PhD in co-tutoring between University of Naples "Federico II" and Aix-Marseille University in the frame Italy-France Vinci Project

PhD in Chemical Sciences (XXXI Cycle) at University of Naples "Federico II"

PhD in Structural Biochemistry (LXII Cycle) at Aix Marseille University.

Understanding the innovative viral glycosylation machinery using a combination of chemical and structural methodologies

Supervisor

Cristina De Castro

Co-Tutor

Chantal Abergel

PhD Student

Anna Notaro

2015-2018

INDEX

Abstracts

Section I: Introduction

Chapter 1

Chapter 2

Thesis objective

Section II: Results and discussion

Chapter 3

Chapter 4

Chapter 5

Section III: Materials and methods

Chapter 6

Abbreviations

AMG: acetylated methyl glycosides

COSY: correlation spectroscopy

diNAcBac: N,N'-di-acetylbacillosamine

DTT: dithiothreitol

EI-MS: Electron ionization-mass spectrometry

GC-MS: gas-chromatography-mass spectrometry

GlcNAc: N-acetylglucosamine

HMBC: Heteronuclear multiple-bond correlation spectroscopy

HPLC: High performance Liquid Cromatography

HSQC: Heteronuclear single-quantum correlation spectroscopy

NMR: nucelear magnetic resonance spectroscopy

NOESY: nuclear Overhauser effect spectroscopy

PMAA: partially methylated alditol acetates

Qui2NAc: N-acetylquinovosamine

Rha: rhamnose

RhaNAc: N-acetylrhamnosamine

RP-HPLC: Reverse Phase- High Performance liquid Cromatography

TFMS: trifluoromethanesulfonic acid, also known as triflic acid

TOCSY: Total correlation spectroscopy

UDP: uridine diphosphate

UDP-diBacNAC: Uridine diphosphate-N,N'-di--acetylbacillosamine

UDP-GlcNAc: Uridine diphosphate-N-acetylglucosamine

UDP-Rha: Uridine diphosphate-rhamnose

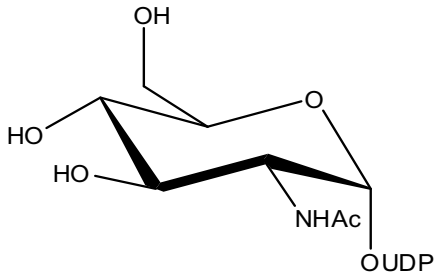
UDP-RhaNAc: Uridine diphosphate-N-acetylrhamnosamine

UDP-Vio. Uridine diphosphate-viosamine

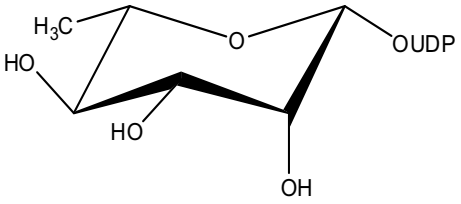
VioNAc: N-acetylviosamine

STD: Saturation-Trnasfer Difference NMR

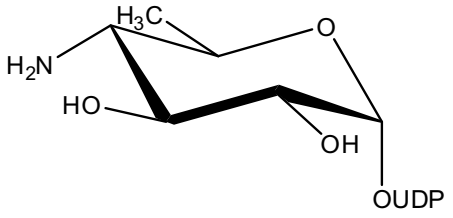
Sugars structures



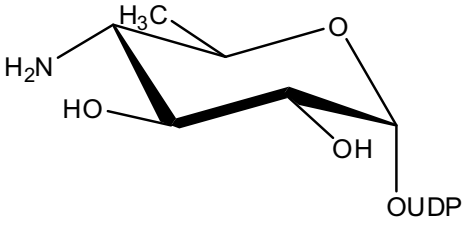
UDP-D-GlcNAc



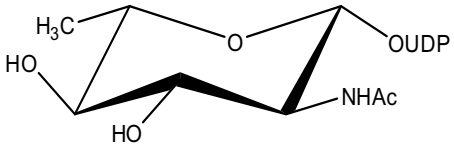
UDP-L-Rha



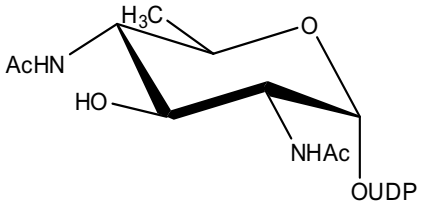
UDP-D-Vio



UDP-L-RhaN



UDP-L-QuiN



UDP-D-N,N'-di-acetylbacillosamine

Abstracts

Abstract

The aim of this thesis is the study of the innovative glycosylation machinery of the *Mimiviridae* family, using *Mimivirus*, *Moumouvirus australensis* and *Megavirus chilensis* as prototypes of lineages A, B and C, respectively.

In 2003 the discovery of *Mimivirus*, the first giant DNA virus infecting amoeba, challenged the traditional view of viruses. *Mimiviruses* are giant viruses due to the size of their virions, easily visible by light microscopy, with a diameter of 700 nm against 200 nm for "traditional virus". Their genomes encode 1000 proteins and count up to 1.2 Mbp, so they are as complex as the smallest free-living bacteria. *Mimiviruses* exhibit heavily glycosylated fibrils surrounding their capsid that differ in length depending on the lineages. Surprisingly, it was evidenced that they encode the proteins involved in their fibrils glycosylation.

The glycosylation of the fibrils was confirmed by the analysis of their sugar content, revealing that the major saccharide components were rhamnose, N-acetylglucosamine, and viosamine for *Mimivirus* and N-acetylglucosamine and N-acetylrhamnosamine for *Megavirus chilensis*. Until now, we lack information on the sugar composition of fibrils from members of the B lineage.

In this thesis, the innovative glycosylation machinery of these giant DNA viruses was investigated combining three different strategies: carbohydrate chemistry, bioinformatic and biochemical methodologies.

The carbohydrate chemistry methodologies allowed to elucidate the structures/composition of the glycans associated to the giant DNA viruses fibrils. *Mimivirus* fibrils are decorated with two distinct polysaccharides, called poly_1 and poly_2. Poly_1 is characterized by a linear disaccharide repeating unit made of 3)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→, with a pyruvic acid branched at position 4,6 of GlcNAc. Poly_2 has a branched repeating unit with the sequence 2)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→ in the linear backbone and rhamnose further branched at

position 3 by viosamine methylated at position 2 and acetylated at position 4. Regarding the novelty of the identified structures, they have no equivalent in eukaryotes, while some components were reported in bacteria. *Megavirus chilensis* has a different sugar composition of its shorter fibrils, with N-acetylglucosamine, N-acetylrhamnosamine and N-acetylquinovosamine as major components. Purification results suggested that *Megavirus* fibrils were decorated by more than one polysaccharides/oligosaccharide species, one having this trisaccharide: α-L-4OMe-RhaNAc-(1→3)-α-L-RhaNAc-(1→3)-α-L-RhaNAc-(1→. A preliminary analysis revealed that *Moumouvirus australensis* fibrils were decorated with glucosamine and quinovosamine in addition to the rare sugar, bacillosamine.

Starting from this experimental data, it was possible to identify new genes involved in glycosylation. As a result, the published nine-gene cluster of *Mimivirus* was extended to thirteen genes. A different cluster of fourteen genes was identified in *Moumouvirus australensis*, representing the first glycosylation gene cluster identified for the B lineage. A comparison of the glycosylation genes in the *Mimiviridae* family reinforced our finding that fibrils glycosylation was lineage specific. However, *Moumouvirus australensis* is an exception as it exhibits a cluster of glycosylation genes that is missing in other member of the B lineage.

Among the genes with the glycosylation cluster, the function of L142 was investigated in vitro, demonstrating that it is a N-acetyltransferase that acetylates the 4 amino group of viosamine. N-L142 represents the first virally encoded N-acetyltransferase.

To conclude, the fibrils of Mimiviridae are heavily glycosylated and the type of sugars and their organization depends on their lineage. The majority of the genes responsible for sugar production, sugar modification and glycosyltransferases were identified, strongly suggesting that Mimiviridae are autonomous for their fibrils glycosylation.

Résumé

The sujet de cette these portait sur la caractérisation de la machinerie originale utilisée par les *Mimiviridae* pour glycosyler les fibrilles entourant leurs capsides en travaillant sur les prototypes des 3 lignées connues, *Mimivirus* (A), *Megavirus chilensis* (B) et *Moumouvirus australensis* (C).

La découverte en 2003 de *Mimivirus,* le premier virus géant ADN infectant l'amibe, a bouleversé notre vision traditionnelle des virus. Les *Mimiviridae* sont des virus géants, visibles au microscope optique en raison des tailles de leurs particules (700nm contre 200nm plus les virus "classiques"). Leurs génomes de ~1.2 Mb codent pour un millier de protéines et sont aussi complexes que les plus petites bactéries non parasites. Les *Mimiviridae* infectant les amibes présentent des capsides couronnées de fibrilles très glycosylées de longueurs variables en fonction des lignées virales. De manière surprenante, ce sont des protéines virales qui sont responsables de la glycosylation des fibrilles. L'analyse de ces fibrilles a révélé que les sucres majoritaires étaient du Rha, GlcNAc et VioNAc pour *Mimivirus*, and GlcNAc and RhaNAc pour *Megavirus chilensis*. A ce jour, on ne connaissait pas la composition en sucres des fibrilles des virus de la lignée B.

Au cours de cette thèse, nous avons étudié la machinerie de glycosylation de ces virus géants en combinant différentes approches : la chimie des carbohydrate, la bioinformatique et la biochimie des protéines.

La chimie des carbohydrate a permis de déterminer la composition en glycans associés aux fibrilles virales et d'en résoudre les structures. Les fibrilles de Mimivirus sont décorées par 2 polysaccharides différents, poly-1 et poly-2. Poly-1 est caractérisé par la répétition d'un disaccaride linéaire fait de 3)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→, avec un pyruvate branché en position 4,6 du GlcNAc. Poly-2 présente une unité répétée branchée de séquence 2)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→ pour le squelette linéaire et du rhamnose branché en position 3 par de la

viosamine méthylée en position 2 et acétylée en position 4. Ces structures originales n'existent pas dans le monde eucaryote, tandis que certains de leurs composants ont déjà été identifiés dans le monde bactérien.

Megavirus chilensis présente une composition en sucres différente. Ses fibrilles plus courtes sont composées en majorité de N-acetylglucosamine, N-acetylrhamnosamine and N-acetylquinovosamine. Après purification, il semble que les fibrilles de Megavirus sont fait de plus d'un type de polysaccharide, l'un ayant présentant un trisaccharide de RhaNAc : α-L-4OMe-RhaNAc-(1→3)-α-L-RhaNAc-(1→3)-α-L-RhaNAc-(1→. L'analyse préliminaire des fibrilles de Moumouvirus révèle la présence de glucosamine et quinovosamine qui décorent les fibrilles et un sucre rare, la bacillosamine.

A partir de ces données expérimentales il devenait possible de rechercher de nouveaux gènes responsables de ces glycosylations spécifiques. Ainsi, le cluster de 9 gènes déjà publié de *Mimivirus* a pu être étendu à 13 gènes. Un cluster de 14 gènes a été d'autre part identifié dans le génome de *Moumouvirus australensis*, le premier cluster de gènes de la glycosylation identifié dans la lignée B. La comparaison des gènes de glycosylation des Mimiviridae renforce le fait que la glycosylation des fibrilles soit lignée spécifiques. Cependant, *Moumouvirus australensis* reste une exception avec un cluster de gènes de glycosylation absent des autres membres de la lignée B.

Parmi les gènes de glycosylation, l'analyse fonctionnelle *in vitro* de la protéine L142 a permis de démontrer qu'il s'agit d'une N-acétyltransferase qui acétyle le groupement amino 4 de la viosamine, le sucre rare composant les fibrilles de Mimivirus. N-L142 est ainsi la première N-acétyltransferase virale identifiée.

En conclusion, les fibrilles des *Mimiviridae* sont lourdement glycosylées and le type de sucres et leur organisation dépend de la lignée considérée. La majorité des gènes responsables de la production de ces sucres, de leur modification ainsi que les glycosyltrasnférases impliquées ont pu être identifiées, argument fort en faveur de

l’autonomie de ces virus par rapport à leur hôte cellulaire dans la glycosylation de leurs fibrilles.

Riassunto

Lo scopo di questa tesi è lo studio dell'innovativo macchinario di glicosilazione della famiglia *Mimiviridae,* usando *Mimivirus*, *Moumouvirus australensis* e *Megavirus chilensis* come prototipi delle cladi A, B and C, rispettivamente.

Nel 2003 la scoperta di *Mimivirus*, il primo virus gigante a DNA, ha cambiato la tradizionale visione sui virus. *Mimiviruses* sono virus giganti a causa delle dimensioni dei loro virioni, infatti sono visibili facilmente al microscopio ottico, con un diametro di 700 nm contro i 200 nm per un "virus classico". I loro genomi di 1.2 Mbp, codificano 1000 proteine e risultano più complessi dei più piccoli batteri non parassiti. *Mimiviruses* esibiscono fibrille altamente glicosilate intorno al loro capside, che differiscono in lunghezza a seconda della clade a cui appartengono. Sorprendentemente, è stato evidenziato che questi virus posseggono le proteine deputate alla glicosilazione delle loro fibrille.

La glicosilazione delle fibrille è stata poi confermata dalla lora analisi degli zuccheri, rivelando che i maggiori componenti saccaridici erano ramnosio, N-acetilglucosammina, e viosammina per *Mimivirus* and N-acetilglucosammina e N-acetilramonosammina per *Megavirus chilensis*. Fino ad oggi, non ci sono informazioni circa la composizione degli zuccheri delle fibrille per la clade B.

In questa tesi, il macchinario innovativo di glicosilazione di questi virus gianti a DNA è stato investigato integrando tre diverse strategie, dalla chimica dei carboidrati alla bioinformatica e biochimica strutturale.

Le metodologie della chimica dei carboidrati hanno permesso di elucidare le strutture o la composizione dei glicani connessi con le fibrille di questi virus giganti. Le fibrille di *Mimivirus* sono decorate con due dstinti polisaccaridi, chiamati poly_1 e poly_2. Poly_1 è caratterizzato da un'unità ripetitiva di 3)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→, e la GlcNAc lega in 4 e 6 l'acido piruvico. Poly_2 ha un'unità ripetitiva ramificata con 2)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→ nello scheletro

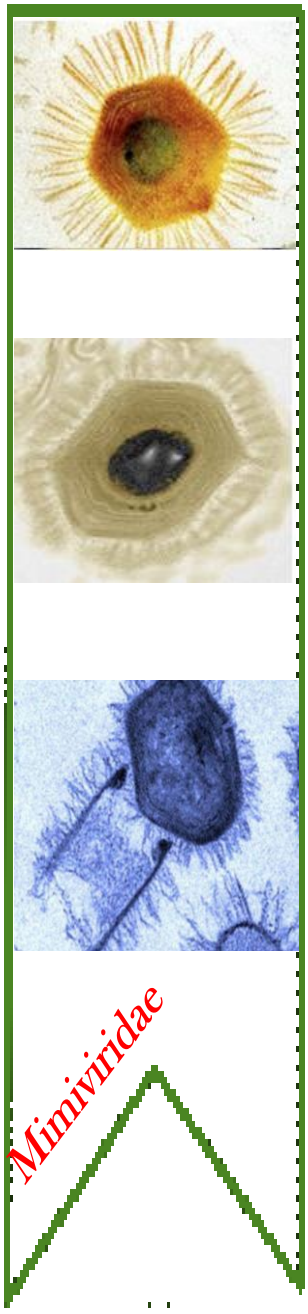
lineare; il ramnosio è ulterirormente ramificato in posizione O-3 con la viosamina che è metilata in posizione 2 e acetilata in posizione 4.Queste strutture non trovano riscontri nel mondo eucariotico, mentre alcuni component sono stati già identificati nel mondo batterico. Le fibrille di *Megavirus chilensis*, più corte di quelle di Mimivirus, hanno una diversa composizione in zuccheri con N-acetilglucosammina, N-acetilramnosammina e N-acetilchinovosammina come principali costituenti. La purificazione dei glicani connessi alle fibrille di *Megavirus* has suggerito che le fibrille erano decorate da più di una specie polisaccaridica o oligosaccaridica, una delle quali presentava questo trisaccaride: α-L-4OMe-RhaNAc-(1→3)-α-L-RhaNAc-(1→3)-α-L-RhaNAc-(1→. Un'analisi preliminare ha mostrato che le fibrille di *Moumouvirus australensis* erano decorate con un altro zucchero raro, la bacillosammina, in aggiunta alla glucosammina e chinovosammina.

A partire da questi dati sperimentali, è sato possible identificare nuvi geni coinvolti nella glicosilazione. Di conseguenza, il cluster di nove geni di Mimivirus, già pubblicato, è stato esteso a tredici geni. Un cluster di quattordici geni è stato identificato in *Moumouvirus australensis*, rappresentando il primo cluster di glicosilazione ad essere stato identificato per la clade B. Un'analisi comparativa dei geni della glicosilazione nella famiglia *Mimiviridae* ha consolidato i dati sperimentali, confermando che la glicosilazione delle fibrille è clade specifica. Inoltre, *Moumouvirus australensis* costituisce un'eccezione, in quanto il suo cluster genico di glicosilazione non è condiviso dagli altri membri della clade B.

Tra i geni della glicosilazione, la funzione di L142 è stata investigata in vitro, dimostrando che è un' N-acetiltransferasi che va ad acetilare la funzione amminica in posizione 4 della viosammina. N-L142 rappresenta la prima N-acetiltransferasi codificata da virus.

In conclusione, le fibrille della famiglia *Mimiviridae* sono altamente glicosilate e il tipo di zuccheri e la loro organizzazione dipende dalla clade a cui appartengono. La maggior parte dei geni responsabili della produzione degli zuccheri, delle loro

modifiche e le glicosiltransferasi sono state identificate, suggerendo fortemente che la famiglia *Mimiviridae* è autonoma per la glicosilazione delle fibrille.



Section I

Introduction

Chapter 1

Discovery of the giant DNA viruses: a new concept of viruses

What is a virus? Are viruses organisms? Are viruses alive? Scientists have tried to answer these questions since the discovery of the first virus, the causative agent of Tobacco Mosaic disease (Beijerinck, 1898; Ivanovski, 1892) in the XIX century. This signed the birth of virology (1886-1898) and we had to wait for 60 years before André Lwoff established the concept of virus (Lwoff, 1957). Interestingly, the viruses were only defined by negative properties that excluded them for the cellular world:

- ✓ Viruses are not retained by the Chamberland filter
- ✓ Viruses have only one type of nucleic acid (DNA or RNA);
- ✓ Viruses are unable to grow and divide;
- ✓ Viruses do not produce the enzymes for their energy production;
- ✓ Viruses have no protein translation apparatus.

In addition, Lwoff clearly stated that viruses were not organisms and not alive by concluding his article by the following statement: "viruses are viruses". One century after this first definition of "virus", these criteria are still accepted by the scientific community. For instance, many authors consider viruses as not living organisms, because they lack autonomy and metabolism (Moreira and Brochier-Armanet, 2008). In 2003 the discovery of *Mimivirus* (Scola, 2003) , the first giant DNA virus infecting amoeba from the Acanthamoeba genus, challenged the traditional view of viruses. Mimivirus was a dogma breaker, shaking the foundation of virology. "We are now entering a new era where the most basic concepts about viruses are revisited" (Claverie and Abergel, 2010). The same old questions were raised again by its discovery: What is a giant DNA virus? What are the boundaries between viruses and

cellular organisms? What is viruses origin? (Abergel et al., 2015; Forterre, 2017; Sharma et al., 2016)

The objective of this chapter is to describe the giant DNA viruses infecting amoebae, stressing their differences with "traditional virus". A detailed dissertation will be done on *Mimivirus* and the *Mimiviridae* family that are the focus of this thesis.

1.1 The first giant DNA virus: *Mimivirus*

Acanthamoeba polyphaga Mimivirus, more commonly known as *Mimivirus*, represents the first giant DNA virus infecting amoeba. In 1992, it was isolated from an amoebal co-culture from a water sample of the cooling tower of the hospital of Bradford, collected after a pneumonia outbreak. It was misidentified as a bacterium, due to the gram-positive staining of the spherical particles visible by light microscopy and was named Bradfordcoccus (Fig.1.1). Subsequently, the team of Rowbotham involved in the investigation of the outbreak, and then the Rickettesia Unit at the School of Medicine (URMITE, Marseille, France), tried different approaches to characterize this microorganism, including the 16S ribosomal gene DNA amplification which all failed. In order to shed light on the nature of this microorganism, it was visualized by electron microscopy in infected cells in 2003. The presence of icosahedral particles in the amoeba, finally, revealed its viral nature (Fig. 1.2)(Scola, 2003). Unexpectedly, the particles reached a diameter of 700 nm. Finally, the sequencing of its dsDNA genome revealed it was as complex as the smallest free-living bacteria with of 1.2 Mbp genome encoding 1000 proteins.

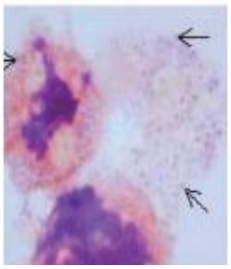


Fig.1.1. *Mimivirus* (arrows) in cytocentrifuged A. polyphaga as Gram-positive particles. Reproduced from La Scola *et al*., 2003.

According to these studies, Bradfordcoccus was renamed *Mimivirus* for "Mimicking microbe virus". *Mimivirus* was the first virus large enough to be visualized by light microscopy. It presented much bigger viral particles than any other known virus at

the time of its discovery (usually 0.2-0.3µm), reaching the size of some bacteria such as Mycoplasma and Rickettsia (300-500 nm), and was also bigger than Nanoarchaeum equitans, the smallest known archaea (400 nm): *Mimivirus* is definitively a giant due to the size of its virions.

The viral particle size has broken one dogma of the virology: viruses were supposed to be smaller than bacteria. It has led to reconsider the strategy for the isolation of viruses from different environments, in order to discover new giant viruses.

Mimivirus is a giant virus not only due to the diameter of its particles, but also by the dimension of its genome (1.2 Mbp). Usually viral genomes are measured in thousands of base pairs which correspond to a handful of genes. Prior to *Mimivirus* discovery, viral genomes bigger than the ones of small bacteria were already identified, such as *Paramecium bursaria* chlorella virus 1 (PBCV-1) with a genome up to 560 Kb (Van Etten et al., 1982).

*Mimiviruse*s is so different from the other viruses, that it started a new viral family, the *Mimiviridae* (Scola, 2003). This family was included in the group of the Nucleocytoplasmic Large DNA Viruses (NLCDVs), which consists of five families of viruses infecting eukaryotes: *Poxviridae*, *Asfarviridae*, *Phycodnaviridae*, *Ascoviridae* and *Iridoviridae* (Iyer et al., 2001; Yutin et al., 2009). The NLCDVs are dsDNA viruses that replicate partly or entirely in the cytoplasm of eukaryotic cells. They share a set of conserved genes. From the preliminary analysis of *Mimivirus* genome it shares 21 of these core genes (Iyer et al., 2001). However, the phylogenetic tree suggests an early divergence from the other viruses' family.

The unique features of *Mimivirus* are described in the following paragraphs.

1.1.1 *Mimivirus* virion morphology

Overview

Mimivirus viral particles from the outside to the inside, exhibit (Fig. 1.2):

- A layer of fibrils of 150 nm (Xiao et al., 2009);
- An icosahedral capsid (Kuznetsov et al., 2010; Scola, 2003; Xiao et al., 2009); the perfect symmetry is broken by the stargate (Zauberman et al., 2008) structure;
- Nucleoid (Kuznetsov et al., 2010; Seibert et al., 2011) containing the dsDNA genome.

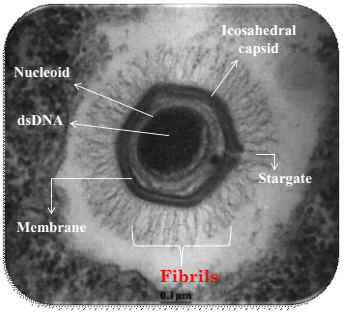


Fig. 1.2 Electron microscopy of Mimivirus in the vacuole of the amoeba cell. Copyright IGS.

Mimivirus fibrils

A dense layer of 150 nm fibrils covers *Mimivirus* mature particle, with the exception of the "stargate"(Xiao et al., 2009). Fibrils have also been observed in other NLCDVs, such as PBCV-1 (Cherrier et al., 2009; Kuznetsov et al., 2005; Zhang et al., 2011), CIV(Yan et al., 2009) and PbV01(Yan et al., 2005). However, *Mimivirus* fibrils are unique in terms of length, abundance, and complexity.

