the eukaryotic GFAT to the inhibitory feedback effects of UDP-GlcNAc. Indeed, our results indicated that, similar to the prokaryotic enzymes, Mimivirus GFAT is not subjected to inhibition by the final product of the pathway. The absence of a feedback inhibition by UDP-GlcNAc may be relevant to viral life cycle, to avoid limitation in the supply of the nucleotide-sugar needed for the production of viral particles. The insertion between the two domains, considered a trademark of eukaryotic sequences, is also absent in putative GFATs from some Protista, specifically Amoebzoa (including the Mimivirus host *A. castellani*), Apicomplexa and Stramenopiles/Heterokonta, which, in this regard, are similar to the viral enzymes. However, GFATs in all these organisms contain an N-terminal extension of variable length, which is not present in enzymes from any other taxa. Preliminary data obtained in our laboratory suggest that GFATs displaying the N-terminal extension (including also the enzyme from *A. castellani*) follow different catalytic strategies compared with the viral enzyme, for which the presence of a free N-terminal Cys is essential for the catalytic activity (C.B., unpublished results). These findings strongly indicate that a recent acquisition of Mimivirus GFAT from its amoebal host is unlikely.

The L316 gene product is clearly an acetyltransferase (GNAT), which, similar to the eukaryotic enzymes, uses GlcN-6P over GlcN-1P as preferential substrate. As expected, best homologs for this enzyme were found among eukaryotes. However, significant homology was observed also with sequences from Archaea, in particular from some Thaumarchaeota. Interestingly, in several Euryarchaeae, the UDP-GlcNAc pathway is similar to the bacterial one, in which the isomerization step precedes the acetyl transfer (Namboori and Graham 2008). The presence of putative acetyltransferases which are homologous to viral and eukaryotic GNATs indicates the need for a more detailed study of the archaean pathways.

For the following isomerization step from GlcN-6P to GlcN-1P, using members of the α-D-phosphohexomutase superfamily which catalyze the intramolecular transfer of phospharyl group in hexoses and to which the eukaryotic PGM1/AGM1 belongs, we failed to identify any clear homologs in the Megaviridae genomes presently available. However, we cannot rule out that one of the many conserved uncharacterized proteins encoded by the Megaviridae genome could assume this function or, as alternative, a host-encoded enzyme could be used. Indeed, typical mutases capable of catalyzing this step are missing also in other organisms, such as *Sulfolobus solfataricus*, which however encodes all the other enzymes of the pathway, suggesting the presence of other still unidentified proteins involved in this reaction (Ray et al. 2005).
The final enzyme of the pathway is a nucleotidytransferase, which transfers UMP from the triphosphate to GlcNAc-1P. This step is present in both prokaryotes and eukaryotes; however, the enzymes that catalyze this reaction are different. In prokaryotes, the GlmU protein is a bifunctional enzyme characterized by the presence of both an acetyltransferase activity, mapped to the C-terminal domain, and an uridylyltransferase activity in the N-terminal domain; acetylation occurs on GlcN-1P and precedes the nucleotide transfer (Brown et al. 1999; Olsen and Roderick 2001). Eukaryotes have a specific enzyme which only possesses the uridylyltransferase activity on the already acetylated amino sugar (Maruyama et al. 2007). In Mimivirus, this activity is performed by the R689, whose sequence is, however, homologous to the bacterial GlmU N-terminal domain and not to the eukaryotic enzymes.

In summary, (i) the L619 sequence is closer to the eukaryotic enzymes, but it also shares some specific features with the prokaryotic ones (shorter sequence with no insertion between the N- and C-terminal domains, absence of a N-terminal prosequence) and, as for prokaryotes, there is no feedback regulation by the final product of the pathway; (ii) the acetylation is done by the L316 enzyme, which is eukaryotic-like in terms of functionality and sequence similarity; (iii) the final step is catalyzed by R689, which sequence is clearly related to the prokaryotic enzymes. These features make the viral pathway intermediate between the eukaryotic and the prokaryotic ones and pose questions about the origin of the viral genes.