The atomic force microscopy (AFM) studies of *Mimivirus* fibrils detached from the virus, revealed some specific features (Fig. 1.3). The fibrils length is uniform, ranging from 125 to 140 nm, suggesting that the fibrils are either precisely "cut to measure" from longer precursor fibrils, or that they are assembled from small precursors until the linear aggregate reaches some predetermined length (Kuznetsov et al., 2010). The fibrils are probably composed of protein/s and should be of helical construction. It was observed that there are globular heads proteins attached to the fibrils shafts (Fig. 1.3). Concerning the fibrils precise anchoring, there are two different hypotheses:

1) The fibrils could be attached to extended polypeptide loops of the major capsid protein (MCP)(Kuznetsov et al., 2010; Xiao et al., 2009);
2) The fibrils could be anchored to a layer of protein positioned immediately above the capsid (Kuznetsov et al., 2010; Xiao et al., 2009).

Until now, it was not possible to discriminate between these hypotheses, but the evidence of multiple fibrils attached to one anchor supported the first hypothesis. Although the number of anchors was estimated to be the same as the number of depressions on the surface of the capsid (~3000), the total number of fibrils remains unknown, because it was not possible estimated the number of fibrils attached for one anchor.

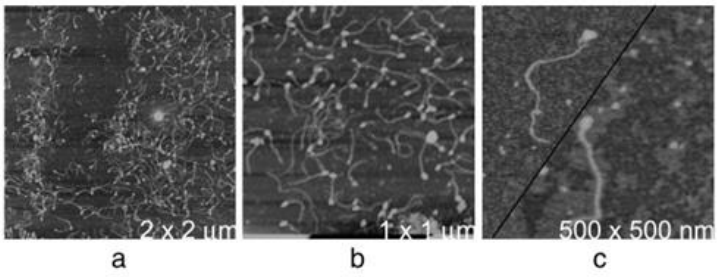


Fig.1.3 In **(a)** is a mass of surface fibers, most still attached to their head groups, shed from *Mimivirus* by a treatment with lysozyme and bromelain. A cluster of fibers still attached to the anchoring polypeptide at their proximal ends is also visible. In **(b)** the fibers are seen at higher magnification. Single fibers are identified due to their more or less uniform lengths and for the ellipsoidal protein heads. In **(c)** are two individual fibers, one of which is surrounded by a penumbra of some diffuse material. Reproduced from Kuznestov *et al* 2010.

Of particular interest is the evidence that the fibrils are heavily glycosylated. In the fig. 1.3 C there is a visible shadow surrounding the fibrils along their lengths, and later experiments suggested that it originated from poly-oligosaccharides decorations.

Indeed, first evidences came from the Gram-positive staining of the viral particles, suggesting the presence of peptidoglycan-like structures (Scola, 2003). However, this hypothesis was later disproved because it appeared that Mimivirus fibrils were retaining any stain.

The glycosylation of the fibrils was next evidenced by the analysis of Mimivirus particles sugar content, which revealed that the major saccharide components were Rha, GlcNAc, Vio. These data have excluded the presence of a true peptidoglycan (Piacente et al., 2012).

The in-depth investigation of *Mimivirus*, and other *Mimiviridae*, fibrils glycosylation is the subject of this thesis and it will be discuss in the following section.

The fibrils function is still under investigation. The current hypothesis is that the presence/absence of fibrils does not impair the viral replication in laboratory controlled condition, while they seem to play an important role in the adhesion process to the host cells (Rodrigues et al., 2015). These fibrils may favor the phagocytosis by the amoebal host and could be essential in the complex natural environment.

Mimivirus capsid organization

Mimivirus capsid consists of 20 large triangular faces joined at their edges to produce the required 12 five-fold vertices (Xiao et al., 2005).

The gene L425 encodes the *Mimivirus* major capsid protein, which is homolog to the Chlorella virus PBCV-1major capsid protein Vp54. Their N-terminal domains (position 1-287) share 46% identity and the C-terminal regions 40% identity at the protein level (positions 483-593). This homology suggested that *Mimivirus*, like other NCLDVs, had hexameric capsomers consisting of double jelly-rolls protein. A model of the L425 MCP was proposed based on the structure of Vp54(Azza et al., 2009; Nandhagopal et al., 2002) (Fig. 1.4).

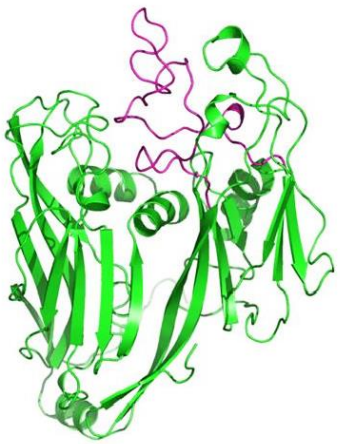


Fig. 1.4 Model of L425 Mimivirus capsid protein based on Vp54, the major capsid protein of Chlorella virus PBCV-1 (in green). In contrast with Vp54, Mimivirus MCP presented an insertion of ~190 amino acids into the DE loop (in pink). Reproduced from Azza *et al*., 2009

In contrast with Vp54, Mimivirus MCP presented an insertion of ~190 amino acids into the DE loop (in pink in Fig. 1.4) of the second jelly-roll fold (the β-strands along the polypeptide of each jelly-roll are named A to E), predicted to be made of 10 β-strands (Azza et al., 2009; Xiao et al., 2009).

In order to elucidate the *Mimivirus* capsid structure, structural studies were performed using *Mimivirus* untreated and enzymatically treated virions to remove the fibrils. The combination of AFM and cryo-EM reconstruction revealed that the capsid surface is composed of trimeric MCP (three monomers of the MCP constitute one capsomer), arranged in a very open honeycomb hexagonal array. These capsomers are organized in p6 plane group instead of the conventional p3 for the others icosahedral NCLDVs. The center-to-center distance between capsomers is 14 nm, and the triangulation number (T), in the capsid architecture, remains uncertain.

The perfect icosahedral symmetry is broken by a five-pronged star structure, called the 'stargate', at one unique vertex of the capsids (Zauberman et al., 2008). This structure is a unique feature of *Mimivirus*, and has been proved to be present on each virion by of sectioned *Mimivirus*-infecting acanthamoeba cells (Zauberman et al., 2008) and by cryo-EM and AFM on fibreless *Mimivirus* particles (Kuznetsov et al., 2010; Xiao et al., 2009) . The stargate arms have a width of about 500 Å, a thickness of about 400 Å and a length of about 2,000 Å. (Fig. 1.5).

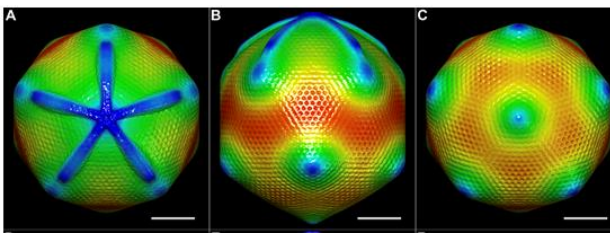


Fig. 1.5 (A–C) Surface-shaded rendering of cryoEM reconstruction of untreated *Mimivirus*. **(A)** Looking down the starfish-shaped feature associated vertex, (B) looking from one side, and **(C)** looking from the opposite side of the "starfish"-associated vertex. Reproduced from Xia *et al*., 2009.

The arms are inserted between two neighboring triangular faces causing a deviation of about five degrees from the geometry of a perfect icosahedron for these five faces. In addition, it was demonstrated that the stargate represents the only point of the *Mimivirus* virion which is not covered by fibrils (Xiao et al., 2009; Zauberman et al., 2008). This feature suggested that the stargate could be involved in the delivery of the dsDNA genome in the cytoplasm of the host cell. The role of the stargate as the genome delivery portal was revealed by the seminal work of Zauberman et al. (2008) followed by the one of Kuznestov et al. (2010). Between the stargate and the genomic nucleoid, there is a space that could be filled by the enzymes required to digest the structural component of the host, to initiate the infectious cycle. The existence of a unique portal for genome delivery reminds tailed bacteriophages.

Mimivirus nucleoid

A membrane surrounded structure, initially called "Seed" corresponds to the nucleoid which encloses the dsDNA genome of *Mimivirus* (Claverie and Abergel, 2010) as well as all the proteins necessary to initiate transcription. The estimated diameter of the nucleoid is about 340 nm and the dsDNA contained inside is 1.2×10^6 bp, which gives a packing density of 0.06 nm^3/bp. The DNA density in the nucleoid is relatively low and it has been suggested that the DNA could be associated with proteins (Kuznetsov et al., 2010). The nucleoid has a defined shape and a fixed position relative to the external capsid (Seibert et al., 2011).

Of particular interest is the presence of 300-500Å gap between the nucleoid and the outer capsid (Xiao et al., 2009). Except for PBCV-1, this space is absent in other viral particles for which the genome is closely surrounded by the capsid. The nucleoid plays an important role during the infection cycle, as discussed later.

Even more surprisingly, fibers with a 7 nm periodicity along their lengths (Fig. 1.6) were evidenced and probably localized in the space between the nucleoid and the outer capsid(Kuznetsov et al., 2010). These fibers are not composed by RNA or DNA. This type of structure was never observed in viral particles before, they are mechanically flexible and resistant to breakage. They could be cushion for the nucleoid. Another hypothesis is that these fibers might be the precursors for the fibrils surrounding the *Mimivirus* capsid. Additional studies are required to elucidate the function of these fibers (Kuznetsov and McPherson, 2011).

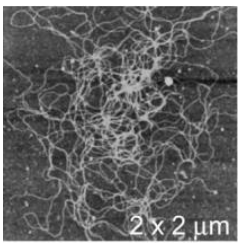


Fig. 1.6 It is a low magnification AFM image of large tangles of the fibers, which appear to have very long lengths. Reproduced from Kuzneston and McPherson, 2011.

1.1.2. Mimivirus genome

Mimivirus has a double stranded DNA genome of 1.18 Mbp encoding more than 1000 proteins, blurring the boundary between viruses and bacteria (Arslan et al., 2011; Legendre et al., 2010; Raoult, 2004; Renesto et al., 2006). *Mimivirus* genome is rich in adenine and thymine (A+T 72%). Codon and amino acid usage is thus dissimilar to the one of its amoebal host and is correlated with the high AT content of the *Mimivirus* genome (Colson et al., 2013a).

Although *Mimivirus* shares several core genes with the NLCDVs, its genome also exhibits genes encountered for the first time in a viral genome, raising the question on its possible origin. The most impressive feature is the presence of several genes related to the translation machinery, a process thought to be exclusively restricted to the cellular world. Some tRNA-like genes have been identified in other dsDNA viruses, such as chlorella virus (Van Etten and Meints, 1999), but *Mimivirus* presents an extended repertoire of genes involved in this process. For instance, it encodes several translation initiation, elongation and termination factors and four aminoacyl-

tRNA synthetase (aaRS): ArgRS, CysRS, MetRS and TyrRS. The aaRS are responsible of the attachment of the appropriate amino acid on its cognate tRNA and it has been demonstrated that these enzymes were all active and specific (Abergel et al., 2007).

In addition to a complete transcription machinery, *Mimivirus* encodes several types of DNA repair enzymes, including a remote homologue of the *E. coli* MutS, that can be used as marker in environmental studies (Wilson et al., 2014). Another unique feature of Mimivirus is the presence of three different topoisomerases of types IA, IB and II. The type IA was never found in viruses before (Raoult, 2004).

Although the *Mimivirus* genes are not interrupted by spliceosomal introns, some include inteins. Additionally, two type I introns were identified in the major capsid protein gene (Azza et al., 2009).

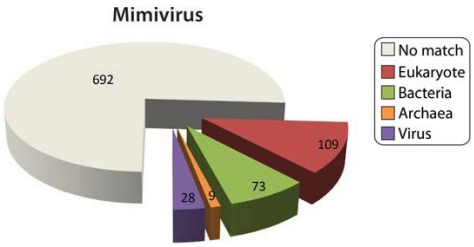


Fig. 1.7 Predicted proteins encoded in the genome of *Mimivirus*. Reproduced from Abergel *et al*., 2015.

Once again, *Mimivirus* invades the cellular world with the presence of genes encoding glycosylation machinery components. The hypothesis is that some of these genes could be involved in *Mimivirus* fibrils glycosylation. The investigation of these genes is part of this thesis and it will be discussed in the Chapter 2 and Section II.

Most of the genes with predicted functions (ORFs) have homologs in bacteria, archaea, eukaryotes or viruses. In addition, it is important to underline that 2/3 of *Mimivirus* genes are ORFans genes encoding unique proteins, without homologues in the cellular or viral world (Fig. 1.7).

This complex mosaic of genes raised again the question on Mimivirus origin, still under investigation.

1.1.3 Mimivirus replication cycle

The viral world is highly heterogeneous and complex, and different general replication strategies have been reported. Until recently, it was assumed that dsDNA viruses replicate in the host nucleus, while RNA viruses replicate in the cytoplasm. The first exception to this "rule" came from *Vaccina virus*, a dsDNA virus belonging to the *Poxviridae* family, for which it has been demonstrated that its replication cycle occurs entirely in the host cytoplasm (Carter et al., 2005). *Mimivirus*, as *poxviruses*, has a cytoplasmic replicative cycle (Mutsafi et al., 2010, 2014). In addition, as all the NLCDVs, *Mimivirus* builds a viral factory in the infected cell cytoplasm, a distinct "organelle", in which transcription and genome replication occurs (Netherton and Wileman, 2011; Suzan-Monti et al., 2007).

The first step of the infection is the adhesion of the viral particle to the surface of its amoebal host. It was suggested early on that this interaction was glycan mediated (Rodrigues et al., 2015). There are evidences that the viral particles are able to interact with the N-acetylglucosamine, which is an adhesion factor on the *Acanthamoeba* surface. In addition it was demonstrated that a *Mimivirus* mutant, M4 (Boyer et al., 2011), without fibrils, has a significantly lower level of attachment to the *Achanthamoeba* surface (Rodrigues et al., 2015) . Accordingly, it was clear that the *Mimivirus* heavily glycosylated fibrils could play an important role in the host adhesion-process. A complete elucidation of *Mimivirus* glycan nature could shed light on this interaction process.

The second step is the entry of *Mimivirus* into the amoeba through phagocytosis. This entry mechanism is related to the heterotrophic nature of *Acanthamoeba*, which feeds on bacteria. *Mimivirus*, mimicking a microbe due to its particle size and "sweet taste", can be phagocytized by *Acanthamoeba*. It was reported that phagocytosis can be triggered by individual particles larger than 0.6 µm (Korn and Weisman, 1967). Particle size of 750 nm could thus provide an evolutionary advantage to *Mimivirus* to be internalized by its amoebal host.

After internalization there is the opening of the stargate and the delivery of the nucleoid, containing the viral DNA and all the proteins needed to initiate the early genes expression, into the cytoplasm (Fig. 1.8 A to I). The biochemical signal responsible for the stargate opening is still unknown, but a recent study of Andrade (Andrade et al., 2017) and colleagues revealed that an important role in *Mimivirus* uncoating could be play by the phagosome acidification.

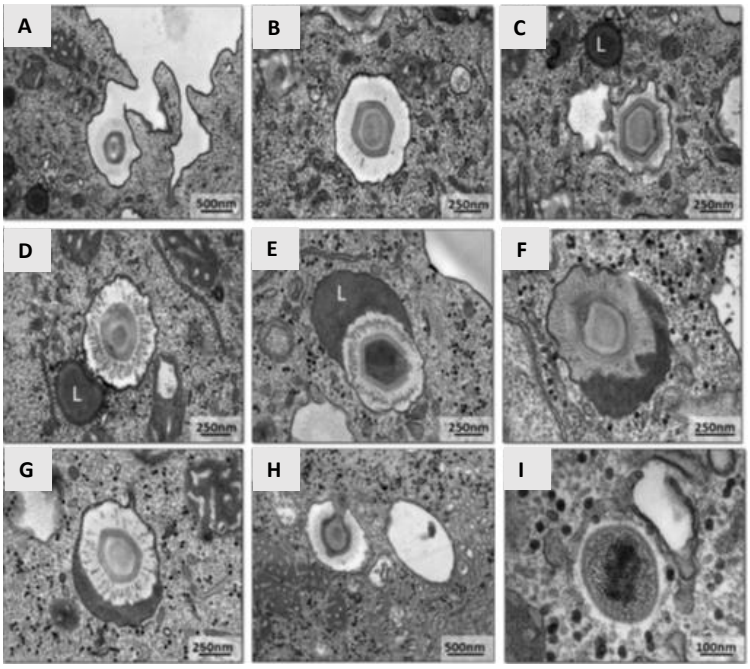


Fig. 1.8 Transmission electron microscopy images of *Mimivirus* particles entering cells by phagocytosis (**A** to **C**), the phagosome-lysosome fusion **(D to G)**, stargate opening **(H)**, and release of the viral seed **(I)**. Reproduced from Andrade *et al*., 2017.

After delivery of the nucleoid, the transcription and DNA replication takes place in the early viral factory (VF) (Mutsafi et al., 2010). The assembly of new viral particles takes place in the periphery of the mature VF.

A model for the capsid formation has been proposed through the combination of different imaging techniques, including scanning transmission electron microscopy

tomography (Mutsafi et al., 2013) and atomic force microscopy (Kuznetsov et al., 2013a) . It was evidenced that, 7.5 h post infection, host cisternae, from which 70 nm vesicles bud out, are recruited to the viral factory. These vesicles probably originate from the host cell endoplasmic reticulum and act as scaffolding for the assembly of the neo-synthesized capsids. In addition it appears that the L425 MCP acts as a scaffolding protein (Mutsafi et al., 2013). The morphogenesis of the capsid starts from the stargate, followed by the thickening of the protein layer in its vicinity (Kuznetsov et al., 2013a). Then the capsid is filled with the DNA through a transient aperture in the center of an icosahedral face, which is distinct from the stargate used for genome delivery: this is a unique characteristic of *Mimivirus*.

The final stage corresponds to the acquisition of the external fibrils, by an unknown process still investigated. To date, two hypotheses have been proposed:

1) In 2013 Kuzestov *et al.* proposed that the surface fibrils were acquired by viral capsids (Fig. 1.9.i) when they passed sequentially through a membrane embedded with a protein sheet, called in this study, integument protein (ii) and then through a protein sheet containing the fibrils (iii and iv) (Fig.1.9).

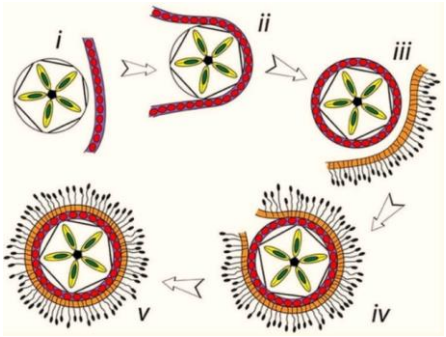


Fig. 1.9 Model for the acquisition fibrils layer. Reproduced from Kuznestov *et al*. 2013.

According to this model, layers of integument and fibrils are acquired as an envelope around the capsid (Fig. 1.9). In addition, this study suggested that the integument protein layer is acquired near the viral factory, while the fibrils are acquired near the cell periphery (Kuznetsov et al., 2013b).

2) In the 2017, Andrade *et al*., used transmission electron microscopy, to reveal at the periphery of the viral factory a less-electron dense region made of fibrils, which was called the fibril acquisition area (FAA). They suggested that the newly formed capsids were acquiring the fibrils layer by passing through this

region. This hypothesis is strengthened by the observation of a reduction of the FAA, after the release of the viral particles from the viral factory (Fig. 1.10) (Andrade et al., 2017).

After the acquisition of the fibrils layer, the *Mimivirus* particles are mature. The replication cycle ends with host cell lysis and the release of approximately 1000 mature particles.

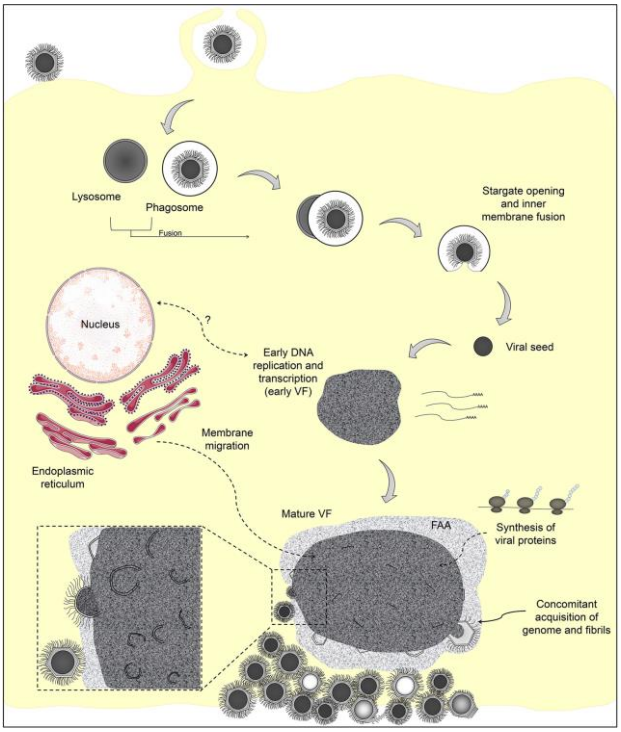


Fig. 1.10 Representative scheme of the *Mimivirus* replication cycle. Reproduced from Andrade *et al*., 2017.

1.2. *Mimiviridae* family

The *Mimiviridae* family was created in 2003 with the discovery of *Acanthamoeba polyphaga Mimivirus* (Scola, 2003), and it was included in the large monophyletic group of the Nucleo-Cytoplasm Large DNA viruses (NCLDVs; proposed order Megavirales (Colson et al., 2013b)). The *Mimiviridae* family is in constant growth and encompasses not only giant DNA viruses infecting amoeba, but also giant viruses infecting unicellular eukaryotes (Claverie and Abergel, 2018) (Fig. 1.11). This family was originally constituted by at list four subfamilies.

The **Mimiviridae group I** (also reported as Megavirinae subfamily (Gallot-Lavallée et al., 2017), this term will be used in this thesis) includes only the giant DNA viruses isolated by culturing on amoeba from water, soil, insect and human samples. Based on phylogenetic data, this group is divided in three lineages (Diesend et al., 2017):

- ✓ Lineage A – The prototype is *Acanthamoeba polyphaga Mimivirus* and it counts 18 members;
- ✓ Lineage B – The prototype is *Acanthamoeba polyphaga Moumouvirus* with a total of five members;
- ✓ Lineage C – The prototype is *Megavirus chilensis* and it includes 12 members.

The sizes of the icosahedral virions from the three lineages range between 400 and 600 nm diameter. They enclose a double stranded DNA genome ranging from 1.02 to 1.26 Mb, rich in A+T, encoding 930-1120 putative proteins. Their replication cycles are the same than the one described for *Mimivirus*.

The **Mimiviridae group II** consists of *Cafeteria roenbergensis* virus (CroV), infecting the marine phagotrophic flagellate *C. roenbergensis* (Fischer et al., 2010). In contrast with *Mimivirus*, CroV exhibits a smaller icosahedral capsid, about 300 nm, not decorated by a layer of fibrils, but it features 30 Å-long surface protrusions that are formed by the loops of the major capsid protein (Xiao et al., 2017). CroV has a dsDNA genome of 730 kbp, encoding 544 putative proteins. As for *Mimivirus*, it encodes translation factors, 22 tRNAs, DNA repair enzymes such as MutS and inteins (Fischer et al., 2010). The inclusion of CroV in the *Mimiviridae* family represents the first demonstration that *Mimiviruses* relatives could infect a wide spectrum of host.

The **Extended Mimiviridae family** (or *Mesomimivirine* family (Gallot-Lavallée et al., 2017)) is another subfamily gathering giant DNA viruses infecting unicellular algae.
The complete genome sequences of *Phaeocystis globosa* virus (PgV)(Santini et al., 2013), of *Aureococcus anophagefferens* virus (Moniruzzaman et al., 2014) and *Chysochromulina ericina virus* (CeV) (Gallot-Lavallée et al., 2017) clearly confirms a common origin with the Megavirinae. Most recently, it has been included in this family the TetV virus infecting *Tetraselmis* (Schvarcz and Steward, 2018), extending the spectrum of algae-infecting *Mimiviridae*.

The Klosneuvirinae is a new subfamily that encompass four linages: *Kloseneuvirus*, *Catovirus*, *Hokovirus*, *Indivirus* (Schulz et al., 2017). These giant DNA viruses have only been identified throughout metagenomics studies. In absence of physical isolated particles, their morphology and replication cycle are still unknown. The Kloseneuvirus's genome is 1.57 Mb, and thus more complex than the fully sequenced genomes of the previously isolated *Mimiviruses*.

A new subfamily could be represented by the two ***Tupanviruses*** strains (*Tupanvirus soda lake* and *Tupanvirus deep ocean*), which is considered as a sister group of the *Mimiviruses* (Abrahão et al., 2018).

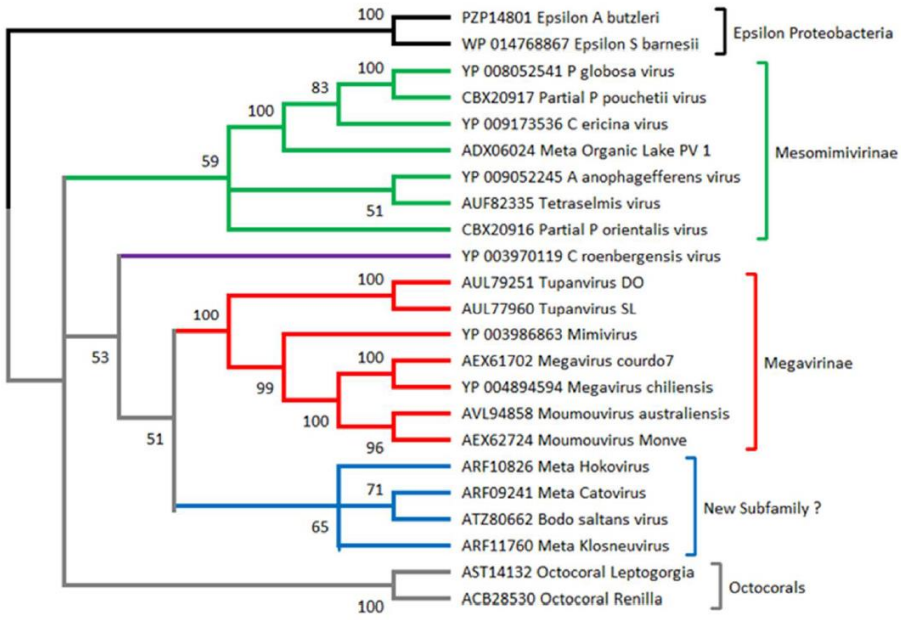


Fig. 1.11 Phylogeny of the Mimiviridae based on MutS7. MutS sequences from Epsilon proteobacteria (from the genus Arcobacter, and Sulfurospirillum) are used as root, as they are most likely the source of the Mimiviridae MutS7 gene (MutS homologs are present throughout the epsilon division). The tree suggests with great confidence that the mitochondrial MutS present in octocorals (in grey) and that present in all Mimiviridae have a common origin. Reproduced from Claverie *et* Abergel, 2018, Viruses.

In conclusion, the *Mimiviridae* family is in constant expansion, including *Mimivirus* relatives infecting hosts other than amoeba. With the discovery of new *Mimiviruses,* the picture become more and more complex, opening the question on their origin. An in depth analysis of these viruses could play a key role to establish the criteria

for the classification of the new members that will be isolated in the future (Claverie and Abergel, 2018).

It is important to precise that to date there is no official classification including the recently identified subfamilies in the *Mimiviridae* family.

In this thesis, I focused my attention on *Mimivirus* (Lineage A), *Megavirus chilensis* (Lineage C) and *Moumouvirus australensis* (Lineage B). In addition, I took in account the newly discovered *Tupanviruses* to perform some comparative analyses of the glycosylation pathways.

1.2.1 Acanthamoeba polyphaga moumouvirus and *moumouvirus australensis*

Acanthamoeba polyphaga moumouvirus represents the prototype of the lineage B of the *Mimiviridae* infecting Acanthamoeba. It was isolated in 2008 by inoculating *A. polyphaga* with water from an industrial cooling tower (France)(La Scola et al., 2010). *A. polyphaga moumouvirus* presents the same morphology as *Mimivirus* and *Megavirus chilensis* virions, except for its smaller capsid size (420 nm, instead of 500-520 nm) and fibrils lengths (100 nm instead of 125-175 nm for *Mimivirus* and *Megavirus,* respectively). Its replication cycle is similar to that of *Mimivirus*. *Moumouvirus* has a dsDNA genome of 1 Mbp, smaller than the ones of *Mimivirus* (1.18 Mb) and *Megavirus* (1.26 Mbp). Based on the phylogenetic data Moumouvirus is closer to *Megavirus*.

The lineage B includes four strains: *Moumouvirus*, *Moumuvirus monve* (La Scola et al., 2010), *Moumouvirus goulette*(Boughalmi et al., 2013) and *Moumouvirus Saudi* (Bajrai et al., 2016).

In addition to these four strains, the IGS laboratory in Marseille identified two new members, named *Moumouvirus australensis* and *Moumouvirus maliensis* (unpublished).

Moumouvirus australensis (Fig. 1.12) was isolated from a sample of muddy water collected in the Ross river mangrove near Townsville. Its morphology and replication cycle are similar to that of the other *Mimivirus*es. The genome of *Moumouvirus australensis* was assembled in one counting of 1,098,002 bp in length with 25% G+C content. It has been attributed to the lineage B, although it is clearly divergent from the other members (Jeudy et al., submitted). As for *Mimivirus*, all the four strains present genes involved in protein glycosylation (Section II).

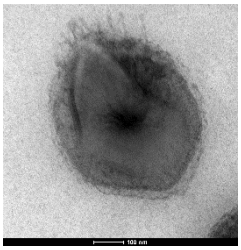


Fig. 1.12 Electron microscopy of *M. australensis* particle. Copyright IGS.

Moumouvirus maliensis was isolated from Mali and the analysis of its genome revealed that is very close to *Moumouvirus australensis* (unpublished*)*.

The B lineage seems to be more divergent from the other two lineages, suggesting that it underwent a relaxed selection and faster evolution.

1.2.2 Megavirus chilensis

Megavirus chilensis is the prototype of the lineage C. It was isolated in 2011 from a water sample collected in the Pacific ocean, off the coast of central Chile, and was able to replicate in the fresh water *Acanthamoeba castellanii* (Arslan et al., 2011). *Megavirus chilensis* is between the Megavirinae (Gallot-Lavallée et al., 2017) the most distant from *Mimivirus*. Although, the morphology o*f Megavirus* resembles that of *Mimivirus*, it is possible to distinguish it based on some specific features (Fig. 1.13).

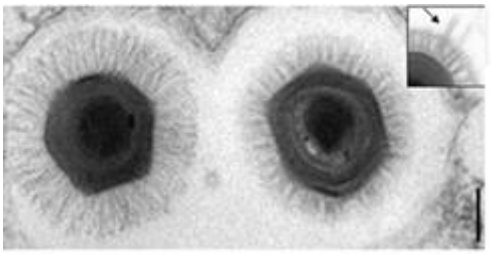


Fig. 1.13 Electron microscopy image of Megavirus compared to *Mimivirus*. (A) *Mimivirus* (Right) and Megavirus (Left) particles are in a same vacuole (coinfection). (Insert) Cowlicks (arrow) are often seen in the Megavirus fibrils layer. (Scale bars200 nm). Reproduced from Arslan *et al*., 2011.

The fibrils layer of *Megavirus* is different in length (fibrils of 75 ± 5 instead of 120 ± 5 nm) and thickness (440 ± 10 nm against 390 ± 10 nm). In addition, *Megavirus* mature particles exhibit one or two patches of slightly longer and denser fibrils, coined "cowlicks" (Fig. 1.13, insert).

The major difference between *Megavirus* and *Mimivirus* is their genomes (Fig. 1.14). In the first place, the *Megavirus* genome is larger than *Mimivirus* (1.26 Mb against 1.18 Mb) and it was the largest until the two *Tupanviruses* were discovered in 2018 (see next paragraph).

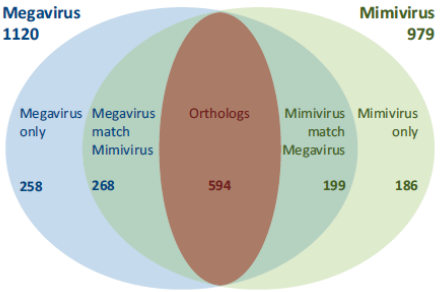


Fig. 1.14 Comparison of *Mimivirus* and Megavirus gene contents. Reproduced from Arslan *et al*., 2011.

Its genome encodes 1120 putative proteins (instead of 979 for *Mimivirus*), of which the 23% have no homolog in *Mimivirus*. *Megavirus* and *Mimivirus* share 50% of their genes with an average sequence similarity of 50% identical residues at the protein level (Fig. 1.14).

As *Mimivirus*, it presents several cell-like genes, including the ones predicted to encode the transcription apparatus. *Megavirus* has three additional aaRS: TrpRS, IleRS and AsnRS, which is the first viral type II aaRS , demonstrating that the viral aaRS are not confined to class I (Arslan et al., 2011) . *Megavirus*, as *Mimivirus*, presents several genes involved in the glycosylation of its fibrils. Part of these genes will be presented in the next chapter. The differences in the glycosylation between *Megavirus* and *Mimivirus* are the subject of this thesis (Section II).

It has been hypothesized from Arslan *et* colleagues that the differences in the genome content of *Megavirus* and *Mimivirus* could be related to: linage specific gains and losses of genes; lineage specific gene family expansion or deletion; insertion/migration of mobile elements (intron, intein).

1.2.3 *Tupanviruses*

In 2018, the long tailed *Mimivirus*es infecting amoeba were discovered: *Tupanvirus soda lake* and *Tupanvirus deep ocean* (Fig. 1.15)(Abrahão et al., 2018). *Tupanvirus soda lake*, as suggested by its name, was isolated from Soda lakes, which are considered the most extreme aquatic environments (high salinity and pH). Similarly, *Tupanvirus deep ocean* was isolated from ocean sediments collected at a depth of 300 m in Brazil.

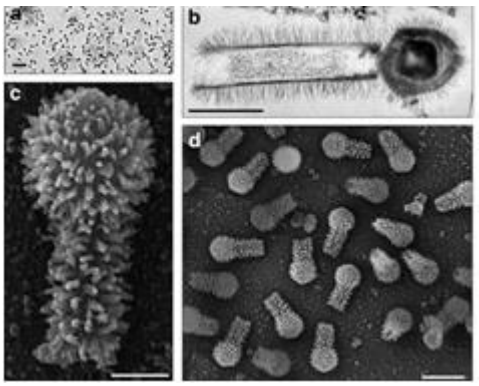


Fig. 1.15 Tupanvirus soda lake particles. **(a)** Optical microscopy of Tupanvirus particles after haemacolour staining (1000 ×). Scale bar, 2 µm. **(b)** Super particle (>1000 nm) observed by transmission electron microscopy (TEM). Scale bar, 500 nm. **(c, d)** Scanning electron microscopy (SEM) of Tupanvirus particles. Scale bars 250 nm and 1 µm, respectively. Reproduced from Abrahão *et al*., 2018.

Regarding their morphology, they exhibit a capsid of 450 nm also decorated with fibrils. In contrast with the others *Mimiviruses*, they present a large cylindrical tail also decorated with fibrils (550 nm extension; 450 nm diameter including fibrils) (Fig. 1.15). The tail of *Tupanvirus* represents the longest one in the virosphere.

The mature viral particles have a length of 1.2 µm, but some can reach up to 2.3 µm due to a variation in the tail lengths.

Tupanvirus soda lake and *Tupanvirus deep ocean* present the most complex genomes in the *Mimiviridae* family, with sizes of 1.43 Mb and 1.51 Mb, respectively. Their

most surprising feature is the presence of the most complex translation machinery in the virosphere, with 70 tRNA, 20 aaRS (against 4 for *Mimivirus* and 7 for *Megavirus chilensis*), 11 factors for all the translational phases, and factor related to tRNA/mRNA maturation and ribosome protein modification. Only the ribosomal genes are lacking from this complete translation machinery.

Are *Tupanvirus* fibrils glycosylated? Does *Tupanvirus* encodes its glycosylation machinery? A glyco-related genes prediction was conducted in Section II.

1.3The discovery of distinct families of large and giant DNA viruses infecting amoeba

The *Mimiviridae* family is not the only family of DNA viruses infecting Acanthamoeba. Starting from 2003 until now, three distinct families have been described: the *Pandoraviridae*, *Pithoviridae* and *Mollivirus.*

Pandoraviridae **(2013)**

Ten years after the discovery of *Mimivirus*, the *Pandoraviridae* family, a new family of giant DNA viruses infecting acanthamoeba, was discovered. Their virions present an amphora-shaped morphology completely different from that of *Mimiviridae* (Fig. 1.16), and unique genomic characteristics. The first two strains, named *P. salinus* and *P. dulcis*, were isolated in 2013 by co-colturing *Acanthamoeba castellanii* with samples recovered from sediments collected on the coast of central Chile and from mud collected from a freshwater pond close to Melburne (Philippe et al., 2013). Subsequently, in 2015, a previously misidentified

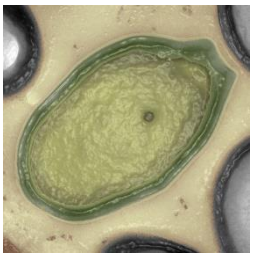


Fig. 1.16 Electron microscopy image of *Pandoravirus* particle. Reproduced from Philippe *et al*., 2013.

parasite (Germany) was sequenced and named *P. inopinatum* (Antwerpen et al., 2015). Most recently, three new strains have been isolated by the IGS laboratory: *Pandoravirus quercus*, isolated from ground soil in Marseille (France); *Pandoravirus neocaledonia*, isolated from the brackish water of a mangrove near Noumea airport (New Caledonia); *Pandoravirus macleodensis*, isolated from a freshwater pond near Melbourne (Australia) (Legendre et al., 2018). Three additional strains were isolated from Brazil, *P. massiliensis*, *P. braziliensis*, and *P. pampulha* (Aherfi et al., 2018).

They all share amphora-shaped particles of about 0.8-1.2 µm in length and 0.5 µm in diameter (Fig.19). Instead of the "stargate" corresponding to the apex of the *Megavirinae* (Gallot-Lavallée et al., 2017), there present an apical pore closed by a diffuse plug. There are no noticeable differences between the different members.

The *Pandoraviruses* have dsDNA genomes ranging from 1.84 to 2.45 Mb, rich in G+C (60%). The numbers of ORFs is ranging from 1552 (*P. inopinatum*) up to 2500 (*P. salinus*) (Legendre et al., 2018). In contrast with the other giant DNA viruses, there is recognizable gene encoding for a major capsid protein. The replication cycle shares with *Mimivirus* only the phagocytosis-dependent entry mechanism.

Genes involved in the glycosylation seem to be present but need to be investigated.

Eighty percent of their genes correspond to ORFans. Based on phylogenetic analysis, it appears that there are two clades within the *Pandoraviridae* family: clade A (*P. salinus, P. dulcis*, *P. quercus*, *P. inopinatum*) and clade B (*P. neocaledonia*, *P. macleodensis*).

Pithoviridae **(2014)**

One year after the discovery of *the Pandoraviridae,* another giant DNA virus infecting amoeba was isolated that was the first member of the proposed *Pithoviridae* family. The founding member, *P. sibericum*, was isolated from a 30000-year-old sample of Siberian permafrost. The family is currently made of at least four clades, one enclosing *Pithovirus sibericum* (Legendre et al., 2014) and its modern relative

Pithovirus massiliensis (Levasseur et al., 2016), isolated from a sewage sample, another enclosing Cedratvirus(Andreani et al., 2016), isolated in *A. castellanii* from an Algerian environmental sample and the last one enclosing Orpheovirus (Andreani et al., 2018), isolated from a rat stool sample using *Vermamoeba vermiformis* as host cell.

The morphology of *Pithovirus* (Fig. 1.17) resembles that of *Pandoraviruses*, but with larger ovoid viral particles. Recently, cryo-EM images revealed that *Pithovirus* virions exhibit a variability in length (major axis of the particle) and in width (minor axis of the particle), ranging from1,350 to 1,650 nm in length and from 750 to 850 nm in width (Okamoto et al., 2017). In contrast with *Pandoraviruses*, the apical pore is plugged by a "cork", composed of highly regular hexagonal honeycomb-like grid. The *Cedratvirus*' viral particle has an ovoid-shaped, but in contrast with the two *Pithoviruses*, exhibits two crocks at each extremity of the virion. Their replication cycle is as for the Mimivirinae, namely cytoplasmic. *Pithovirus'* genome (600 Kbp) is smaller than *Pandoraviruses* and *Mimiviruses* genomes. It is rich in A+T and encodes about 450 proteins. *Pithoviruses* also seem to present genes involved in glycosylation.

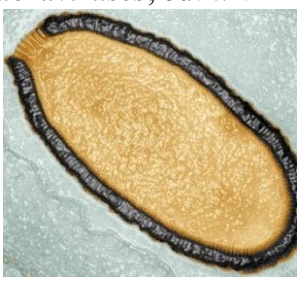


Fig. 1.17 Electron microscopy image of Pithovirus particle. Copyright IGS.

Mollivirus **(2015)**

Mollivirus sibericum represents a new giant DNA virus infecting amoeba, which was isolated from the same 30,000-years-old Siberian permafrost sample as *P. sibericum* (Legendre et al., 2015). Its morphology (Fig. 1.18) is very different in comparison with those of previously isolated viruses, with a spherical particle of 500-600 nm, covered by 2-3 layers of fibrils.

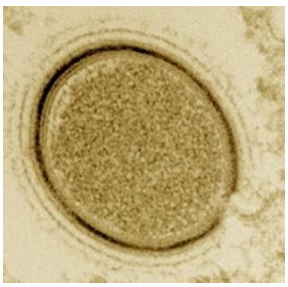


Fig. 1.18 Electron microscopy image of *Mollivirus* particle. Copyright IGS.

Mollivirus has a 651 Kbp G+C rich dsDNA genome (60%), encoding 523 putative proteins out of which 65% are ORFans.

Are *Mollivirus* fibrils glycosylated? Does *Mollivirus* have genes involved in glycosylation? These questions still need to be addressed.

Chapter 2

Viral glycosylation

Glycosylation is one of the most abundant modification of proteins and it is classified as N-, O- and C-glycosylation, based on the type of amino acid to which the sugar moiety is attached. Glycosylation influences the protein stability, solubility, rigidity, interactivity, transport and signaling. This proteins' modification occurs not only in all domains of the life (Eukarya, Bacteria, Archaea) but also in viruses. It has been reported in literature that eukaryotic viruses use the ER and Golgi apparatus of the host cell to add and to remove sugar to their nascent glycoproteins located on the envelope/surface. Viral proteins can be glycosylated at relatively few sites or can be densely glycosylated with N-linked or O-linked glycans. For instance, *Ebola* virus glycoprotein is modified by a very high content O-linked glycans (Jeffers et al., 2002), while the HIV-1 glycoprotein gp160 presents multiple N-linked glycans (Pabst et al., 2012).

Viral glycosylation play a key role during the virus infection (Bagdonaite and Wandall, 2018) (Fig. 2.1):

- ✓ Attachment and entry: viruses, such as Ebola virus, SARS coronavirus, use the N-linked glycans as a bite to interact with cellular lectins or mannose receptors, which later promote their uptake in the cells. Less representative are the cases in which the O-linked glycans participate in the host-virus interaction;
- ✓ Assembly and exit: glycosylation is involved in viral particle formation and release;
- ✓ Immunity: glycans shield immunodominant epitopes from immune recognition of the host cell.

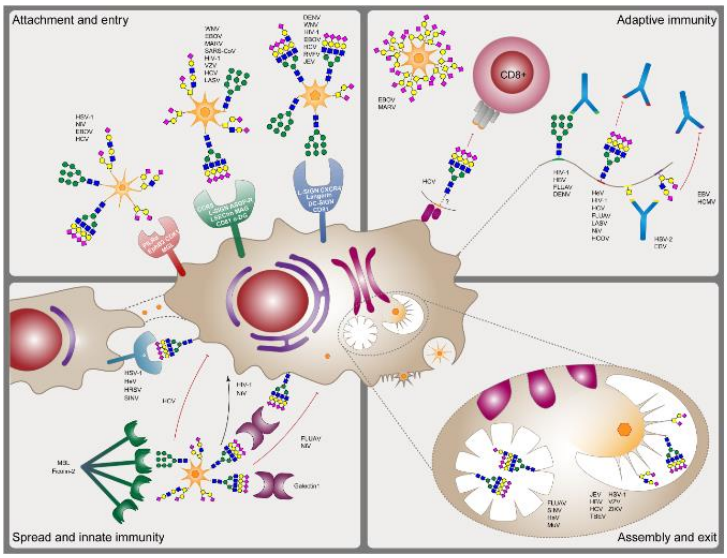


Fig. 2.1 Roles of glycosylation in the biology of enveloped viruses. Reproduced from Bagdonaite *et al*., 2018.

Due to the importance of the glycosylation in virus biology, viruses have developed several strategies to exploit the host glycosylation machinery. For instance, several viruses are able to change the expression level or activity of some host glycosylation elements, in order to modify the final glycan structure. In addition, several viruses to increase their survival and propagation include in their genome glycosyltransferase genes. The bacteriophages T2, T4 and T6 of *E. coli* encode a glycosyltransferase that transfer glucose residue to the hydroxymethylcytosines (HMC) of their DNA, protecting it against specific restriction endonucleases encoded by the chromosome of the host or its plasmid (Markine-Goriaynoff et al., 2004) . Some *baculoviruses* have developed a different strategy to improve their propagation. *Baculoviruses* infecting *Lepidoptera* larvae express a glucosyltransferase able to attach glucose residues to the hormone ecdysone, inactivating the hormone and subsequently the pupation of the host. This strategy allows these viruses to complete their infection cycle (Markine-Goriaynoff et al., 2004).

2.1 **The giant DNA viruses encode an autonomous glycosylation system.**

In the last years, it has become increasingly evident that, in contrast to other viruses, some members of Nucleo-Cytoplasmic Large DNA Viruses (NCLDVs) group encode most, if not all, the machinery required for the glycosylation of their structural proteins. The concept of autonomous viral glycosylation system emerged from *Paramecium bursaria* chlorella virus 1 (PBCV-1) which encodes the glycosylation machinery for its major capsid protein (Vp54), whose glycans structure was disclosed only recently(De Castro et al., 2013) . Vp54 presents four N-linked glycosylation sites. The structure of the N-linked glycans revealed unique feature:

- ✓ The glycan is bound to asparagine (Asn) via β-glucose, instead of N-acetylglucosamine or N-acetylgalactosamine typical of N-linked glycoproteins;
- ✓ The presence of L-Fucose fully substituted (it was observed only in the phosphoglycan epitope of *Trypanosoma cruzi*) and D-Rhamnose (a very rare sugar);
- ✓ The Asn acceptor is not in the classical consensus sequence Asn-X-Ser/Thr of N-glycosylation both in eukaryotes and in prokaryote;
- ✓ The Vp54 N-linked glycan structure is unique, does not resembles any structure found in the three domains of life.

PBCV-1 not only encodes for several glycosyltransferases presumably involved in the glycan structure of Vp54, but also has the enzymes to produce the nucleotide-sugars used as substrate for glycan formation. PBCV-1 has a functional pathway for GDP-L-Fucose and D-Rhamnose, which represents the first example of nucleotide-sugar metabolism described in viruses (Tonetti et al., 2003).

Starting from PBCV-1, enzymes for glycan biosynthesis have been identified in the giant DNA viruses of amoeba, in particular in members of the *Mimiviridae* family,

as well as in new families such as the *Pandoraviridae*, *Pithoviridae*, and *Mollivirus*, a list in constant growth. To date, functional pathways for nucleotide-sugars have been identified for *Mimivirus* and *Megavirus chilensis* (see next paragraphs).

The subject of this thesis is the investigation of the innovative glycosylation system of some members of *Mimiviridae* family (***Mimivirus***, ***Megavirus chilensis***, ***Moumouvirus australensis***) combing chemical, biochemical and bioinformatics strategies. Starting from the experimental knowledge acquired on *Mimivirus*, *Megavirus* and *Moumouvirus australensis*, an extended prediction of glycosylation related genes was done for all the members of the *Mimiviridae* family for which the complete genome was available, including the recently discovered *Tupanviruses*.

The study of the glycobiology of these giant DNA viruses is important for the annotation of new genes, the usage of their enzyme for biotechnological purposes, and for the advancement of glycobiology.