Phylogenetic study of the three enzymes reveals three important features of the evolution of the viral UDP-GlcNAc pathway. First, for each enzyme family, the Mimivirus sequences form a monophyletic clade with other members of the Megaviridae family, indicating that they were inherited from a Megaviridae common ancestor. Thus, the Megaviridae ancestor most likely possessed a functional UDP-GlcNAc pathway. Secondly, sequence database mining and subsequent phylogenetic reconstruction do not support the hypothesis that the corresponding viral genes have recently been acquired from a single cellular host. Thirdly, the apparent deep branching of the Megaviridae clades in the three phylogenetic trees is compatible with the hypothesis of an ancient origin of the viral genes. However, this hypothesis must be taken with caution because the statistical support values associated with the virus clade branches were low in two of the three reconstructed phylogenetic trees. Virus sequences may have emerged as a result of lateral gene transfers with different ancient cellular organisms or by vertical inheritance from a cellular ancestor. Indeed, two opposite hypotheses have been proposed for the evolution of giant virus genes: They could originate from a primitive cellular organism, which lost genes through a reductive process common to all parasites or, alternatively, most viral genes were acquired by horizontal gene transfer from their hosts (Filee 2007; Moreira and Brochier-Armanet 2008; Claverie and Abergel 2013). While this point is still debated, the genome sequencing of the recently identified Megavirus chilensis argues in favor of the first scenario (Arslan et al. 2011). If this scenario holds true, the enzymes of the UDP-GlcNAc biosynthetic pathways in Mimivirus may represent the remnants of an ancient glycosylation machinery, different from and intermediate between the modern prokaryotic and eukaryotic ones. Identification and genome sequencing of other giant viruses will provide further opportunities to analyze and solve this hypothesis.

Increasing evidence indicates that large DNA viruses encode complex systems for glycoconjugate formation, which include both enzymes for the production of the nucleotide-sugar substrates and the glycosyltransferases. The recent characterization of the glycans associated with the major capsid protein of the chlorella virus PBCV-1 highlights that large DNA viruses produce novel structures, never described in cellular organisms (De Castro et al. 2013). However, several open questions are still present about the formation of viral glycoproteins, including the mechanisms for glycan production and their subcellular localization, and about the origin of the viral enzymes and their relationships with those from cellular organisms. The identification of the viral enzymes involved in these pathways represents the starting point to clarify all these issues.

Materials and methods

Sequence and phylogenetic analyses

ORFs encoding enzymes possibly involved in the UDP-GlcNAc biosynthetic pathway were searched in the Mimivirus genome with the BLAST (Basic Local Alignment Search Tool) program using known sequences of Eukaryota, Prokaryota and Archaea enzymes as query (Badet et al. 1987; Mengin-Lecreulx and van Heijenoort 1996; Brown et al. 1999; Mio et al. 1999; Olsen and Roderick 2001; Milewski 2002; Milewski et al. 2006; Maruyama et al. 2007; Raczyńska et al. 2007; Namboori and Graham 2008). The identified Mimivirus candidates were then aligned back against Genbank “nr” database at NCBI using BLASTP. The best scoring homologs in viruses, eukaryotes, prokaryotes and Archaea were retrieved and aligned using the MAFFT (Multiple Alignment using Fast Fourier Transform) program (Katoh and Standley 2013). Gapped positions were removed before phylogenetic tree reconstruction using the maximum likelihood (ML) approach implemented in PhyML3.1 (Guindon et al. 2010) and the phylogeny.fr website (Dereeper et al. 2008). Selection of the best fitting protein substitution model for each phylogenetic reconstruction was performed with the ProtTest program (Abascal et al. 2005). ProtTest is a java program to find the best model of amino acid replacement for a given protein alignment; models included are empirical substitution matrices that indicate relative rates of amino acid replacement, and specific improvements (+G, +I and +F) to account for the evolutionary constraints imposed by conservation of protein structure and function. The Akaike Information Criterion (AIC) was used to find which of the candidate models best fits the data. The selected models were LG + gamma (G) + I, WAG + G and WAG + G for GFAT, GNAT and UAP protein families, respectively, with “LG (Le and Gascuel)” and “WAG (Whelan And Goldman)” standing for their corresponding substitution matrices, “G” standing for gamma correction for multiple substitutions and “I” standing for application of a proportion of invariant sites in the evolution model. The gamma shape parameter and proportion of invariant sites were estimated from the data.