2.1.1 The *Mimivirus* UDP-L-Rhamnose pathway

The UDP-L-Rhamnose pathway represents the first functional pathway identified in the giant DNA virus *Mimivirus* (Parakkottil Chothi et al., 2010). In this pathway involves the products genes of *R141* and *L780*, whose activity has been demonstrated in vitro. *R141* encodes for a UDP-D-glucose 4,6-dehydratase (UGD), which catalyzes the removal of a water molecule from the carbon 4 and carbon 6, leading to an intermediate UDP-4-keto-6-deoxy-D-glucose. This compound is the substrate of L780, a NADPH-dependent 3,5-epimerase/4-reductase (UGER), generating the UDP-L- Rhamnose (Fig. 2.2).

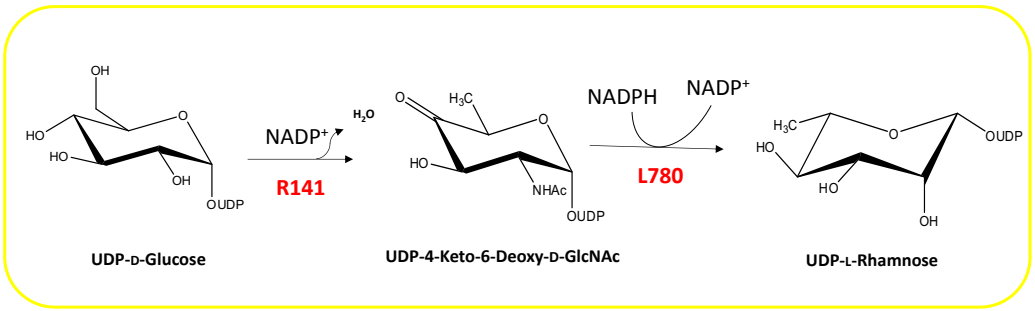


Fig. 2.2 UDP-L-Rhamnose pathway in *Mimivirus*. This biosynthetic mechanism follows the eukaryotic pathway for the UDP-L-Rhamnose. *R141* and *L780* (in red) have an eukaryotic origin.

L-rhamnose is a very common sugar in the lipopolysaccharides of the gram-negative bacteria and it is a component of the cell wall in plants. However, the biosynthesis is quite dissimilar in bacteria and eukaryote. The major difference regards the reaction of epimerization and reduction that is made by one bifunctional enzyme in plants and by two different enzymes in bacteria. UDP-L-Rhamnose production in *Mimivirus* resembles the mechanism used by plants.

An in-depth analysis of *R141* and *L780* genes was done in order to understand the origin of this pathway. The G+C content of these two genes is similar to that of *Mimivirus* genome, excluding the hypothesis of a horizontal genes transfer from its amoebal host. Phylogenetic analysis has suggested that *Mimivirus* probably acquired this pathway from a protozoan ancestor.

The genes involved in the UDP-L-Rhamnose production are expressed at the late stage of *Mimivirus* infection, thus suggesting which is the timing of the fibrils decoration.

2.1.2 The *Mimivirus* UDP-D-Viosamine pathway

Starting from the identification of UDP-L-Rhamnose pathway, several glycan related genes were found, including glycosyltransferase and sugar-modifying enzymes. A nine-gene cluster, involved in glycans formation was identified (Piacente et al., 2012) (Fig. 2.3). Inside this cluster, only the function of three genes has been demonstrated in vitro: R135 is a GMC-type oxidoreductase that is located outside of *Mimivirus* fibrils, which seems to play a role in the virus entry (Klose et al., 2015); R141 that is an UDP-D-glucose 4,6 dehydratase (UGD); L136 that is an aminotransferase. In addition, in the cluster there are several putative glycosyltransferases (L137, L138, and L139, R140), a putative N-acetyltransferase (N-terminal domain of L142) and a putative polysaccharide pyruvyltransferase (L143). All the genes inside the cluster are expressed starting from the intermediate phase of virus infection and last until the late phase, suggesting that they are related with *Mimivirus* virion maturation and fibrils formation.

R141 and L136 are responsible for the UDP-D-Viosamine production (Fig. 2.3).

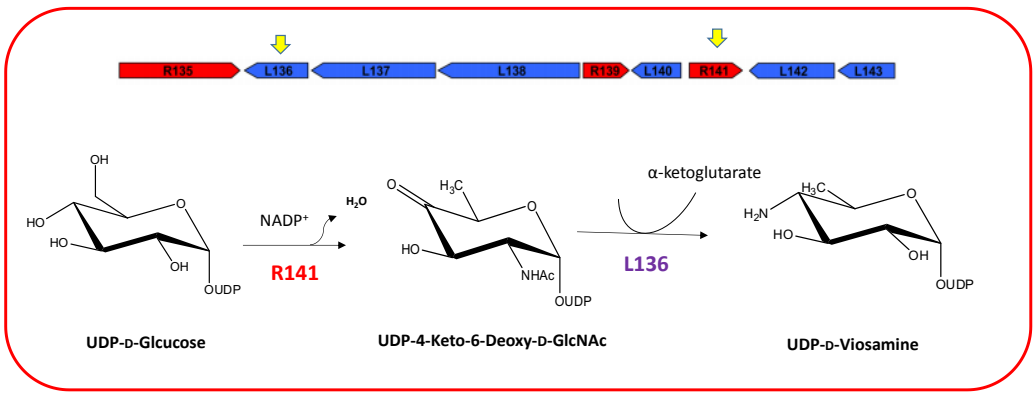


Fig. 2.3 Upper part: the nine gene cluster involved in glycan formation (reproduced from Piacente *et al*., 2012); down: UDP-D-Viosamine pathway in *Mimivirus*. *R141* (in red has an eukaryotic origin), and *L136* (in purple).

The protein encoded from *R141* transforms UDP-Glc in UDP-4-keto-6-deoxy-D-Glucose, which is the substrate of L780 (outside of the nine genes cluster) and L136. It has been proved that L136 is a pyridoxal phosphate-dependent sugar aminotransferase, which catalyzes the transfer of an amino group from the glutamate to the carbon four of the UDP-4-keto-6-deoxy-D-Glcucose resulting in the production of UDP-D-Viosamine (4-amino-4,6-dideoxy-D-glucose) and α-ketoglutarate. The reaction of transamination of the UDP-4-keto-6-deoxy-D-Glucose made by L136 is irreversible. The D-viosamine, in contrast with L-rhamnose, is a very rare sugar, which is found only in some bacteria.

Accordingly, the gene L136 has a bacterial origin, and a phylogenetic analysis revealed that it is not the result of a recent horizontal gene transfer from bacteria, but that it was acquired very early during the evolution. Thus, Mimivirus uses two genes with different origins (Eukaryotic and prokaryotic) to produce the D-Viosamine, opening the question about the origin of this pathway.

2.1.3 The *Mimivirus* biosynthesis of UDP-D-N-acetylglucosamine

A preliminary analysis of the sugar content of *Mimivirus* fibrils revealed that the major saccharides component are rhamnose, viosamine and glucosamine (Piacente et al., 2012). This result has been supported the in vitro data about the presence of a functional pathway for UDP-L-Rhamnose and UDP-D-Viosamine. In addition, this analysis showed that N-acetylglucosamine is the most abundant component of the fibrils. Accordingly, the *Mimivirus* genome was scrutinized to verify if *Mimivirus* has also the genes to produce the N-acetylglucosamine. *L619*, *L316* and *R689* were identified, cloned and expressed, and altogether demonstrated that *Mimivirus* has a functional pathway for the N-acetylglucosamine production (Piacente et al., 2014a) (Fig.2.4).

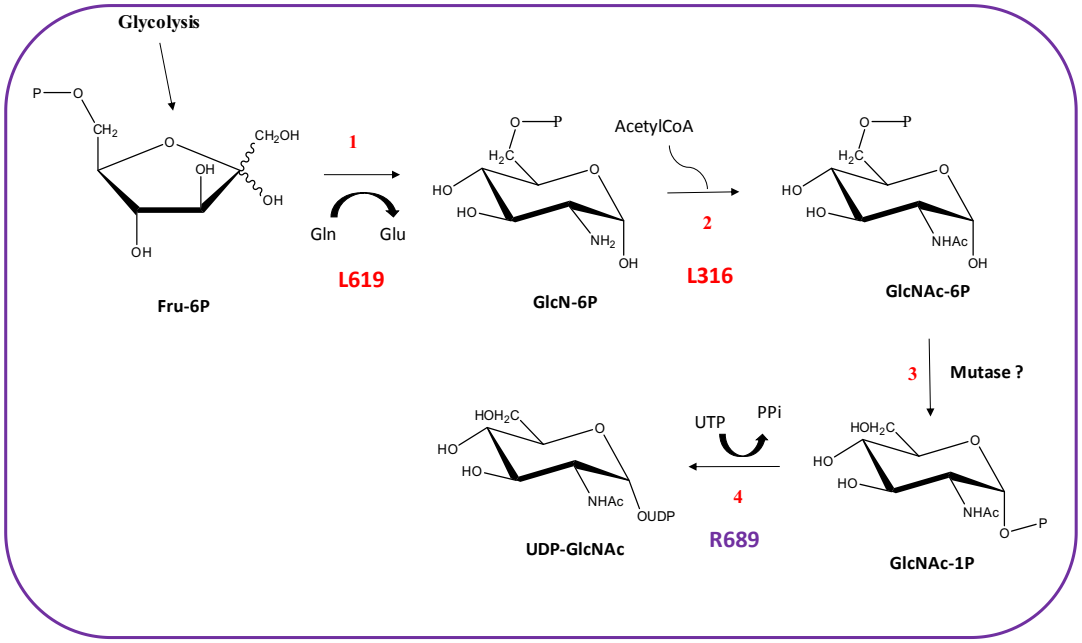


Fig. 2.4 *Mimivirus*' UDP-D-N-acetylglucosamine pathway. *Mimivirus* adopts the eukaryotic strategy for the biosynthesis of the UDP-D-N-acetylglucosamine. The first two genes (in red) have an eukaryotic origin, while the last one (in purple) has a prokaryotic origin.

L619 is a glutamine-fructose-6-phosphate transaminase that converts the fructose-6-P in glucosamine-6-P, using the glutamine as ammonia donor. Then L316, a glucosamine-6-phosphate N-acetyltransferase, adds an acetyl group from the Acetyl-coA to the amino function, leading to N-acetylglucosamine-6-P. This substrate is converted in N-acetylglucosamine-1-P from a mutase. The gene encoding for the mutase was not identified, due to the absence of homologs with characterized mutases in the cellular world. The last step, made by R689, is the addition of an uridyl moiety to the phosphorylated sugar, generating UDP-D-N-acetylglucosamine. This pathway follows the eukaryotic strategy; in addition, the first two enzymes are related to their eukaryotic counterpart but the last one resembles bacterial proteins. Once again and as found for UDP-Vio, the genes encoding for the UDP-D-N-acetylglucosamine have different origins and the phylogenetic analysis suggested that this pathway was acquired early during the evolution.

D-N-acetylglucosamine is a very common sugar, present in all domains of the life and that plays an important role in several processes. The identification of this pathway in *Mimivirus* was surprising, due to the presence of this pathway in its host. Therefore, the production of UDP-D-N-acetylglucosamine seems a strategy for *Mimivirus* to decrease the dependency from the host machinery and to have its own supply of this nucleotide sugar.

2.1.4 The *Megavirus chilensis* UDP-L-Rhamnosamine pathway

The glycans related genes are not confined to *Mimivirus* but are present in all the *Mimiviridae* family. A functional pathway for another uncommon acetamido sugar, UDP-L-Rhamnosamine, was identified in *Megavirus chilensis* (Piacente et al., 2014b). Similar to *Mimivirus* UDP-D-Viosamine production, the genes involved in the UDP-L-Rhamnosamine are organized in cluster. The first gene of the cluster is *mg534*, which encodes for an inverting 4,6-dehydratase that catalyzes the removal of a water molecule from the carbon 4 and carbon 6 and the epimerization of the carbon 5 of the UDP-D-GlcNAc, leading to UDP-4-keto-6-deoxy-N-acetyl-L-idosamine. This compound is then epimerized at carbon 3 and reduced at carbon 4 by mg535 with the formation of UDP-L-RhaNAc (Fig. 2.5). In addition, the presence of this rare sugar was confirmed from a preliminary sugar analysis of *Megavirus* particles.

Inside the cluster, the gene *mg536* has a high similarity with *V. cholera O37* WbvD (50%) and *P. aeruginosae* WbjD (47%), strongly suggesting that it encodes for an UDP-D-GlcNAc 2-epimerase, thus generating UDP-L-Quinovosamine. Despite, only the activity of the first two genes has been demonstrated in vitro and the sugar analysis in vivo has not confirmed yet the presence of L-quinovosamine. Therefore, further investigation was undertaken demonstrate the presence of L-RhaNAc and L-QuiNAc in *Megavirus* fibrils.

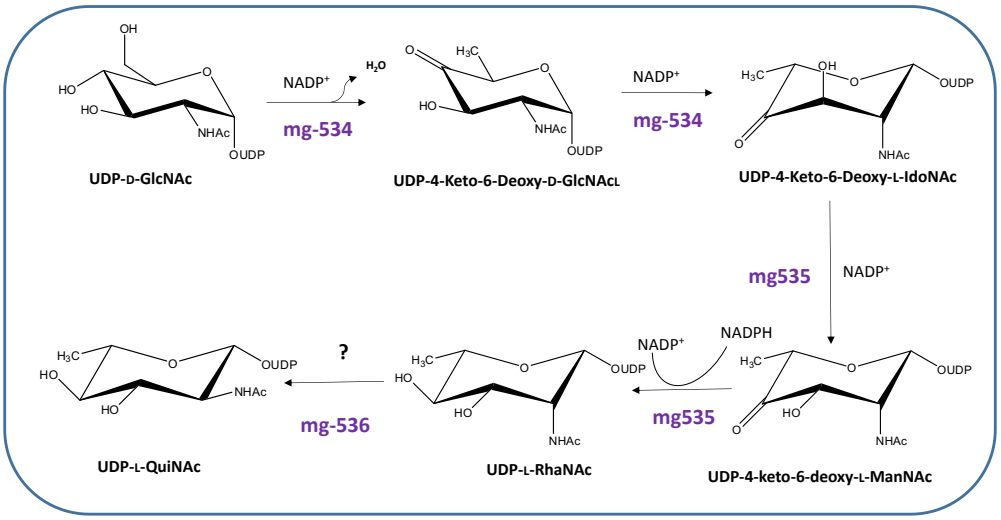


Fig. 2.5 UDP-L-Quinovosamine pathway in *Megavirus chilensis*. All the genes involved in this pathway have a prokaryotic origin. This pathway is reported in Piacente *et al*., 2015.

2.1.5 Glycosyltransferases

In addition to the genes involved in nucleotide-sugar metabolism, several genes (up to seven) encoding for glycosyltransferase have been identified in the *Mimiviridae* family. Until now, L230 represents the only viral glycosyltransferase for which the activity has been demonstrated in vitro. This enzyme catalyzes the transfer of a glucose unit to the hydroxylysine of *Mimivirus* collagen-like proteins (Luther et al., 2011).

Most interesting is the presence of genes for putative glycosyltransferases inside the cluster for the nucleotide sugar production. The only way to elucidate the function of each glycosyltransferase is the knowledge of the structures of the glycans associated to the fibrils of *Mimiviruses* so that a reasonable guess about the nature of the donor and of the acceptor can be done. Accordingly, the investigation of the glycans structures of some members of the *Mimiviridae* family is the object of this thesis.

Thesis objective

The understanding of the innovative glycosylation system of the giant DNA viruses

The giant DNA viruses, in contrast with the "traditional viruses", encode, if not all, the machinery required for their own glycosylation. The first evidences of glycan-related genes, come from two members of the *Megavirinae* (Gallot-Lavallée et al., 2017), *Mimivirus* (lineage A)(Scola, 2003) and *Megavirus chilensis* (Arslan et al., 2011) (lineage C), which have several genes encoding enzyme for sugars-nucleotide metabolism, sugar-modifying enzyme and glycosyltransferase. It has been demonstrated that the glycans were attached to the fibrils surrounding their capsids (Piacente et al., 2012), as suggested from the Gram-positive staining of the viral particles, indicating that the fibrils were probably composed by peptidoglycan-like structures. The glycosylation of the fibrils was confirmed from the analysis of their sugar content, revealing that the major saccharide components were Rha, GlcNAc, and Vio for *Mimivirus* (Parakkottil Chothi et al., 2010; Piacente et al., 2012, 2014b) and GlcNAc and RhaNAc (Piacente et al., 2014a) for *Megavirus chilensis*. These data have excluded the presence of a true peptidoglycan. Until now, the information about the sugar composition of the fibrils sugar of a member of the lineage B, for which the prototype is *Moumouvirus* (La Scola et al., 2010), is lacking.

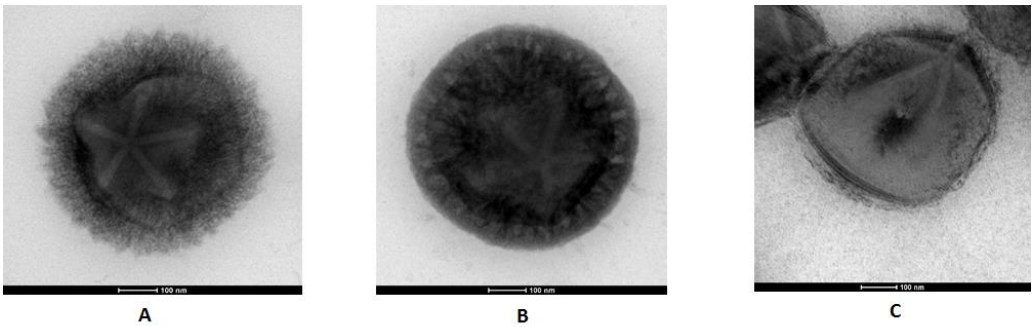
It has been hypothesized that the differences in sugar composition could be connected to the different length of the fibrils, which are longer in *Mimivirus* than in *Megavirus* and *Moumouvirus*.

Many questions are still opened:

- ✓ Are the fibrils glycosylated by poly-oligosaccharide or by monosaccharides?
- ✓ Is the fibrils sugar composition of the linage B similar to that of the other two lineages? Or is it completely different?
- ✓ Is the fibrils sugar composition conserved in each lineage?

- ✓ Are the giant DNA viruses completely autonomous for the fibrils glycosylation?
- ✓ Which are the glycosylated proteins composing the fibrils?
- ✓ What is the role of the glycosylation in the fibrils?

The target of this thesis is the understanding of the innovative glycosylation system of the *Mimiviridae* family, using *Mimivirus* (Lineage A), *Megavirus chilensis* (Lineage C) and *Moumouvirus australensis* (lineage B) as prototypes (Fig. A, B and C).



Electron microscopy of **A)** *Mimivirus*; **B)** *Megavirus chilensis*; and **C)** Mouomouvirus australensis particles. *Mimivirus'* fibrils are the longest compared to those of *Megavirus chilensis* and *Moumouvirus australensis*.

To address these questions a combination of three different methodologies was adopted:

1) **Carbohydrate chemistry**: to elucidate the structure of the glycans associate to the fibrils of the giant DNA viruses (Chapter 3);
2) **Bioinformatic strategies:** to discover new genes involved in the fibrils glycosylation, starting from the experimental data (Chapter 4);
3) **Biochemical strategies:** to validate in vitro the function of the new glycan-related genes (Chapter 5).

The integration of these different strategies is necessary to shed light on the glycosylation system of the giant DNA viruses, which will enable to:

1) Make progresses in the field of glycobiology;

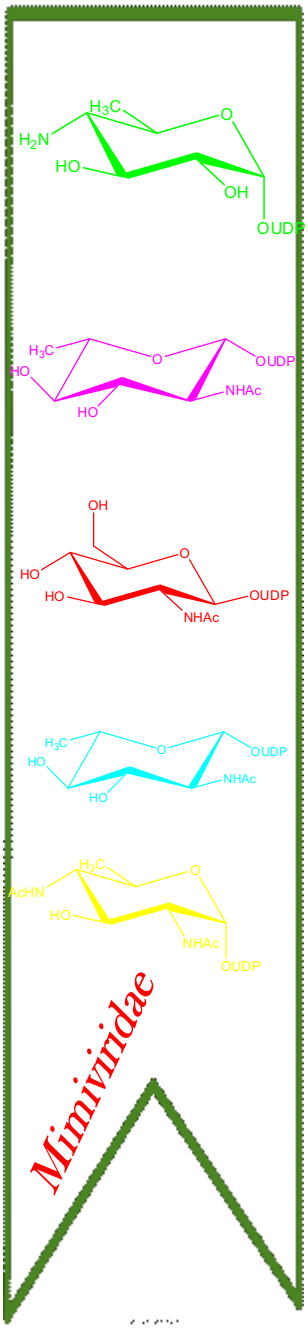
2) Elaborate new methods for the purification of carbohydrates and glycoproteins;

3) Increase the number of genes of known function in giant viruses;

4) Extend these approaches to other giant viruses such as the *Pandoraviruses*, *Pithoviruses*, *Mollivirus*, a list in constant growth;

5) Expand the current knowledge about environmental viruses;

6) Correlate the viral glycosylation machinery to that of other organisms.



Section II

Results and discussion

Chapter 3

Characterization of the glycans associated to the fibrils of *Mimivirus*, *Megavirus chilensis* and *Moumouvirus australensis.*

The carbohydrate chemical methodologies were applied to elucidate the glycans structures associated with the fibrils of these giant DNA viruses. The experimental data have been used as a starting point, for the identification of new glycan related gens (Chapter 4), to be later validated for their function in vitro (Chapter 5). In addition, the understanding of the fibrils glycans structure, at the later point, could shed light on the host-interaction process, clarifying the role of the fibrils in the giant DNA viruses' biology.

3.1 The fibrils surrounding the giant DNA virus Mimivirus capsid are heavily glycosylated

3.1.1 Production and purification of *Mimivirus* virions

A combination of in vitro and in vivo data has disclosed the nature of the sugars associate to Mimivirus fibrils as rhamnose, N-acetylglucosamine and viosamine (Parakkottil Chothi et al., 2010; Piacente et al., 2012, 2014b) (Chapter 2). These data have prompted the question about the fibrils glycans structures. Are Mimivirus fibrils decorated by single monosaccharides or by oligo-polysaccharides? To complete these data, an in vivo analysis was carried out on the sugars associated to Mimivirus particles. Accordingly, a huge number of viruses (about 3x 10^{12} viral particles) was produced and purified (Fig. 3.1).

Production and purification of *Mimivirus* virions

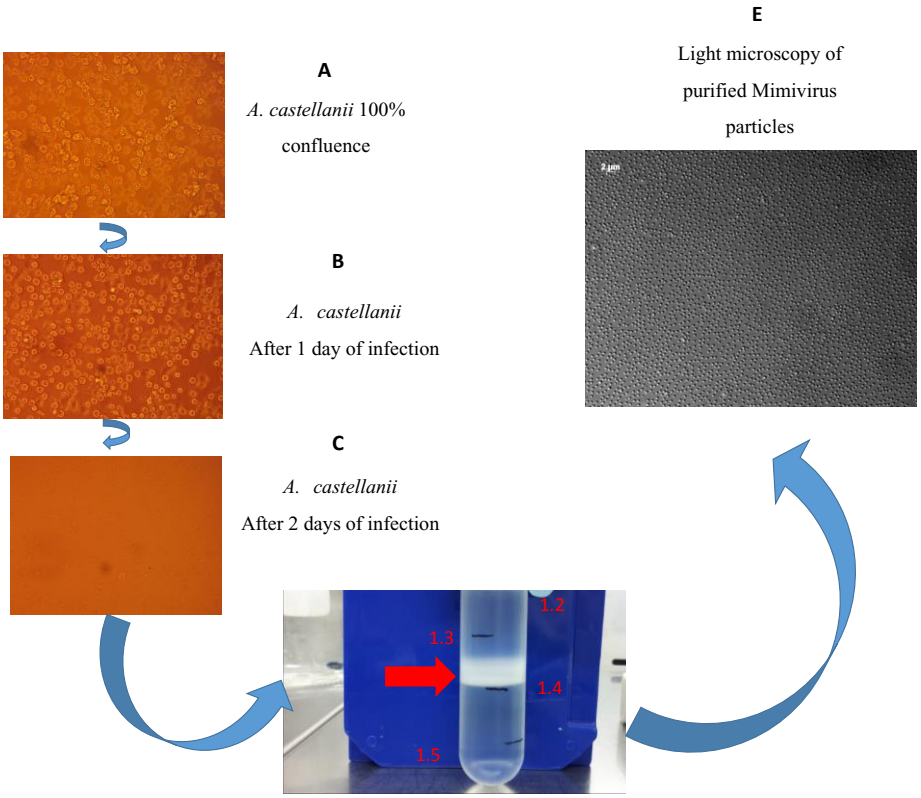


Fig. 3.1 M*imivirus* production and purification. Light microscopy observation of. **A)** A. castellanii cells at 100% of confluence, prior infection. **B)** A. castellanii cells after one day of infection with Mimivirus. C) A. castellanii cells completely lysed. **D)** Mimivirus particles purified on CsCl gradient, the viral particles appeared as a withe disk, indicated from the red arrow. **E)** Light microscopy observation of purified Mimivirus particles; the scale bar is 2 µm.

Mimivirus was cultivated in the laboratory in *A. castellanii* (Douglas) Page (ATCC® 30010). The amoeba cells were cultivated until 100% of confluence (Fig. 3.1 A) and infected with *Mimivirus*. The infection was followed by light microscopy

observation. One-day post infection, most of the cells presented the infected phenotype, indeed the cells were around (Fig. 3.1 B). Two days post infection, the majority of amoebae cells were completely lysed from the virus. (Fig. 3.1 C). Then *Mimivirus* mature particles were separated from the cellular debris and purified on a gradient of CsCl. Mimivirus particles appeared as a white disk of a density ranging from 1.4 to 1.3 (Fig. 3.1 D). The viral particles were recovered and washed with pure water several times. The possible presence of bacteria was excluded by checking the purified viral particles by light microscopy (Fig. 3.1 E), where appeared as spherical particles.

3.1.2 Isolation of the *Mimivirus* fibrils

To isolate the glycans attached to the fibrils, a protocol to detach the fibrils from the viral particle was set up. Until now is not clear how the fibrils are connected to the viral capsid and the protein making the fibrils are unknown, getting more difficult the choice of the treatment to adopt. The *Mimivirus* viral particles were treated with dithiothreitol (DTT), which is one of the strongest reducing agents, in two different conditions (Fig.3.2):

- ✓ DTT 50 mM 2h R.T.
- ✓ DTT 50 mM 2h 100°C

The efficiency of each condition was evaluated by transmission electron microscopy (TEM) and by nuclear magnetic resonance spectroscopy (NMR).
TEM observations disclosed that both conditions were suitable for removing the fibrils from the viral capsid, but with different efficiency (Fig. 3.2).

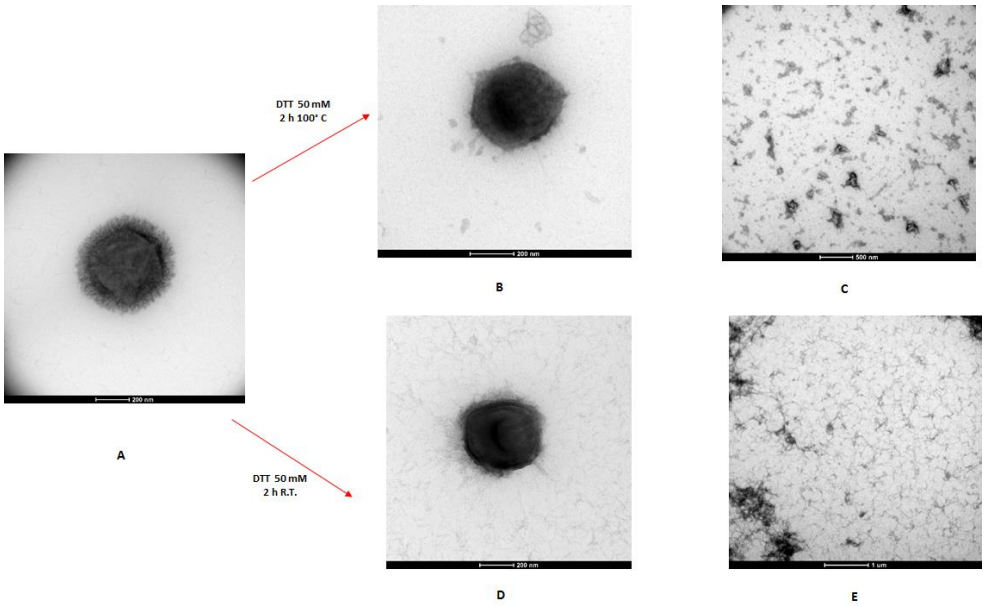


Fig. 3.2 Transmission electron microscopy of *Mimivirus*: **A)** virus untreated as control; **B)** virus particle without fibrils, after treatment DTT 50 mM, 2 h ,100°C; **C)** fibrils detached after treatment DTT 50 mM, 2 h, 100°C, they appeared aggregate probably due to the heating; **D)** virus particle after DTT 50 mM, 2 h, R.T. there are still some fibrils attached to the viral capsid; **E)** fibrils detached after treatment DTT 50 mM, 2 h, R.T, they appeared similar to those connected to the viral particle untreated.

Under the strongest conditions the major part of the viral particles was devoid of the fibrils (Figs. 3.2 A, B, C), although some of them were damaged.

On the contrary, TEM analysis of the defibered viruses in milder conditions disclosed that the removal of the fibrils was non-complete but that there was a minor number of damaged viruses (Figs. 3.2 A, D, E).

The fibrils obtained from the different treatments appeared quite dissimilar by TEM observations. The fibrils isolated in strongest conditions appeared as aggregated, probably due to the heating, while those obtained in mild conditions resembled to those connected to the viral particle untreated.

The fibrils underwent mono-dimensional DOSY analysis, in order to evaluate the presence of the sugars and to understand if the different treatments had an implication

on the glycans isolation. This kind of spectrum was preferred to the sequence with solvent presaturation because it did not introduce any distortion in the signals around the residual water signal. A mono-dimensional DOSY on the fibrils differently obtained (Fig. 3.2), displayed a complex pattern of anomeric signals (^{1}H range 5.1-4.5 ppm), a crowded carbinolic area (^{1}H range 4.4-3.1 ppm), two main acetyl signals (~2.0 ppm), a methyl group of the pyruvic acid (1.5 ppm) and a group of methyl signals characteristic of 6-deoxysugars at 1.3 ppm (Fig. 3.3). There were non-significant differences between the two samples, implicating that the two treatments were comparable for the glycans isolation.

Therefore, further analyses were performed on the fibrils from the treatment at 100°C because obtained with a better yield (10 mg against 5 mg).

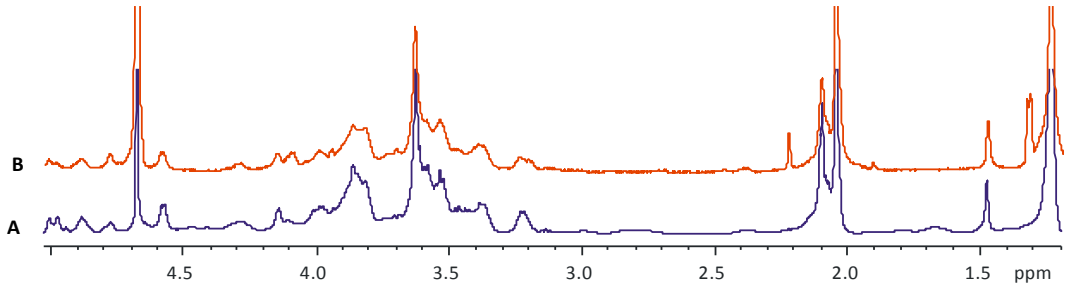


Fig. 3.3 (600 MHz, 310 K) Mono-dimensional DOSY spectra of isolated fibrils glycans in different conditions: **A)** DTT 50 mM, 2h, 100°C; **B)** DTT 50 mM, 2h, R.T. There is a complete overlap between the fibrils glycans isolated in strong and mild conditions.

3.1.3 Characterization of Mimivirus glycans structure

The sugars associated to the fibrils were analyzed by two different strategies:

- ✓ Chemical analysis: Sugar composition and linkage analysis;
- ✓ NMR analysis: ^{1}H NMR, COSY, TOCSY, NOESY, HSQC, HMBC

Sugar composition of the fibrils

The chemical characterization of the Mimivirus fibrils was carried out by transforming 0.5 mg of the sample in the corresponding acetylated methyl glycosides (AMG), sufficiently volatile species that can be analyzed by gas chromatography-mass spectrometry (GC-MS). Each acetylated methyl glycoside is eluted in a range of the chromatogram and at each peak corresponds a specific fragmentation pattern, which enables the identification of the monosaccharide (Lönngren and Svensson, 1974). The elution time compared with those of standards monosaccharides permits to distinguish between sugars of the same class, for instance between glucose and mannose.

This analysis disclosed that the major fibrils components were rhamnose (Rha), 2O-methyl-viosamine (2OMe-Vio), along with non-methylated viosamine (Vio), N-acetylglucosamine (GlcNAc) (Fig. 3.4), and mannose (Man) and glucose (Glc) as minor components.

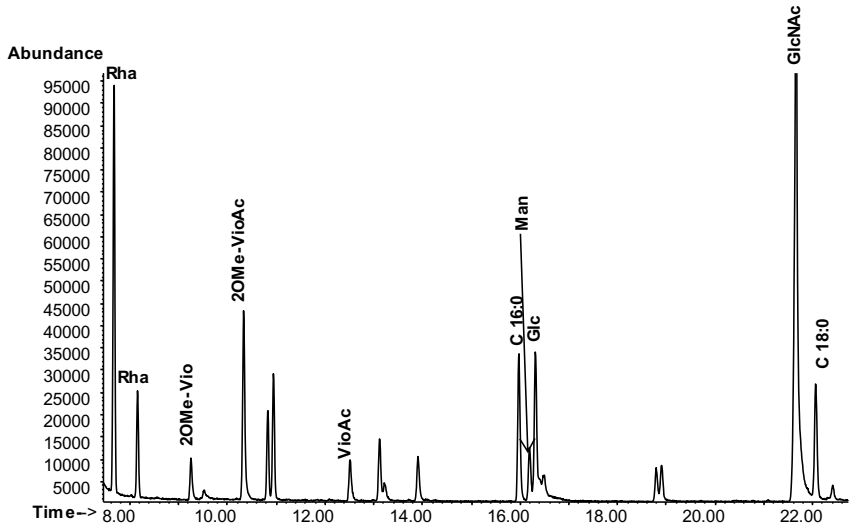


Fig. 3.4 GC-MS analysis of acetylated methyl glycosides (AMG) of fibrils glycans. This analysis revealed that the major saccharides components are Rhamnose (Rha), Viosamine (Vio), Viosamine methylated in C2 (2OMe-Vio) and Glucosamine (GlcNAc). As minor components: Mannose (Man), Glucose (Glc). In addition, there were traces of C16:0 and C18:0.

Identification of the peak at 10 min as 2-OMe-Vio was possible by applying the fragmentation rules, indeed, the EI-MS spectrum contained a fragment at m/z 244 consistent with the oxonium ion of a 6-deoxy-aminosugar substituted with a methyl group. In addition, the fragment at m/z 88 (Fig. 3.5) was diagnostic of two vicinal methoxyl groups, indicating that the anomeric methoxyl group was flanked by another at position two.

These results agreed with the presence of a functional pathway for the UDP-L-Rha, UDP-D-GlcNAc and UDP-D-Vio (as discussed in Chapter 2).

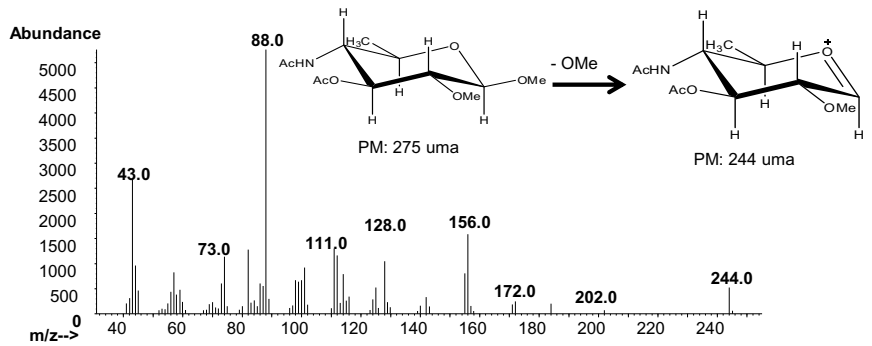


Fig. 3.5 EI-MS spectrum of the 2OMe-Vio (RT 10 min). The fragment at m/z 244 consistent with the oxonium ion of a 6-deoxy-aminosugar substituted with a methyl group. In addition, the fragment at m/z 88 was diagnostic of two vicinal methoxyl groups, indicating that the anomeric methoxyl group was flanked by another at position two.

Sugar linkage analysis of the fibrils

GC-MS analysis of partially methylated alditol acetates (PMAA) led to the identification of monosaccharides branching points. The chromatogram was

interpreted based on the fragmentation pattern (Ciucanu and Kerek, 1984) , relative to each signal, and retention times, compared with appropriate standards.

This analysis revealed the presence of 3-Rha (3-linked rhamnose), 2,3-Rha, t-VioAc, 3-GlcNAc and 3,4,6-GlcNAc (Fig. 3.6). This type of analysis cannot discriminate between Vio and 2OMe-Vio, because both appear as terminal residues. In addition, minor peaks, corresponding to less representative sugars, were not identified.

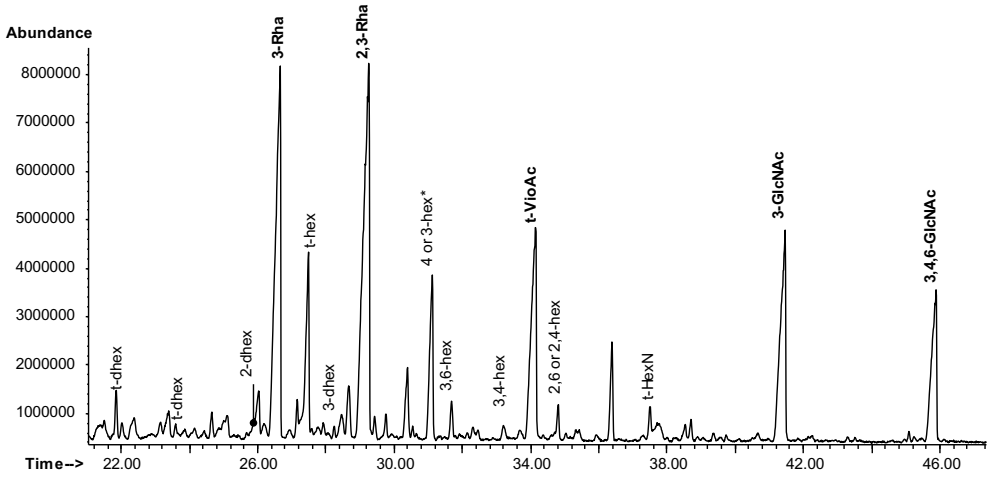


Fig. 3.6 GC-MS analysis of partially methylated alditol acetates (PMAA) of fibrils glycans. In bold there are the main sugar of the fibrils: Rha linked in 3, Rha linked in 2 and 3, Vio as terminal sugar, GlcN linked in 3 and GlcN fully substituted. In italic there are the less representative sugars.

NMR analysis of the fibrils

To elucidate the structure of the glycans, the complete NMR analysis (Table 1) was carried out on the fibrils by applying 1D and 2D homo- and heteronuclear sequences.

The ^{1}H NMR spectrum (Fig. 3.7 A) had five main anomeric signals (5.1-4.6 ppm) labeled with letters A-E, a crowded carbinolic area (4.4-3.2 ppm), and several signals of methyl groups in the high field region, due to N-acetyl groups (2.09-2.04ppm), to a pyruvate substituent (1.5 ppm) and to the six position of 6-deoxy-sugars (1.3 ppm).

Residue **A** has been attributed to α-Rha. Indeed, COSY spectrum contained a weak H-1/H-2 correlation while all the others were intense, with H-6 found at 1.24 ppm (Table 1) and diagnostic of a 6-deoxysugar. This information agreed with the pattern found in the TOCSY spectrum which showed all the correlations starting from H-2, and not from H-1, as found for monosaccharides with the *manno* stereochemistry, rhamnose in this case. Integration with HSQC spectrum information (Fig. 3.7 D), resulted that C-2 and C-3 (78.9 and 82.2 ppm, respectively) were shifted at low field compared to the reference values because glycosylated(Bock and Pedersen, 1983) .Accordingly, NOESY spectrum (Fig. 3.7 C) correlated H-1 of **A** with H-3 of **B** and H-1 of **E** with H-3 of **A**. Finally, the configuration at the anomeric carbon was α as inferred by the diagnostic carbon chemical value of C-5 (70.6 ppm). The anomeric proton of **C** almost overlapped with that of **D** (Fig. 3.7 B), and the TOCSY spectrum from H-2 position (3.94 ppm) indicated the positions of the other protons of the residue. Analysis of the carbon chemical shifts values indicated that this rhamnose unit was α-configured and linked at position 3, as proved by the C-3 value (81.8 ppm) and supported by the NOE effect among its H-3 and H-1 of **D**. The residue **B** and **D** were identified respectively as 3-β-GlcNAc and 3,4,6-β-GlcNAc.

Residue	Nucleus	1	2	3	4	5	6
A	^{1}H	5.01	4.13	3.85	3.41	3.97	1.24
2,3-α-Rha	^{13}C	101.0	78.9	82.2	72.0	70.6	18.4
B	^{1}H	4.87	3.82	3.6	3.59	3.36	3.87/3.89
3-β-GlcAc	^{13}C	102.5	56.6	82.6	69.4	76.9	61.7
C	^{1}H	4.79	3.94	3.87	3.42	4.29	1.24
3-α-Rha	^{13}C	101.9	71.4	81.8	72.1	69.8	18.4
D	^{1}H	4.78	3.90	3.75	3.60	3.47	3.71/4.10
3,4,6-β-GlcNAc	^{13}C	104.1	57.0	79.0	75.8	67.0	65.86
E	^{1}H	4.58	3.22	3.55	3.63	3.55	1.23
β-2OMeVioNAc	^{13}C	105.4	84.3	74.0	57.7	71.9	18.4
Pyruvic acid	^{1}H	-	-	1.47			
	^{13}C	175.6	102.7	26.2			

Table 1. (600 MHz, 329 K, D_2O) ^{1}H and ^{13}C chemical shifts of the two polysaccharides associated to *Mimivirus* fibrils. The two Glucosamine (residues B and D) are acetylated to the amino function (^{1}H 2.09, ^{13}C 23.8). The Viosamine is acetylated to the four amino function (1H 2.07, 13C 23.5) and methylated at carbon 2 (^{1}H 3.63, ^{13}C 60.8).

The *gluco* stereochemistry of **B** and **D** determined the efficient magnetization transfer in the TOCSY spectrum. Accordingly, the **B** and **D** anomeric signals displayed correlations up to H-6s and integrating this information with those from COSY and HSQC spectra, the proton and carbon chemical shift of **B** and **D** were determined (Table 1).

Indeed, H-2/C-2 values of **B** and **D** (3.82/56.6 and 3.91/57.1 ppm, respectively) underlined that these units were N-acetylated, each anomeric carbon was β configured as indicated by characteristic H-1/H-3 and H-1/H-5 correlations in the NOESY spectrum. For residue **B**, the low-field displacement of its C-3 signal (82.6 ppm, Fig. 3.7 D) indicated glycosylation at this position, in agreement with the correlation between H-3 and H-1 of **A** in the NOESY spectrum (Fig. 3.7 C).

As for **D**, the chemical shifts of C-3, C-4 and C-6 (79.0, 75.8 and 65.9 ppm, respectively), indicated that this residue was fully substituted. The NOE spectrum revealed that H-1 of **C** correlated with H-3 of **D**. Pyruvic acid was linked at C-4 and C-6 as acetal, with the S configuration (Fig. 3.8) based on the characteristic $^{1}H/^{13}C$ chemical shifts (1.47/26.0 ppm) reported in literature (Garegg et al., 1980; Toukach, 2011).

Residue **E** was identified as a not substituted β-2OMe-VioNAc. Combination of COSY and TOCSY spectra, established the sequence of the different protons, while HSQC identified the corresponding carbon chemical shifts. H-2 (3.22 ppm) correlated with a carbon at low field (84.3 ppm) because methylated at the corresponding hydroxyl function (O-CH3 at 1H/13C 3.63/60.9 ppm), as confirmed by the $\mathbf{E}_{2,2OMe}$ and $\mathbf{E}_{2OMe,2}$ cross peaks in the HMBC spectrum (Fig. 2.3.7 E). H-3 and H-5 were found almost coincident, as suggested from TOCSY and NOESY spectrum (Fig. 3.7 B) which had one intense cross peak at 3.55 ppm embracing both H-3 and H-5. H-4 and H-6 were found respectively at 3.64 and 1.24.

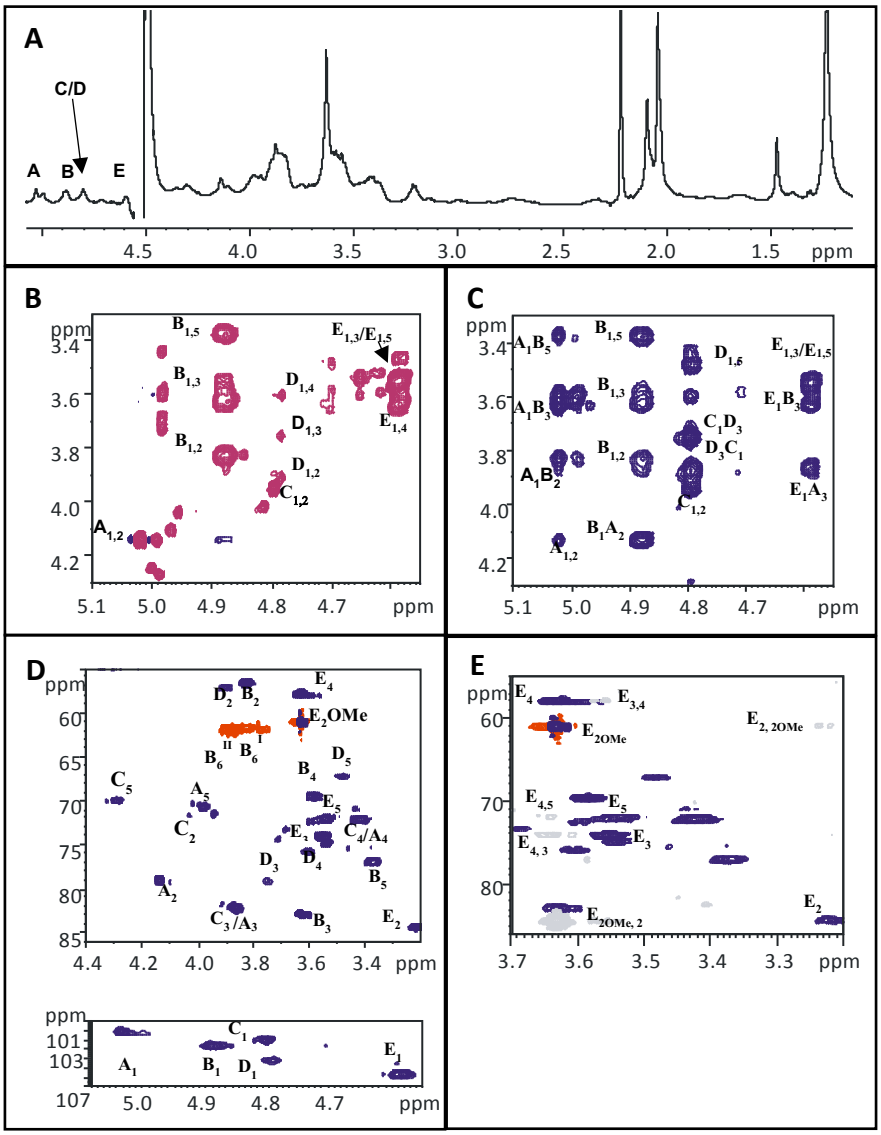


Fig. 3.7 NMR analysis of the glycans fibrils (600 MHz, 329 K, D2O). **A)** Expansion of proton spectrum; The ^{1}H NMR spectrum shows five main anomeric signals (5.1-4.6 ppm) labeled with letters A-E, a crowded carbinolic area (4.3-3.2 ppm), and several signals of methyl groups in the high field region, due to N-acetyl groups (2.09-2.04ppm), to a pyruvate substituent (1.5 ppm) and to the six position of 6-deoxy-sugars (1.3 ppm). Letters used for densities attribution follows the system of Table 1; **B)** Expansion of TOCSY spectrum; **C)** Expansion of NOESY spectrum; **D)** Expansion of HSQC spectrum: upper carbinolic region; lower anomeric region; **E)** Expansion of HMBC spectrum.

The values of H-4/C4 (3.64 and 57 ppm) were diagnostic of an acetylated amino function. This information completed the evidence of our in vitro studies (Chapter 4, publication attached). The residue E is linked at O-3 of A as inferred from the diagnostic correlation in the NOESY spectrum (Fig. 3.7 C).