Reagents, vectors and enzymes

All reagents used to express recombinant proteins and test enzymatic activities were provided by Sigma-Aldrich. Vectors
pGEX-6P-1 and pET-28a, from GE Healthcare (Milano, Italy) and Novagen (Merck Millipore, Darmstadt, Germany), respectively, were used for protein expression. Pfu polymerase was from Promega. Restriction enzymes, Antarctic phosphatase and Rapid Ligation kit were purchased from New England Biolabs (Ipswich, MA, USA).

Cloning

Mimivirus genomic DNA was purified as described (Piaccete et al. 2012). L619 (YP_003987136), L316 (YP_003986819) and R689 (YP_003987216) coding regions were amplified by PCR using Pfu polymerase and they were then cloned in pGEX-6P-1 using BamH I and Xho I restriction enzymes following the standard procedures. The three correct sequenced constructs were then used for transformation in E. coli BL-21 GOLD cells (Stratagene). L619 was also cloned in pET-28a but since a Nco I site was present in the insert, we used Afl III (which generates compatible ends with Nco I) and Xho I restriction enzymes to cut the L619 coding sequence. Nco I and Xho I were then used to cut the vector. The vector was sequenced and then transformed in E. coli BL-21(DE3) cells.

Recombinant proteins expression and purification

Expression of the GST fusion proteins was performed as previously described (Piaccete et al. 2012). Cells were grown at 22°C; isopropyl β-D-1-thiogalactopyranoside was added at the final concentration of 0.1 mM, when the culture OD600 reached 0.6. After induction cells were further incubated overnight at 22°C. Purification and proteolytically removal of the GST-fusion protein were done as reported (Piaccete et al. 2012). The recombinant His-tagged L619 protein was expressed as described above. Cell lysis and the subsequent purification steps using Nickel-charged ProBond™ resin (Invitrogen) were performed according to the supplier’s protocols. The purified proteins were concentrated and imidazol was removed using centrifugal filters Amicon® Ultra (Millipore). Final protein concentration was estimated by UV absorbance, using the following ε280: L619, 62830 M⁻¹ cm⁻¹; L316, 10430 M⁻¹ cm⁻¹; R689, 16390 M⁻¹ cm⁻¹ (Gasteiger et al. 2005).

Enzymatic activities

The GFA T activity of L619 was initially assayed by analyzing separately the isomerase activity on Fru-6P to produce Glc-6P and the glutaminase activity on glutamine to form glutamate and ammonia. The Glc-6P formation was analyzed by coupling the Glc-6P dehydrogenase (G6PD) reaction and following the NADPH formation. Reactions mixture contained 500 µM Fru-6P, 2 mM MgCl2, 1 mM NADP⁺, 5 U/ml G6PD (from baker’s yeast) in 50 mM KCl, 100 mM KH2PO4, pH 7.5, and the L619 gene product at different concentrations. Glutaminase activity was analyzed using the coupled reaction of GDH monitoring the increase in NADPH fluorescence. Reaction conditions were as follows: 500 µM Fru-6P, 2 mM MgCl2, 1 mM NADP⁺, 2 mM glutamine, 10 U/ml GDH in 50 mM KCl, 100 mM KH2PO4, pH 7.5. The kinetic of NADPH production was analyzed using a PerkinElmer LS 50B fluorescence with λex 340 nm and λem 450 nm.

The formation of GlcN-6P was tested using the Morgan–Elson colorimetric assay for acetyl-hexosamines (Reissig et al. 1955; Levy and McAllan 1959; Takahashi et al. 2003), with a protocol optimized for our needs. Samples, blanks and standards were prepared with 500 µM Fru-6P, 2 mM MgCl2, 2 mM glutamine in 50 mM KCl, 100 mM KH2PO4, pH 7.5, and were incubated for 30 min at 25°C. The L619 protein was added to the reaction samples, whereas the standards were prepared by addition of GlcN-6P at 12.5, 25, 50, 100, 200 and 400 µM concentrations. The standard curve was linear (r² > 0.99) in the range of concentration used for the assay. Negative controls contained the L619 protein in the buffer mentioned above without the substrates. The qualitative formation of glucosamine was also confirmed by GC-MS as described (Piaccete et al. 2012).