Taken all the spectroscopic information together, it was possible to define the structure of two glycans: polysaccharide 1 (or poly_1, Fig. 3.8 A) has a linear repeating unit made of 3)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→, with GlcNAc further having at position 4,6 a pyruvic acid. Polysaccharide 2 instead (poly_2, Fig. 3.8 B) has a branched repeating unit, with 2)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→ in the linear backbone, with Rha is further branched at O-3 with a terminal β-D-2OMe-VioNAc.

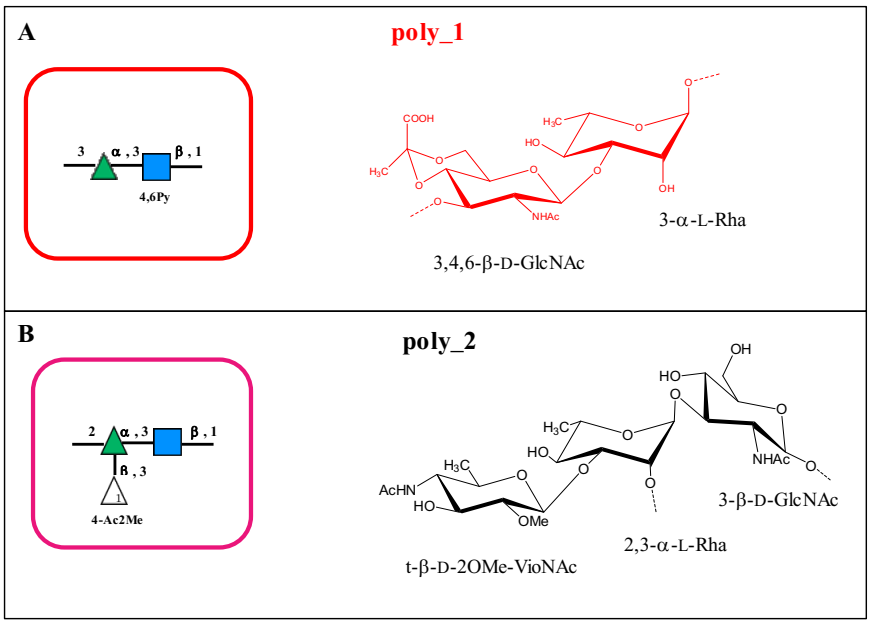


Fig. 3.8 Structure of the two polysaccharides. **A)** The polysaccharide 1 has a repeating unit made of 3-α-L-Rha and 3,4,6-β-D-GlcNAc; the GlcNAc links in 4 and 6 a pyruvic acid. In the red square there is the symbolic representation of the Poly-1. **B)** The polysaccharide 2 has a branched repeating unit on a linear backbone made of 3-β-D-GlcNAc and 2-α-L-Rha; Rha is further branched at O-3 with β-D-2OMe-VioNAc, which is the terminal sugar. In the pink square is reported the symbolic representation of the Poly_2. Monosaccharides symbols follow the SNFG (Symbol Nomenclature for Glycans) system.

However, the finding of two polysaccharides repeating units raised the question about how these are organized. Is there a single polymer with blocks each made by one repeating unit (Fig. 3.9 A) or are they assembled as separate polysaccharides? In the latter case, are each polysaccharide linked to the same protein at different positions (Fig. 3.9 C) or branched to two different proteins (Fig. 3.9 B)?

Answer to these questions was inferred by purifying the fibrils by two main approaches, anion exchange chromatography and Reverse Phase-HPLC.

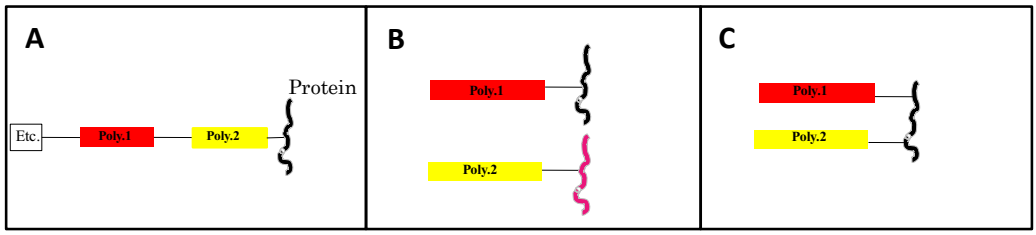


Fig. 3.9 Possible polysaccharides organizations. **A)** One single filament in a block-wise fashion; **B)** Each polysaccharide is linked to different proteins; **C)** Distinct polysaccharide branched on a same protein.

3.1.4 Separation of *Mimivirus* polysaccharides by anion exchange chromatography

The strategy adopted to separate the two polysaccharides was based on their different chemical properties, as the poly_1 is an acid polysaccharide for the presence of the pyruvic acid and the poly_2 is a neutral polysaccharide. The anion exchange chromatography was applied on the fibrils either intact or after enzymatic digestion with proteinase K.

In general, polysaccharide material was eluted at NaCl concentration between 10 and 400 mM and as expected from the chemical properties of the polysaccharides, the Poly_2 has always been eluted at low concentration of NaCl, because it was hardly retained by the chromatographic column, while the Poly_1 has been eluted at higher

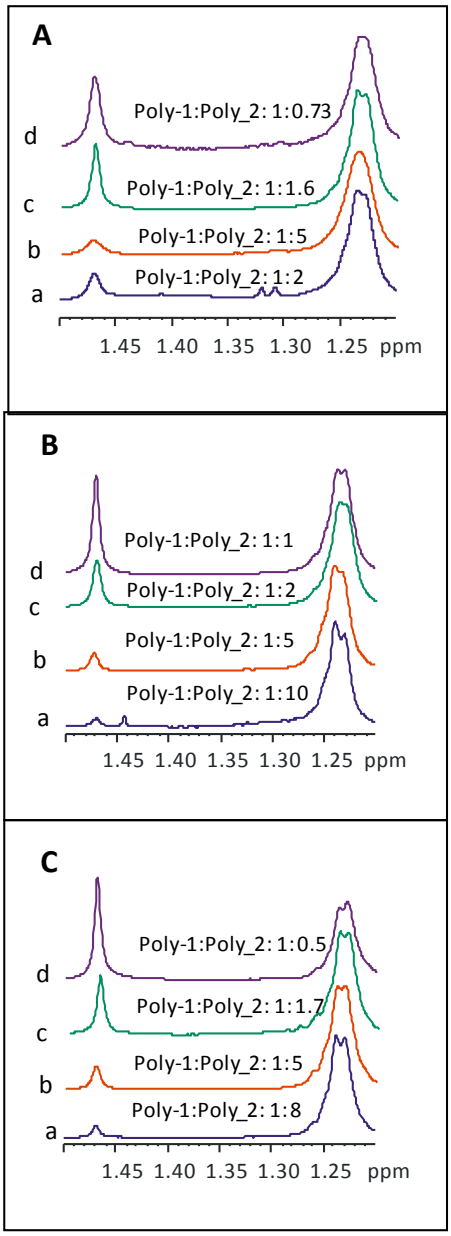


Fig. 3.10 (600 Mz, 310 K) Expansion of mono-dimensional DOSY spectra of the anion exchange fractions in different condition: untreated fibrils (A), fibrils digested with protease K (B) and double digested fibrils (A). For all conditions a, b, c, d represent the elution at these concentration of NaCl: 10 mM, 100 mM, 200 mM and 400 mM.

concentration. The proportion between the two species was determined by analyzing the mono-dimensional DOSY spectrum of each fraction. The methyl signal at 1.47 ppm, arising from the pyruvate group, was used as reporter group of poly_1, while the signal at 1.24 ppm was used to determine poly_2 proportion after that the contribute from poly_1 was deducted.

The anion exchange chromatography of the untreated fibrils produced polysaccharide 2 in pure form (fraction 10 and 100 mM NaCl,), while the polysaccharide 1 was never obtained alone, although increased by increasing the anionic strength (Table 2, 3, Fig. 3.10 A).

With the intent to improve polysaccharide separation, the fibrils were digested with proteinase K and purified by ion exchange chromatography. Again, the poly_2 content was predominant in the fractions at low ionic strength but higher than that observed for the same experiment with the untreated fibrils. For instance, poly_2 was in tenfold excess with respect to poly_1 in the 10 mM while in the untreated fibrils its amount only doubled that of poly_1 (Tables 2, 3, Fig. 3.10 B).

Thus, poly_2 was almost pure in 10 and 100 mM NaCl fractions, while the proportion of poly_1 steadily increased with the increase of NaCl concentration of the eluent.

To further improve the separation between the two species, fibrils were digested twice with proteinase K and separated by anion exchange chromatography. NMR analysis of the resulting fractions (Tables 2, 3, Fig. 3.10) disclosed a slight enrichment of poly_1 in the fraction eluted with 400 mM NaCl. In addition, the amount of poly_2 recovered in the 10 mM NaCl fraction (0.28 mg) increased compared to that obtained with only one protease K digestion (0.13 mg), although it had an eight- (and not ten-) fold excess respect poly_1.

$[NaCl]_{mM}$	Untreated fibrils	Proteinase K digestion	Double proteinase K digestion
10	1:2	1:10	1:8
100	1:5	1:5	1:5
200	1:1.6	1:2	1:2
400	1:0.7	1:1	1:0.5

Table 2: poly_1: poly_2 ratio as determined by methyl signal integration in the DOSY spectrum of the fractions obtained by ion exchange chromatography of the fibrils with or without proteinase K treatment.

$[NaCl]_{mM}$	Untreated fibrils (mg)	Proteinase K digestion	Double proteinase K digestion
Starting material	5.6	5	5
10	0.48	0.13	0.28
100	0.14	1.34	0.38
200	0.29	0.18	0.61
400	0.10	0.29	0.5

Table 3: Yields of Mimivirus polysaccharides purified by ion exchange chromatography starting for the fibrils with or without proteinase K treatment. Fractions eluted at NaCl concentration higher than 400 mM did not contain carbohydrate material. Polysaccharide composition of the different fractions is given in Table 2.

3.1.5 Separation of *Mimivirus* polysaccharides by Reverse Phase HPLC

The reverse phase HPLC on the intact fibrils was used as a complementary purification approach to isolate poly_1. This method in contrast with the ion exchange chromatography, allowed to isolate before the most polar polysaccharide (poly.1 Fig 3.8 A) and then the less polar (poly_2 Fig. 3.8 B). The same approach described previously was used to evaluate the proportion between the two polysaccharides in each eluted fraction. The HPLC profile showed several peaks (Fig. 3.11 A) and the first eluted (R.T. 2.92 min.), albeit representing a minimal part of the whole sample mixture loaded, contained poly_1 in pure form (poly_1: poly_2: 1:0,34) (Fig. 3.11 B). All the other fractions contained the two polysaccharides with varying proportions and further protease treatments, did not increase the yield in poly_1 or a better separation between the two polysaccharides. Indeed, this approach, although valuable in isolating the acidic polysaccharide, was not explored further.

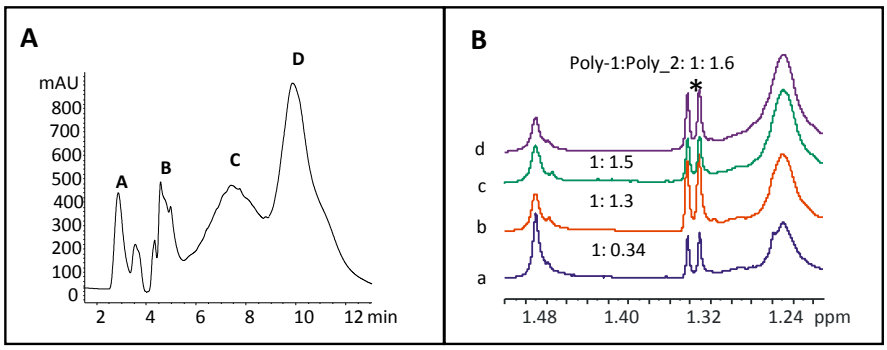


Fig. 3.11 A) Reverse phase HPLC analysis of untreated fibrils. The elution of the polysaccharide material was monitored by UV 206 nm. The peaks containing the polysaccharide material are labelled with letter A to D**; B)** (600 MHz, 298K) expansion of ^{1}H NMR spectra of the products of reverse phase HPLC analysis. a consists primarily of acid polysaccharide, as indicated from the ratio poly_1: poly_2 = 1: 0.34; b, c, and d contain a mixture of the two polysaccharides. The signals at 1.32 ppm, indicated by *, are impurities.

3.1.6 NMR analysis and molecular weight determination of the Mimivirus *fibrils* polysaccharides

To confirm the structure of the two polysaccharides, a complete NMR analysis was carried out on the polysaccharides purified by ion exchange chromatography. A comparison between the mono-dimensional DOSY spectra of the untreated fibrils and the fractions where the two polysaccharides were alternatively enriched, strongly highlighted the efficiency of the anion exchange chromatography as purification method (Fig. 3.12). In the fraction 100 mM NaCl, come from one digestion (poly_1: poly2: 1:5), the anomeric signals of the Poly_2 were increased (5.1-4.6 ppm), while the signal at 4,78 ppm, corresponding to the anomeric signals of the Poly_1, was almost absent. Contrary, the fraction 400 mM NaCl, come from a double digestion (poly_1: poly_2:1:0.5), disclosed that the anomeric signals of the poly_1 at 4,78 ppm and the signal at 1.47 ppm (diagnostic of a pyruvic acid) were strongly augmented compared with the untreated fibrils and the fraction 100 mM NaCl.

Despite the anomeric signals of the poly_2 were less represented, but still visible.

Definitively, the structures of the two polysaccharides were confirmed through a complete set of bi-dimensional spectra, both homo and heteronuclear.

The inspection of HSQC spectrum of the fraction 100 mM NaCl (Fig. 3.12 B) disclosed the presence of three anomeric signal instead of five, and the region of carbinolic signals was less crowded, demonstrating that the polysaccharide 2 was in pure form. The values of protons and carbons of each residue were comparable with the values of the mixture (Tab 1).

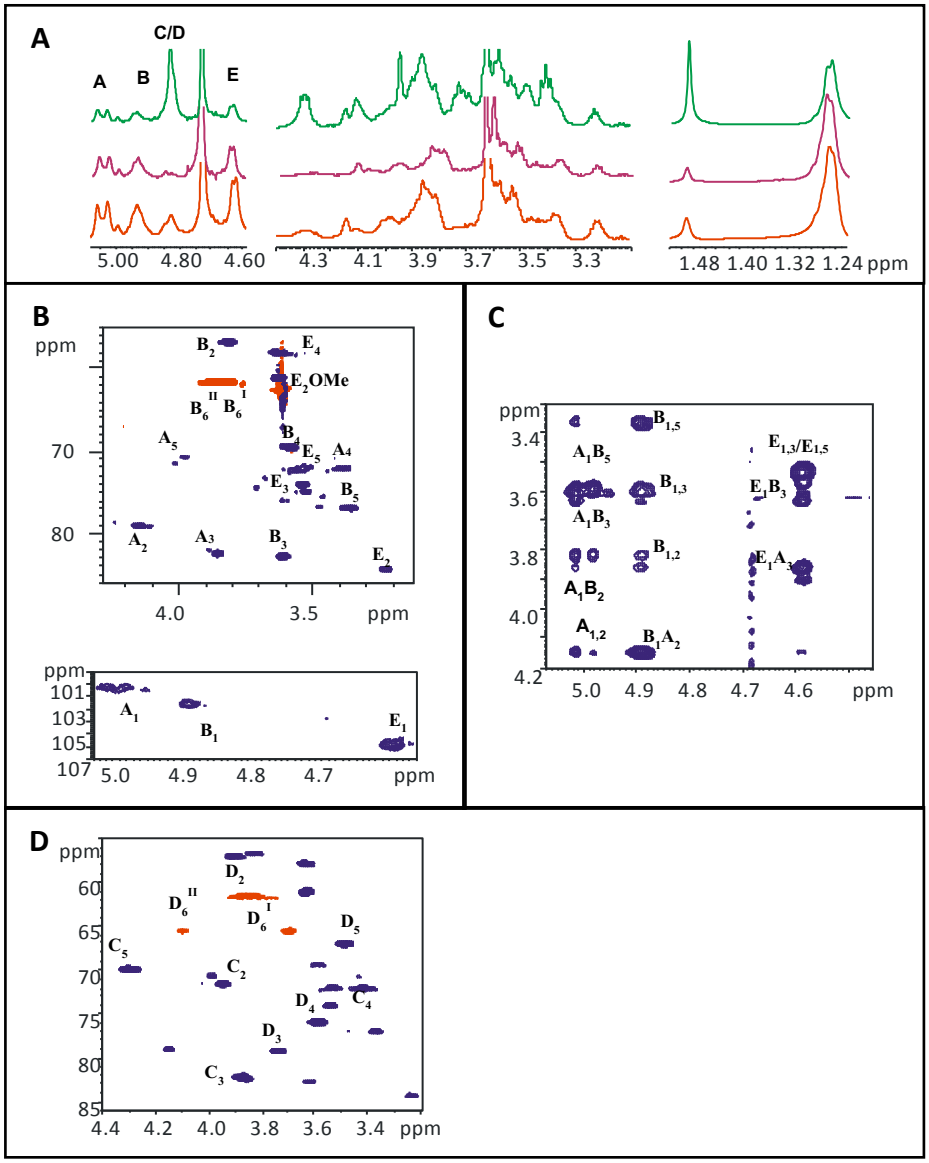


Fig. 3.12. **A)** (600 MHz, 310K) Expansion of mono-dimensional DOSY spectra of the untreated fibrils (red), the fraction 100mM NaCl of the anion exchange on digested fibrils (violet) and the fraction 400 mM NaCl (green).; **B)** Expansion of HSQC spectrum of the fraction 100 m NaCl: upper carbinolic region; lower anomeric region; **C)** Expansion of NOESY spectrum of the fraction 100 mM NaCl; **D)** Expansion of the HSQC spectrum of the fraction 400 mM NaCl. Only the densities of the poly_1, which is enriched in the fraction 400 mM NaCl, are labeled. Letters used for densities attribution follows the system of Table 1.

The integration of a complete set of spectra allowed to confirm the structure of polysaccharide 2 (Fig. 3.12 C).

The analysis of HSQC spectrum of the fraction 400 mM NaCl (Fig. 3.12 D) revealed that the signal of poly_1 were enriched compared to those of poly_2 allowing the identification of the carbon six of the GlcNAc (residue), which was not visible in the HSQC of the mixture due to the lower amount of this polysaccharide. The value is reported in Tab. 1.

In addition, the fractions where the poly_1 and poly_2 were in pure form have been used to determine the molecular weight of the two polysaccharides, trough SEC-HPLC. The comparison with known molecular weight dextrans revealed that the molecular weight of ply_1 (79 KDa) was more than twice that of the poly_2 (33,8 KDa) Tab. 4.

	MW(Da)	log MW	Time(min)	Elution volume
Dexstran	5000	3,70	12,5	10
Dextran	50000	4,70	10,7	8,56
Dexstran	150000	5,18	9,8	7,84
Dexstran	410000	5,61	9,2	7,36
Dexstran	670000	5,8	8,7	6,96
poly_1	79442,82	4,9	10,3	8,24
poly_2	33884,42	4,53	11	8,8

Table 4: Molecular weight of the dexstran used to calculate the molecular weight of the poly_1 and poly_2. Dexstran with molecular weight from 5000 Da to 670000 Da were inject on the TSK column (50 µl of a solution 1 mg /ml) and were used to construct a calibration curve with the following straight-line equation: y= 0.7082x + 10.764. Each polysaccharide was injecting on TSK column (50 µl of a solution 1 mg /ml) and the molecular weight was determined by solving the straight-line equation.

3.1.7 Identification of the protein/s attached with the two polysaccharides

The identification of the protein/s connected with the two polysaccharides is inserted in a very complex scenario, due to the lack of information on the fibrils. Indeed, many questions required to be address:

- ✓ Are the fibrils composed by one or more proteins? Is there one glycosylated protein or more than one?
- ✓ Which are the globular heads proteins attached to the fibrils shafts?
- ✓ Which are the proteins involved in the fibrils formation?

In attempt to identify the glycosylated protein/s composing the fibrils, a comparative proteomic analysis was carried out on the isolated fibrils, the defibered virus and total virus (control), appropriately lysed. For a first evaluation of the different protein content of each fraction, an SDS Page was performed on these samples. The proteins were visualized by conventional Blue Coomassie staining and PAS staining to visualize the glycosylated proteins. This allowed to identify the glycosylated and non-glycosylated proteins that were most enriched in the fibrils (Fig. 3.13).

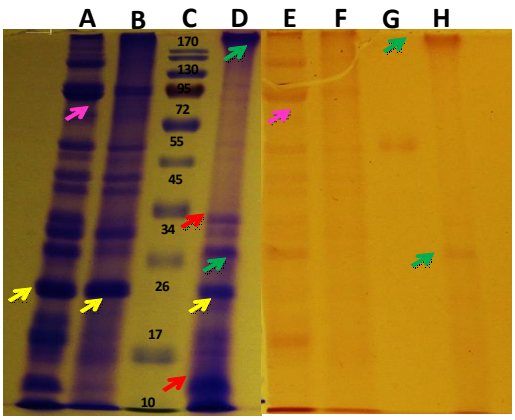


Fig. 3.13 SDS Page, gel 12%. Comassie Blue staining **A** to **D** and periodic acid-Schiff method **E** to **H**; **A** and **E** are Mimivirus total(control), B and F are the defibered virus; **D** and **H** are the isolated fibrils; **C** is a proteins marker; **G**) is the peroxidase as positive control for glycoprotein detection. The green arrows indicate the glycosylated proteins. The red arrows indicate the main proteins that are enriched in fibrils compared to the defibered virus. The yellow arrows indicate one protein that is abundant in each sample. The fuchsia arrow indicates one band that could be the major capsid protein, also glycosylated.

Comparing the fibrils with the total virus, made clear that most of the proteins with MW between 150 kDa and 30 kDa were absent from the fibrils fraction. In the fibrils

there were two main enriched proteins (marked with red arrow Fig. 3.13), one < 34KDa and one < 17KDa, compared to the defibered virus, that were not glycosylated, suggesting that they could be involved in the fibrils formation. In addition, there was a protein ~23KDa (labelled with the yellow arrow Fig. 3.13) that was abundant in all fractions, thus its contribute to the fibrils was not clear. The comparison between the Coomassie Blue staining and the PAS staining highlighted two main proteins in the fibrils that were glycosylated (marked with green arrow Fig. 3.13), one at high molecular weight (> 170 KDa) and one at lower molecular weight (~28 KDa). The band a lower molecular weight was compatible with the presence of small oligosaccharides, while the high band could correspond to one or more proteins connected with the two polysaccharides. Indeed, the heterogeneity of this band > 170 KDa can be explained by the molecular weight heterogeneity of the polysaccharides. A comparative proteomic analysis was performed on the isolated fibrils, the defibered virus and the total virus (as control) in collaboration with the Institute of Grenoble. The proteomic data were interpreted using a software developed at IGS, allowing the comparison between the different fractions. The proteins related to the fibrils were identified by looking at the proteins that decreased in the defibered virus compared to the total virus (Tab. 5) and at the proteins that increased in the fibrils with respect to the defibered virus (Tab. 6). This analysis disclosed that at list seven proteins were enriched in the fibrils samples: R135, L724, R362, L567, L719, R610 and L829. The proteins L829 and R362 had a mild decrease in the defibered virus compared to the total virus, as indicated from the p-Value (Tab. 5), but their content increased significantly in the fibrils fractions compared to the defibered virus, as suggested form the p-value (Tab. 6). Based on this finding, these proteins were considered as significant for the construction of the fibrils. Of interest R135 and L829 are absent in M4, a mutant of Mimivirus without fibrils, supporting our idea that both could be involved in fibrils composition. All the identified proteins were hypothetical proteins, except for R135 and R362, which are putative oxidoreductases. In addition, these proteins are very well conserved in the

Mimiviridae family, suggesting that could play an important role in Mimivirus biology.

These are only preliminary studies and required to be confirmed. Indeed, is working in progress a proteomic analysis on the fibrils isolated in mild condition and dialyzed against water, using a dialysis tube of 1 MDa, to avoid the presence of possible contaminants proteins, along with the proteomic analysis of the defibered virus and the total virus. This analysis is the starting point to identify all the proteins related to the fibrils, and further studies are necessary. In particular, it is important to identify the position where glycosylation takes place and to clarify the type of glycosylation (N- or O-linked). In this frame, it is necessary to perform a comparative proteomic analysis between the native (and glycosylated) proteins and those deglycosylated.

Rank	Protein	Mimivirus total	Defibered virus	p-Value	Distance
3	R135	175	88	5,86E-07	14,35
5	L724	84	35	1,51E-05	11,10
24	R610	49	26	0,0090492	4,71
9	L567	100	54	0,0004772	7,65
1	L719	48	8	5,31E-08	16,75
87	R362	82	61	0,088911	2,42
217	L829	79	71	0,4068	0,90

Tab.5 Comparison of the protein content of the Mimivirus total and the defibered virus using a software developed at IGS, Aix Marseille University. This tool is based on a distance and a p-value, higher is the distance between the number of peptides identified for the two samples and lower is the p-value, giving the best rank. The proteins that are decreased in the defibered virus compared to the total virus, should be related to the fibrils.

Rank	Protein	Defibered virus	Fibrils	p-Value	Distance
1	R135	88	449	3,96E-70	159,87
2	L724	35	206	4,18E-36	81,46
10	R610	26	114	2,04E-17	38,43
3	L567	54	211	3,71E-28	63,16
5	L719	8	95	4,45E-23	51,57
6	R362	61	196	3,91E-22	49,29
7	L829	71	200	1,16E-19	43,60

Tab.6 Comparison of the protein content of the defibered virus and the fibrils using a software developed at IGS, Aix Marseille University. This tool is based on a distance and a p-value, higher is the distance between the number of peptides identified for the two samples and lower is the p-value, giving the best rank. The proteins that are increased in the fibrils respect to the defibered virus, should be related to the fibrils.

3.1.8 Conclusions

Previous studies have shown that Mimivirus contains different sugars, Rha, GlcNAc and Vio, and have also established that this virus has the biosynthetic machinery to produce them (Parakkottil Chothi et al., 2010; Piacente et al., 2012, 2014b), leaving open the questions of how these monosaccharides are linked and where they are located.

First, fibrils were detached from the virus by the protocol set up in this thesis work, and monosaccharide analysis clearly revealed that the sugars are attached to the fibrils and are not part of the body of the virus.

Chemical analysis confirmed the presence of Rha, GlcN, VioNAc and 2OMe-VioNAc, and disclosed their linkage pattern, with rhamnose 3- or 2,3-linked, viosamine only terminal and glucosamine 3- or 3,4,6-linked.

NMR analysis was performed on the fibrils and combining the information from different 2D sequences, the repeating units of two polysaccharides were detected (Fig. 3.7) and named poly_1 and poly_2, respectively (Fig. 3.8).

The first polymer is characterized from a linear disaccharide repeating unit made of 3)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→, with GlcNAc further having at position 4,6 a pyruvic acid. The finding of the pyruvate substituent agrees with the bioinformatic analysis of Mimivirus genome, which assigned to the gene L143, the putative function of a polysaccharide pyruvyltransferase domain; of note, this gene is the last gene of Mimivirus nine-gene cluster (Piacente et al., 2012) where many of the genes related to the glycan biosynthetic machinery are located (as discussed in Chapter 2 and 4). The second polysaccharide has a branched repeating unit with the sequence 2)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→ in the linear backbone, with rhamnose further branched at position 3 with viosamine, methylated at position 2 and acetylated at position 4. The presence in vivo of the viosamine acetylated at the amino function allowed to totally demonstrate the role of the N-terminal domain of L142 as an N-acetyltransferase (Piacente et al., 2017) (see Chapter 5, published work). The viosamine is also methylated at C-2, but no putative methyltransferase appears in the nine-gene cluster involved in Mimivirus glycans formation. Does Mimivirus have the gene for the viosamine methylation? (see Chapter 4).

The occurrence of two different repeating units prompted further questions about the way they were assembled, namely if they appeared as blocks within one polysaccharide chain (Fig. 3.9 A) or if they built separate polysaccharides each attached to a distinct protein (Fig. 3.9 B) or to same protein (Fig. 3.9 C).

In order to get some insights into these issues, the fibrils were purified by two complementary approaches: anion exchange chromatography and reverse phase HPLC.

First, results obtained on the intact fibrils excluded the hypothesis of a polysaccharide chain containing the two repeats (Fig. 3.10 A) because poly_1 and

poly_2 were isolated in a rather enriched form, which suggest that each polysaccharide was attached to a protein, as depicted in figure 3.9 B. Moreover, the isolation of fractions containing both polysaccharides suggested that they could also be linked at the same protein as represented in fig. 3.9 C.

To verify this possibility, the fibrils were treated with proteinase K to degrade the peptide backbone and this treatment improved the isolation yields in both polysaccharides, presumably because not anymore interconnected by the same polypeptide backbone. In addition, a second digestion further improved the yields and the purity of both species, suggesting that one single protease digestion was not exhaustive probably due to the presence of the polysaccharide, maybe increasing the number of protease digestions, the separation between the two species will improve further. The structures of the two polysaccharides were confirmed through a complete set of 2D spectra on the fraction of the anion exchange chromatography, where they were almost in pure form (Fig. 3.12).

Regarding the novelty of the structures found, they have no resemblance with any in the eukaryotic domain, while some matches can be found in those reported for bacteria.

Indeed, search on the Bacterial Carbohydrate Structure Database (Toukach, 2011) found that the O-antigens of some strains of *Klebsiella pneumoniae* (Knirel' and Kochetkov, 1994), *Serratia marcescens* (Knirel' and Kochetkov, 1994) and *Agrobacterium tumefaciens* (De Castro et al., 2004) were identical to poly_1 devoid of the pyruvate substituent.

Concluding, the Mimivirus fibrils are decorated with two distinct polysaccharides and understanding if the two polysaccharides are linked to different proteins (Fig. 3.9 B) or to the same protein (Fig. 3.9 C) is under investigation. A comparative proteomic analysis was done on the fibrils, the defibered virus and the total virus (control) in order to identify the glycosylated proteins composing the fibrils. Preliminary results disclosed that at list seven protein were enrich in the fibrils sample (Tab 5 and 6) but these results required to be confirmed. The next step will

be a comparative proteomic analysis between the glycosylated fractions and the de-glycosylated one, to identify the peptide were the two polysaccharides are branched and consequentially clarify the type of glycosylation (N- or O-glycosylation). At the later point, the glycosylated proteins and possibly other proteins most commonly found in isolated fibrils can be followed in vivo using the available cellular biology techniques developed at IGS. This study could lead to the understanding of the mechanism for fibril assembly on the viral capsids that remain uncertain until now.

3.2 The fibrils of the giant DNA virus *Megavirus chilensis* exhibit a different glycosylation pattern compared to *Mimivirus*.

3.2.1 Production and purification of *Megavirus chilensis* virions

In order to elucidate the structure of the glycans associated to the Megavirus fibrils, a huge number of viral particles (~$3x10^{12}$ viral particles) was produced and purified, as described for Mimivirus. The purified viral particles appeared as a white band of a density ranging from 1.4 to 1.3, as for Mimivirus (Fig. 3.14). The production was controlled by light microscopy, and verified the absence of bacteria, was used for the chemical characterization of *Megavirus* fibrils glycans.

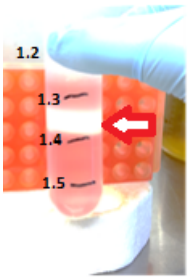


Fig. 3.14 *Megavirus chilensis* viral particles purified on CsCl gradient. The viral particles appeared as a white disk of density between 1.4 and 1.3 (arrow).

3.2.2 Isolation of *Megavirus chilensis* glycans

The Megavirus glycans were isolated using the same protocol set up for Mimivirus. Also in this case, the treatment in stronger condition (DTT 50 mM,
2h, 100°C) was the most effective for the removal of the fibrils (Fig. 3.15), despite most of the viral particles were damaged. The isolated glycans were checked by ^{1}H NMR, disclosing a region of anomeric signals (^{1}H range 5.2-4.92 ppm), and carbinolic area not well resolved (^{1}H range 4.4-3.1 ppm), two main acetyl signals (~2.0 ppm), a methyl group of the pyruvic acid (1.47 ppm) and a group of methyl signals characteristic of 6-deoxysugars at 1.25 ppm. (Fig. 3.16).

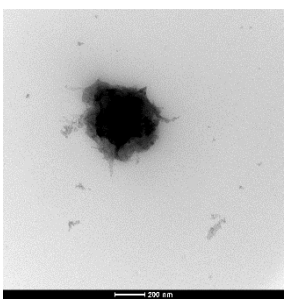


Fig. 3.15 Electron microscopy of *Megavirus* viral particle treated with DTT 50 mM, 2 h at 100°C. The fibrils were completely removed, but the viral particle was damaged.

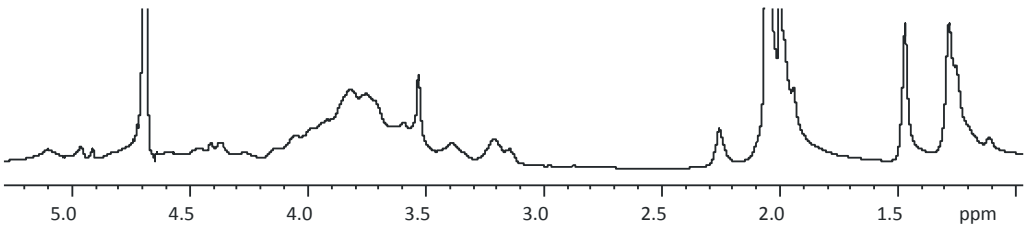


Fig. 3.16 (600 MHz, 298 K, D_2O).^{1}H NMR of *Megavirus chilensis* glycans fibrils.

3.2.3 Characterization of Megavirus *chilensis* glycan structure

The glycans connected to the *Megavirus* fibrils were investigated, as for *Mimivirus*, by chemical analysis and NMR analysis.

Sugar composition and linkage analysis of the glycans fibrils

Previous studies have been demonstrated that *Megavirus chilensis* has a functional pathway for UDP-D-GlcNAc (Piacente et al., 2014b) and for UDP-L-RhaNAc (Piacente et al., 2014a) and have been hypothesized that the UDP-L-RhaNAc could be converted in UDP-L-QuiNAc by the product gene of *mg536*. Preliminary glycans analysis of the viral particles have confirmed the presence of GlcNAc and RhaNAc, instead was less clear the presence/absence of the QuiNAc (Piacente et al., 2014a). To complete these data, the fibrils were transformed in the corresponding acetylated methyl glycosides (AMG) derivatives and analyzed by GC-MS. This analysis confirmed the presence of GlcNAc and unequivocally revealed the presence of the QuiNAc, through a comparison with a QuiNAc standard (Fig. 3.17).

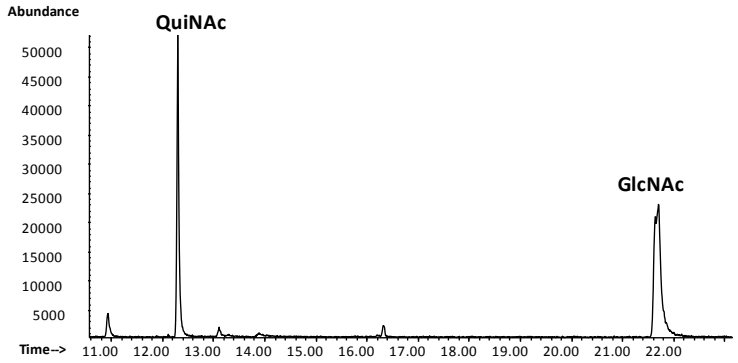


Fig. 3.17 GC-MS analysis of acetylated methyl glycosides (AMG) of Megavirus *chilensis* glycans fibrils. This analysis confirmed the presence of GlcNAc and identify the presence of the Qui2N, despite the RhaNAc could not be detected with this type of derivatization.

Despite, with this analysis the peak corresponding to the RhaNAc was not detected, suggesting that the RhaNAc could be N-linked, because methanolysis cannot break this type of linkage.

The GC-MS analysis of the fibrils, transformed in the corresponding partially methylated alditol acetates (PMAA), revealed the monosaccharides branching points as 3-dhexN, t-dhexN, 3-GlcN, 3,4-GlcN and 3,4,6-GlcN (Fig. 3.18).
Due to the lack of appropriate standards, was not possible to attribute the t-dhexN and the 3-dhexN to RhaN or QuiN.

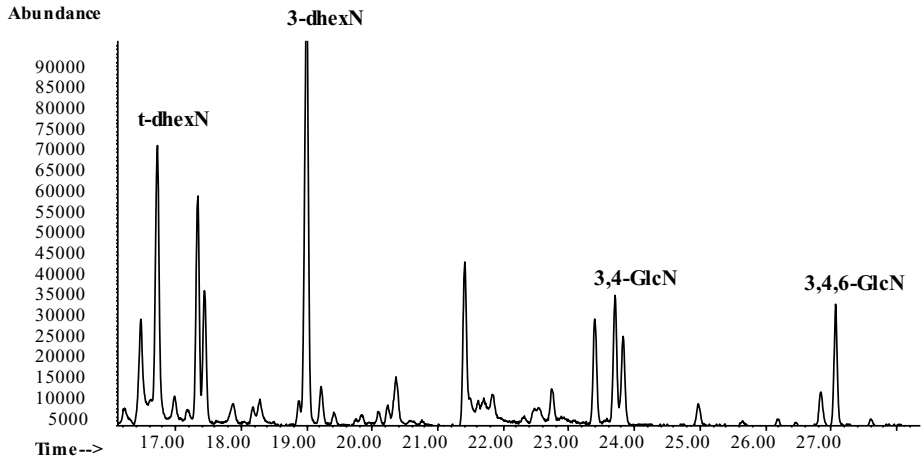


Fig. 3.18 GC-MS analysis of partially methylated alditol acetates (PMAA) of Megavirus *chilensis* glycans fibrils. This analysis revealed the monosaccharides branching points as 3-dehexN, t-dehxN, 3,4-GlcN and 3,4,6-GlcN. Despite was not possible to discriminate between RhaN and Qui2N.

NMR analysis of the fibrils

To complete the data of the chemical analysis, the complete set of bi-dimensional NMR spectra, both homonuclear and heteronuclear, was performed. Unfortunately, the resolution of the NMR spectra was not appropriated, and no structural information could be deduced (Fig. 3.16).

3.2.4 Purification of the *Megavirus chilensis* fibrils glycans

With the intent to obtain NMR spectra with good resolution to resolve the glycan structure, the fibrils were digested with proteinase K, reducing the interference of the protein part. This treatment did not improve the resolution of the spectrum (does not shown). So, the digested fibrils were purified by ion exchange chromatography. The glycans connected to the fibrils were separated by increasing the anionic strength of the eluent NaCl and the eluted fractions were analyzed by ^{1}H NMR (Fig. 3.19). The carbohydrate material was eluted between 100 mM and 400 mM NaCl.

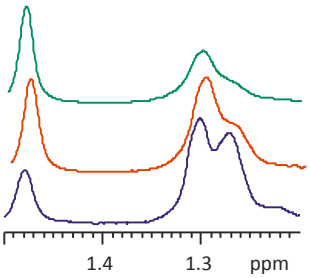


Fig. 3.19 (600 Mz, 310 K) Expansion of ^{1}H NMR spectra of the anion exchange fractions on Megavirus digested fibrils. In blue 100 mM NaCl; In red 200 mM NaCl; In green 400 mM NaCl.

The ^{1}H NMR analysis disclosed that the fraction eluted at low NaCl concentration (100 mM) was enriched of the signal of the methyl group of the six-deoxy-sugars (Fig. 3.19, profile in blue), while, increasing the strong of the eluent, the signal diagnostic of the pyruvic acid (1.47 ppm) increased (Fig. 3.19, fraction 200 mM in red and fraction 400 mM in green), suggesting that more than one polysaccharide was attached to the fibrils (Fig. 3.19).

The fraction 100 mM NaCl (Fig. 3.19 in blue), had a better resolved carbinolic region (4.1- 3.2 ppm) and was analyzed by NMR (1D, 2D).

The ^{1}H NMR displayed three anomeric signals (4.98, 4.96 and 4.92 ppm) labelled with the letter A to C, a carbinolic region (^{1}H range 4.4-3.1 ppm), two main acetyl signals (~2.0 ppm), a lower signal of the methyl group of the pyruvic acid (1.47 ppm), and a group of methyl signals characteristic of 6-deoxy-sugars at 1.3 ppm (Fig. 3.20 A). The protons and carbons chemical shifts for each residue are reported in Tab. 7.

Residue	Nucleus	1	2	3	4	5	6
A	^{1}H	4.98	4.33	3.94	3.41	3.56	1.29
3-α-RhaNAc	^{13}C	101.6	53.8	77.4	72.2	70.4	18.0
B	^{1}H	4.96	4.37	3.92	3.54	3.7	1.29
3-α-RhaNAc	^{13}C	101.6	54.0	77.8	72.2	70.4	18.0
C	^{1}H	4.92	4.41	4.01	3.75	3.71	1.29
t–α-4OMe-RhaNAc	^{13}C	102.2	54.2	69.2	83.2	69.4	18.0

Table 7. (600 MHz, 310 K, D_2O) ^{1}H and ^{13}C chemical shifts of the fraction 100 mM of the anion exchange chromatography on Megavirus digested fibrils. The three Rhamnosamine (residues A, B and C) are acetylated to the amino function (^{1}H 2.09, ^{13}C 23.8). The residue C is methylated at carbon 4 (^{1}H 3.71, ^{13}C 69.4).

The residues **A** and **B** have been attributed as α-RhaNAc. The anomeric proton and carbon of the residue **A** was almost coincident with that of **B** and the TOCSY spectrum from H-2 position (4.33 and 4.37 ppm respectively) indicated the positions of the other protons for each residue (Tab. 7). TOCSY spectrum showed intense correlations starting not from H-1, but from H-2, indicating the *manno* stereochemistry of these residues (Fig. 3.20 B). The values of the protons 6 (1.29 ppm) revealed that all were six-deoxysugars (Tab. 7). H-2/C-2 values (4.38/53.8 ppm and 4.37/54.0 ppm respectively) were diagnostic of an acetylated amino function, identifying these residues as RhaNAc.

Integration with HSQC spectrum information, disclosed that for both residues the C-3 (77.48 and 77.85 ppm, respectively) were shifted at low field compared to the reference values because glycosylated (Bock and Pedersen, 1983). Accordingly, NOESY spectrum (Fig. 3.20 C) correlated H-1 of **B** with H-3 of **A**. Finally, the

configuration at the anomeric carbon was α as inferred by the diagnostic carbon chemical value of C-5 (70.42 and 70.39 ppm respectively).

The residue **C** was identified as not substituted α-4OMe-RhaNAc. A combination of COSY, TOCSY and HSQC allowed the identification of all protons and carbons (Tab. 7). The inspection of HSQC spectrum revealed that the C-4 was a low field (83.24 ppm), because methylated at the corresponding hydroxyl function. The configuration at the anomeric carbon was α as indicated by the carbon chemical value of C-5 (69.4 ppm). This residue was linked to the O-3 of **B** as was evident from the NOESY correlation (Fig. 3.20 C).

A combination of all the spectra allowed the identification of the glycan structure as a trisaccharide of α-L-RhaNAc: 3)-B-(1→3)-A-(1→? in the linear backbone, with B further branched at O-3 with a terminal α-L-4OMe-RhaNAc (residue C) (Fig. 3.20 D).

Despite, it was not possible to understand where the anomeric carbon of the residue **A** was linked. One hypothesis could be that **A** is linked with another sugar that was not visible due to the low resolution of the spectrum (apparently trisaccharide), or that it is linked to a protein, and given its ^{13}C anomeric chemical shift (101. 6 ppm) it should be O-linked. This information contrasts with the absence of this unit in the AGM analysis, and further investigations are necessary to address the nature and the location of this oligosaccharide.

The NMR analysis of the fraction 100 mM, allowed to attribute the t-dhexN, found in the PMAA analysis of the fibrils, as 4OMe-RhaNAc and the 3-dhexN as 3-RhaNAc.

The anion exchange chromatography was preformed also on the double digested fibrils with protease K, but this purification did not improve the separation between the different species and did not add any further information about the glycans.

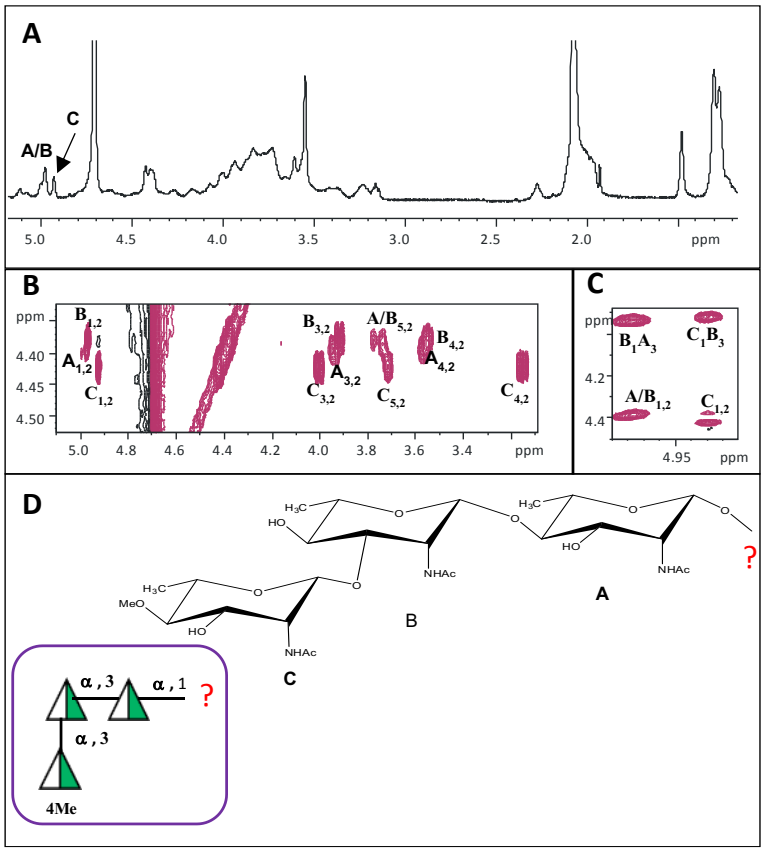


Fig. 3.20 (600 MHz, 310 K, D_2O) **A)** ^1H NMR of the fraction 100 mM NaCl; **B)** Expansion of the TOCSY spectrum; C) Expansion of the NOESY spectum; **D)** Structure of the trisaccharide of RhaNAc connected to the Megavirus fibrils. In the purple square there is the symbolic representation of the trisaccharide structure. Monosaccharides symbols follow the SNFG (Symbol Nomenclature for Glycans) system.

3.2.5 De-polymerization of *Megavirus* fibrils glycans

The chemical analysis of the fibrils has revealed the presence of RhaNAc, GlcNAc and QuiNAc, meaning that in addition to the trisaccharide of RhaNAc, identified by NMR, there should be at least another polysaccharide in the fibrils, as suggested from

the presence of other sugars in the composition analysis and in the purification results (3.19)

In order to elucidate the complex mixture of carbohydrates attached to the Megavirus fibrils, a strong acid hydrolysis was carried out on Megavirus viral particles ($1x10^{11}$) by using triflic acid (or trifluoromethanesulfonic acid, TFMS, for 4 h at 4°C). Due to the low amount of fibrils available, the solvolysis was performed directly on the intact particles. Triflic acid was used because it is able to cleave in a mild way the glycosidic linkages of the amino sugars without removing the acyl at the amino function. Importantly, depending on the temperature used, this treatment has some selectivity and may yield to the recovery of oligosaccharide fragments.

After the solvolysis, the sample was neutralized with ammonia 33% and purified by size exclusion chromatography (Biogel P2). The different fractions were analyzed by ^{1}H NMR (Fig. 3.21), revealing two fractions (C and D) with a better resolution of the anomeric and carbinolic regions. These fractions were subsequently purified by Reverse Phase HPLC and peaks collected were checked by ^{1}H NMR.

Despite the different HPLC profile (Figs. 3.22and 3.23 A), in both cases the^{1}H NMR analysis (Figs. 3.22 and 3.23B) disclosed the presence of all monosaccharides, as confirmed from a complete set of bi-dimensional spectra on the fraction II of the HPLC on the fraction C (Fig. 3.22 B) and of the fraction I of the HPLC on the fraction D (Fig. 3.23 B).