GNAT activity of the L316 gene product was tested using both the Morgan–Elson colorimetric assay, as described above, and the Ellman colorimetric assay, as described (Riddles et al. 1983) to determine the formation of free thiol groups. For the Morgan–Elson assay, reactions and the blank were prepared with 400 µM GlcN-6P, 5 mM MgCl2, 400 µM Acetyl-CoA in Tris–HCl 50 mM, pH 7.8, and were incubated at 30°C. Different amounts of the enzyme were added to the reaction mixture and aliquots were withdrawn at different time points. In this case, the acetylation step with acetic anhydride was omitted, in order to solely detect the GlcNAc-6P formed by the L316 acetyltransferase activity. GlcNAc-6P at the concentrations of 12.5, 25, 50, 100, 200 and 400 µM was used for standards. Negative controls contained only the enzyme or acetyl-CoA, respectively. For the Ellman assay, reaction conditions were set as above and a standard curve was prepared using 12.5, 25, 50, 100 and 200 µM GSH. Concurrently, negative controls were prepared using either L316 or Acetyl-CoA. GlcN-1P was also used as substrate in the same reaction conditions.

The UAP of the R689 gene product was analyzed by anion exchange HPLC. Chromatographic conditions were as described (Parakkottil Chothi et al. 2010; Piaccete et al. 2012). Identity of the formed nucleotide-sugars was confirmed by electrospray ionization - mass spectrometry (ESI-MS) as described (Piaccete et al. 2012). The nucleotidyltransferase activity was tested using 1 mM GlcNAc-1P and 1 mM UTP, GTP, CTP, ATP and dTTP as nucleotide donors. The GlcN-1P and Glc-1P were also tested as substrates, using UTP as donor. The reactions were performed in 50 mM Tris–HCl, pH 7.4, containing inorganic pyrophosphatase (10 U/ml). The reverse pyrophosphorylase reaction was analyzed using 1 mM UDP-GlcNAc or UDP-Glc incubated in the presence of 5 mM inorganic pyrophosphate. The effects of divalent cations, 5 mM MgCl2, CaCl2 and MnCl2, were also determined. All samples were incubated at 25°C; at the designated time points, aliquots were withdrawn and the enzyme was inactivated at 100°C for 3 min. After clarification, the samples were analyzed by HPLC.

Supplementary data

Supplementary data for this article is available online at http://glycob.oxfordjournals.org/.

Funding

This work was in part supported by Regione Liguria. C.B. is supported by a fellowship from Regione Liguria [P.O. C.R.O. 2015-2018].
**Abbreviations**

AIC, Akaike Information Criterion; ATP, adenosine triphosphate; BLAST, Basic Local Alignment Search Tool; CoA, coenzyme A; CTP, cytidine triphosphate; CryV, *Caterina roenbergensis* virus; CyS, cysteine; ESI-MS, Electrospray Ionization - Mass Spectrometry; Fru-6P, fructose-6-P; G, gamma; GC-MS, Gas Chromatography - Mass Spectrometry; Glc-6P, Glc-6-P dehydrogenase; GDH, glutamate dehydrogenase; GFAI, glucosamine-6P synthase; GIC-1P, glucose-1P; GlcN-6P, glucosamine-6P; GlcNAc, N-acetylglucosamine; GlmM, glucosamine-6P mutase; GNAT, glucosamine-6-phosphate N-acetyltransferase; GST, glutathione transferase; HPLC, high pressure (or high performance) liquid chromatography; LG, Le and Gascuel; MAFFT, Multiple Alignment using Fast Fourier Transform; NCLDV, nucleocytoplasmic large DNA viruses; PBCV-1, *Paramecium bursaria* chlorovirus 1; PGM, phosphoacetylglucosamine mutase; UAP, N-acetylglucosaminyltransferase and UDP-glucose dehydrogenase enzymes. **References**