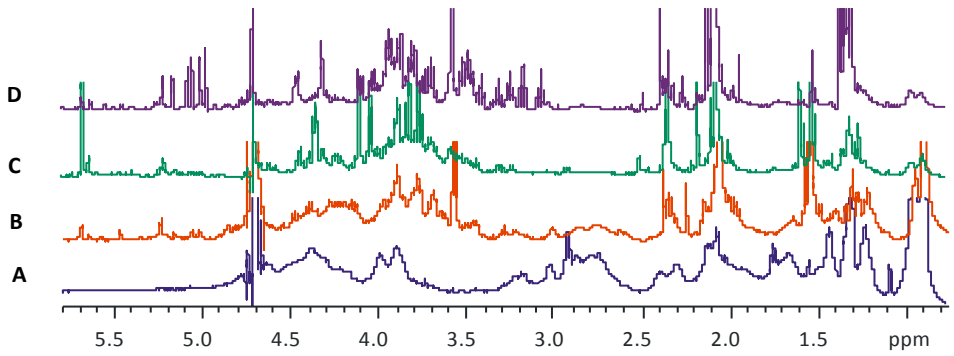


Fig. 3.21 (600 MHz, 298 K) ^{1}H NMR of the eluted fractions of the size exclusion chromatography on the Megavirus fibrils hydrolyzed wit TFMS 4h, 4 °C.

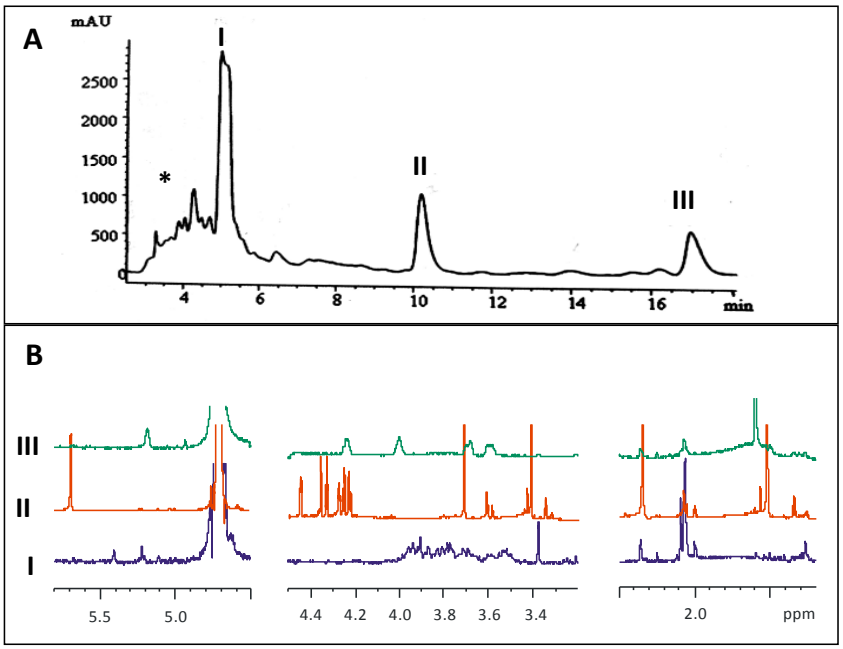


Fig. 3.22 A Reverse phase HPLC analysis of fraction C of the size exclusion chromatography on the fibrils treated with TFMS. The elution of the polysaccharide material was monitored by UV 206 nm. The peaks containing the carbohydrate material are labelled with the number I to III; The * indicate the HPLC fractions that were not well separated each from another and for which the NMR profile was not relevant (does not shown) **B** a **(**600 MHz, 298K) Expansion of ^{1}H NMR spectra of the products of reverse phase HPLC analysis on the fraction C . This analysis revealed some typical spectra of monosaccharides.

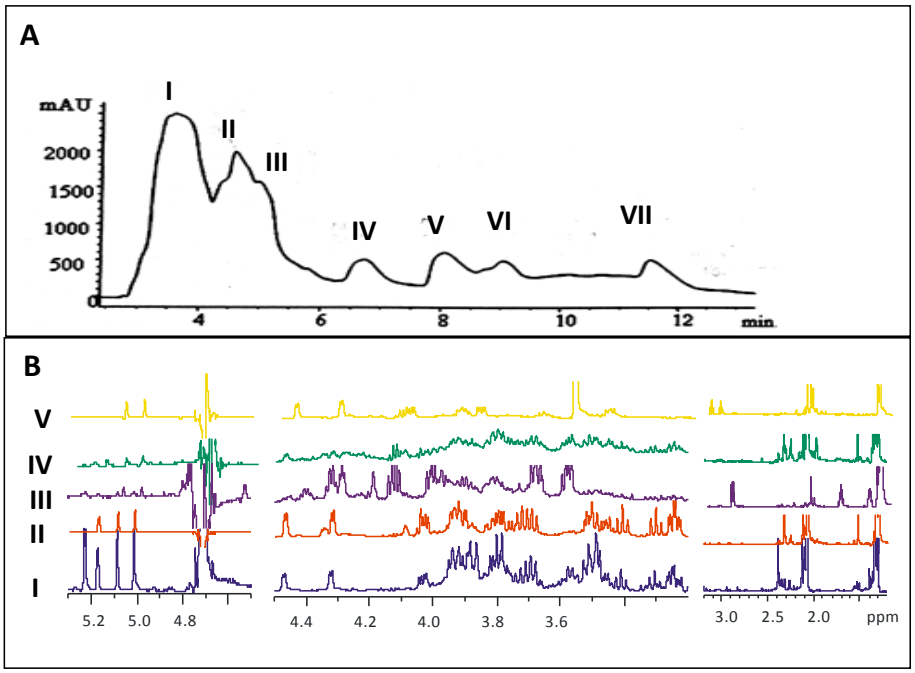


Fig. 3.23 A) Reverse phase HPLC analysis of fraction C and D of the size exclusion chromatography on the fibrils treated with TFMS. The elution of the polysaccharide material was monitored by UV 206 nm. The peaks containing the carbohydrate material are labelled with the number I to VII respectively; **B) (**600 MHz, 298K) Expansion of ^{1}H NMR spectra of the products of reverse phase HPLC analysis on the fraction D. The peaks from V to VII showed the same NMR profile. This analysis revealed a typical spectra of monosaccharides.

In order to recover small oligosaccharides, the Megavirus viral particles were hydrolyzed with TFMS in milder condition (4 h, -20 °C), with no appreciable change in the nature of the products collected, namely, only monosaccharides were recovered.

3.2.6 NMR analysis on the RP-HPLC products

A complete set of 2D homonuclear and heteronuclear NMR spectra was performed on fraction II (Elution time 10 min) of the RP-HPLC (Fig. 3.22 A), to elucidate the nature of the monosaccharide linked with the pyruvic acid (signal at 1.47 ppm) (Fig. 3.24).

The residue **A** was identified as α-GlcNAc. The *gluco* stereochemistry of **A** determined the efficient magnetization transfer in the TOCSY spectrum. In combination with the information from COSY and HSQC spectra, the proton and carbon chemical shift were determined (Tab. 8). The atypical values of the anomeric carbon (83.9 ppm) and the C-2 (67.7 ppm) compared to a not substituted α-GlcNAc (C-192 ppm; C-2 55 ppm) suggested the formation of a linkage between the acetal nitrogen with the anomeric carbon. This hypothesis is supported by HMBC correlation between the anomeric carbon and a carbon at 170 ppm. The configuration at the anomeric carbon was α, indeed the value of the C-5 (73.03 ppm) was the typical value of α-GlcNAc. The H-3 (4.32) correlated with a carbon at low field (80.31 ppm) because linked with the pyruvic acid, as revealed from the long range correlation $A_{3,}Pyr_2$ (Fig. 3.24 D). The HMBC did not show any other correlation with the pyruvic acid, but it was possible to assume that the Pyruvic acid was linked in 3 and 4, because the C-4 value was a low field (78.38 ppm) compared to the value for the α-GlcNAc reported in literature (72 ppm). For the pyruvic acid the configuration at the acetal carbon atom was S, as confirmed from the comparison of characteristic chemical shifts of $^{1}H/^{13}C$ (1.49 and 24.6 ppm respectively) reported in literature (Garegg et al., 1980).

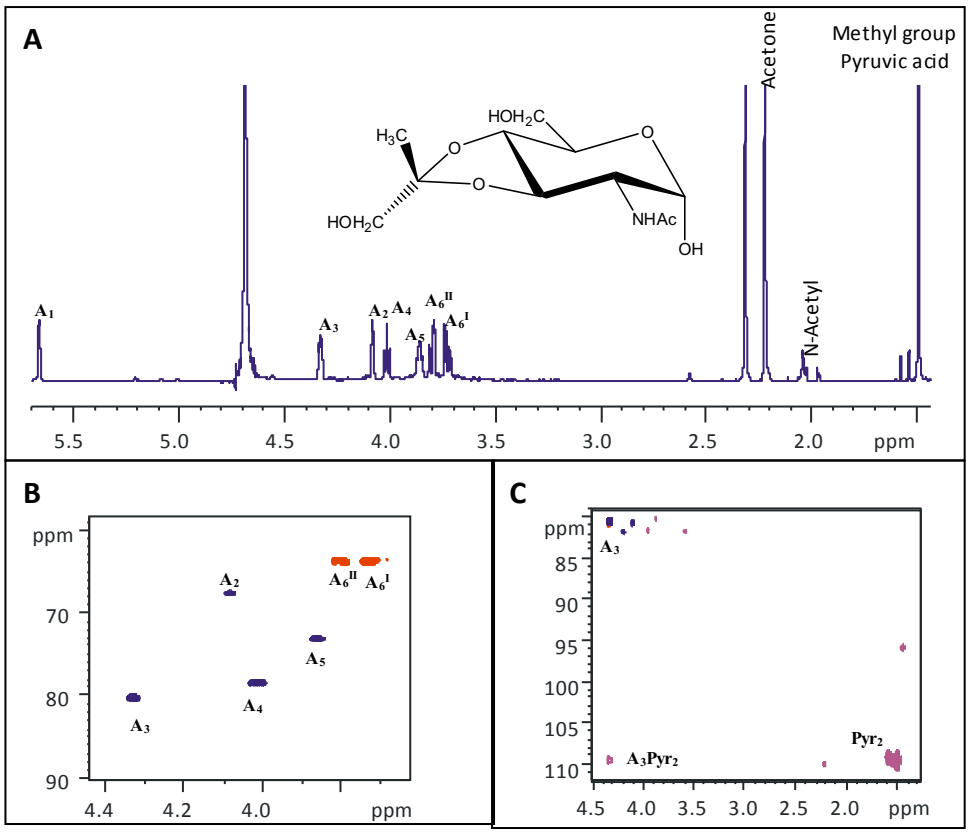


Fig. 3.24 (600 MHz, 310 K) **A)** ^{1}H NMR of the fraction II (Elution time 10 min) come from the RP-HPLC of the fraction C of the size exclusion chromatography. The ^{1}H NMR revealed that was a monosaccharide; **B)** Expansion of HSQC spectrum; **C)** Expansion of HMBC spectrum, showing the long range correlation A_3,Pyr_2. A combination of all the 2D spectra allowed to identify this sugar as α-GlcNAc linked in 3,4 with the pyruvic acid with the S configuration at the acetal carbon. The chemical shifts are reported in Tab. 8.

Residue	Nucleus	1	2	3	4	5	6
A	^{1}H	5.66	4.08	4.32	4.01	3.85	3.73/3.79
3,4-α-GlcNAc	^{13}C	83.9	67.6	80.3	78.4	73.0	63.7
Pyruvic acid	^{1}H	-	-	1.49			
	^{13}C	178.0	109.5	24.5			

Tab. 8 (600 MHz, 310 K, D_2O) ^{1}H and ^{13}C chemical shifts of the fraction II (Elution time 10 min) come from the RP-HPLC of the fraction C of the size exclusion chromatography. The Glucosamine was acetylated to the amino function (^{1}H 2.04, ^{13}C 23.4). The residue C is methylated at carbon 4 (^{1}H 3.71, ^{13}C 69.4).

Fraction I (Elution time 3.5 min), of the RP-HPLC on the fraction D (3.23 A) , was analyzed by NMR to understand the nature of the other monosaccharides obtained from this treatment.

The ^{1}H NMR (Fig. 3.25 A) displayed the presence of six anomeric protons (5.2-4.6 ppm), a crowded carbinolic region (4.5-3.2 ppm), and several signals of methyl groups in the high field region, due to N-acetyl groups (2.09-2.04ppm) and to the six positions of 6-deoxy-sugars (1.3 ppm).

The residues **A** and **E** have been attributed as α and β-GlcNAc respectively. All the protons were identified using the information of the COSY and TOCSY spectra. The attribution of the corresponding carbons values was possible by integration of HSQC (Fig. 3.25 B) and HSQC-TOCSY information.

The residues **B** and **F** were identified as α and β-QuiNAc respectively. The protons values for both residues were easily detected due to the efficient magnetization transfer in the TOCSY spectrum, indicating their *gluco* stereochemistry. The value of the protons six at 1.28 ppm, was diagnostic of six-deoxysugars, and the C-2 value (55.6 and 58.2 respectively) was typical of an acetylated amino function, revealing that **B** and **F** were Qui2NAc (Fig. 3.25). Finally, the residue F was β configured at

the anomeric carbon, as indicated by characteristic H-1/H-3 and H-1/H-5 correlations in the T-ROESY spectrum, while the residue B was α, as evident from the absence of these correlations.

The residues **C** and **D** have been attributed to α and β-RhaNAc. The identification of all protons and carbons was made as illustrated in the paragraph 3.2.4. The values of the C-5 allowed to identify the residue **C** as α (69.4 ppm) and the residue **D** as β (73.7 ppm) (Fig. 2.3.25). On the basis of these results, it was possible to assign also the spectra of the fraction D of the size exclusion chromatography (Fig. 3.23 C and D), which contained α-Qui2N and RhaNAc α and β; fraction V, VI and VII the RhaNAc α and β methylated at O-4.

Residue	Nucleus	1	2	3	4	5	6
A	^{1}H	5.19	3.88	3.75	3.49	3.85	3.84/3.79
α-GlcNAc	^{13}C	92.0	55.5	72.30	71.3	72.9	61.8
B	^{1}H	5.14	3.89	3.7	3.2	3.9	1.3
α-Qui2NAc	^{13}C	91.9	55.6	71.7	77.0	68.9	18.0
C	^{1}H	5.05	4.30	4.00	3.38	3.90	1.27
α-RhaNAc	^{13}C	94.2	54.7	69.6	73.7	69.4	18.1
D	^{1}H	4.98	4.44	3.77	3.27	3.43	1.29
β-RhaNAc	^{13}C	94.0	55.5	72.9	73.4	73.7	18.0
E	^{1}H	4.69	3.67	3.53	3.46	3.46	3.73/3.90
β-GlcNAc	^{13}C	96.2	58.1	75.2	71.2	77.3	62.1
F	^{1}H	4.68	3.67	3.47	3.21	3.49	1.3
β-Qui2Nac	^{13}C	96.0	58.2	74.9	76.4	73.3	18

Tab. 9 (600 MHz, 310 K, D_2O) ^{1}H and ^{13}C chemical shifts of the fraction I (Elution time 3.5 min) come from the RP-HPLC of the fraction D of the size exclusion chromatography. The Glucosamine, rhamnosamine and quinovosamine were acetylated to the amino function (^{1}H 2.04, ^{13}C 23.4).

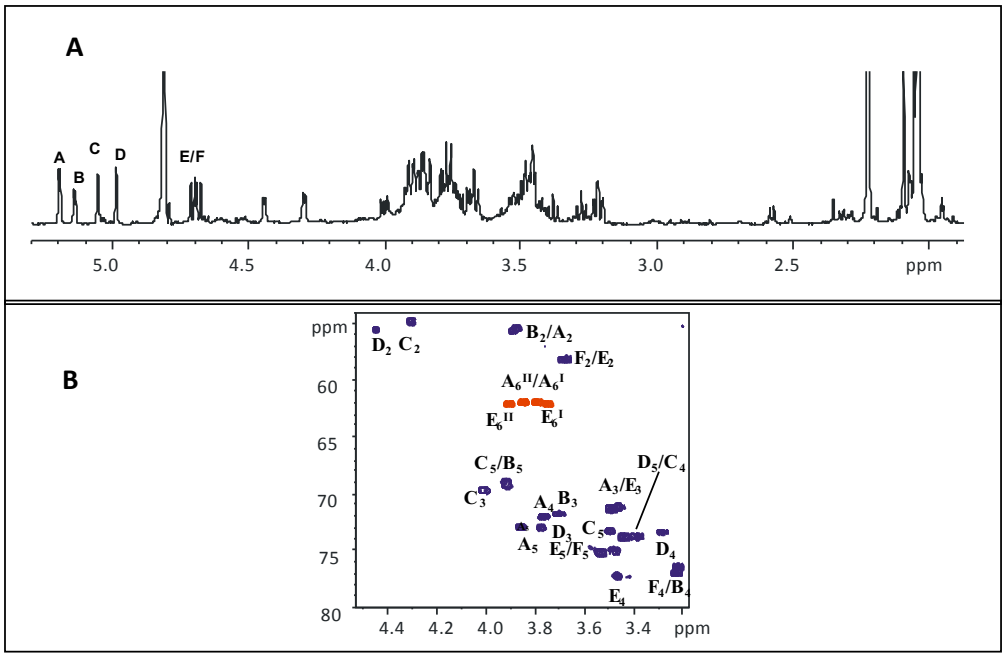


Fig. 3.25 (600 MHz, 310 K) **A)** ^{1}H NMR of the fraction I (Elution time 3.5min) come from the RP-HPLC of the fraction D of the size exclusion chromatography; **B)** Expansion of HSQC spectrum; This analysis revealed that was a mixture of monosaccharides.

3.2.7 Conclusions

Previous data have been demonstrated that the major sugars components of the *Megavirus chilensis* fibrils were GlcNAc and RhaNAc and the genes responsible for their production were also identified (Piacente et al., 2014b, 2014a). In this thesis work, these data were confirmed and in addition have confirmed that Qui2NAc is produced from the virus. How these sugars are organized is still under investigation. Purification results on the digested fibrils suggested that the fibrils should present different polysaccharide species, with one having the trisaccharide of RhaNAc: α-L-4OMe-RhaNAc-(1→3)-α-L-RhaNAc-(1→3)-α-L-RhaNAc-(1→, unfortunately, was not possible to understand where this trisaccharide was linked, or how the other monosaccharides, GlcNAc and QuiNAc, contribute to the formation of other polysaccharide species.

From the solvolysis of the viral particles, in strong and mild conditions, it was possible to obtain only monosaccharides. This analysis allowed to confirm the sugars fibrils composition based on the chemical analysis. In addition, the identity of the sugar linked with the pyruvic acid was revealed as GlcNAc. So, Megavirus chilensis shares with Mimivirus the GlcNAc modified with the pyruvic acid, but the type of linkage is different. Indeed, in Mimivirus the GlcNAc has the pyruvic acid linked to O-4 and O-6, while in Megavirus it is linked in O-3 and O-4. The pyruvic acid is a common component in bacteria extracellular polysaccharides, and the linkage in O-4 and O-6 occurs more often than the linkage in O-3 and O-4(Garegg et al., 1980). In addition, the pyruvic acid linked in O-3 and O-4 was found only with galacto and rhamnopyranosyl residues and not with glucopyranosyl residue as in Megavirus. The difference in the linkage of the pyruvic acid between Mimimivirus and Megavirus opens the question about the origin of the genes for the pyruvyltransferase in these viruses.

Concluding, the *Megavirus chilensis* fibrils seem to be composed from more than one polysaccharide species. An alternative hypothesis to the presence of more polysaccharide species is that the fibrils are densely substituted with small oligosaccharides or with monosaccharides only, thus explaining why we could never isolate an oligosaccharide after solvolysis and also why these fibrils appeared resistant to protease treatment.

3.3 Glycans composition of *Moumouvirus australensis* fibrils.

3.3.1 Production and purification of *Moumouvirus australensis* virions

In contrast with *Mimivirus* and *Megavirus chilensis*, the fibrils sugars composition of the members of the linage B was never investigated before. In attempt to give an answer, ~3x10^{12} viral particles of *Moumouvirus australensis*, a new isolate of the linage B, were produced and purified, using the same protocol adopted for *Mimivirus* and *Megavirus*. Surprisingly, the purified viral particles appeared as a white disk of a density of ~ 1.3, maybe due to the difference in the fibrils thickness and length (Fig. 3.26). The production was checked by light microscopy and used for preliminary analysis of the fibrils sugar content.

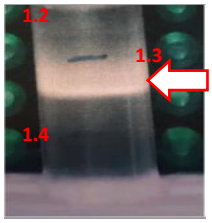


Fig. 3.26 Moumouvirus australensis viral particles purified on CsCl gradient. The viral particles appeared as a white disk of density of ~ 1.3 (arrow).

3.3.2 Isolation of *Moumouvirus australensis* fibrils glycans

The treatment 50 mM DTT, 2h at 100°C was effective also for the removal of M. australensis fibrils as shown from the TEM analysis of the treated virus particle (Fig. 3.27). The isolated glycans were analyzed by ^{1}H NMR, displaying a very complex situation due to the not resolved region of the anomeric and carbinolic signals (Fig. 3.28). In addition, there was a crowded region of N-acetyl group (2 ppm). In contrast with Mimivirus and Megavirus, the signal of the methyl group of the pyruvic acid (1.47 ppm) was absent.

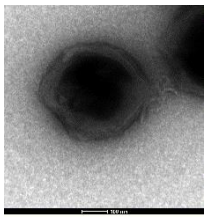


Fig. 3.27 Electron microscopy of Moumouvirus australienis viral particle treated with DTT 50 mM, 2 h at 100°C. The fibrils were completely removed.

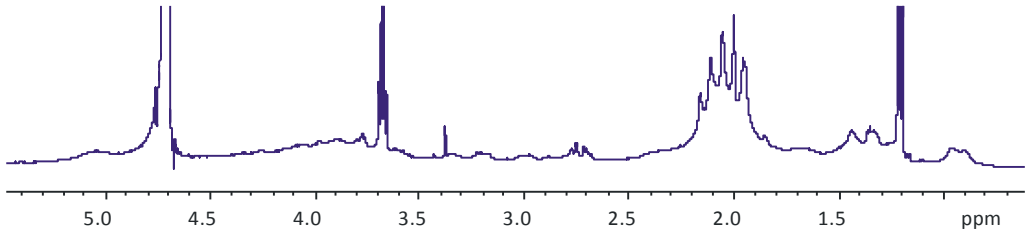


Fig. 3.28 (600 MHz, 310 K, D2O) ^{1}H NMR of the isolated glycans connected to the fibrils of *Moumouvirus australensis.*

3.3.3 Sugar composition of *Moumouvirus australensis* fibrils glycans

0.5 mg of fibrils were transformed in the corresponding acetylated methyl glycosides and analyzed by GC-MS. This analysis revealed the presence of GlcNAc, Qui2N and bacillosamine (2,4-diacetamido-2,4,6-trideoxy-glucopyranose or N',N'-di-acetylbacillosamine or diNAcBac) as major components and hexose and inositol as minor components (Fig. 3.29). The peak at 20 min was identified as diNAcBac, applying the fragmentation rules, indeed, the EI-MS spectrum contained a fragment at m/z 271 consistent with the oxonium ion of six-deoxy-sugar with two amino functions (Fig. 3.30).

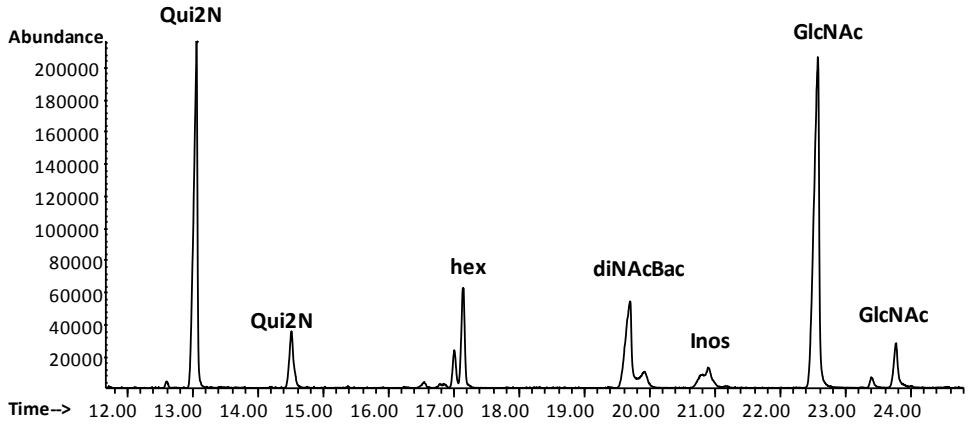


Fig. 3.29 GC-MS analysis of acetylated methyl glycosides (AMG) of Moumouvirus australensis isolated fibrils glycans. This analysis revealed that the major saccharides components were QuiNAc, GlcNAc, diNacBac and as minor component hexose and inositol.

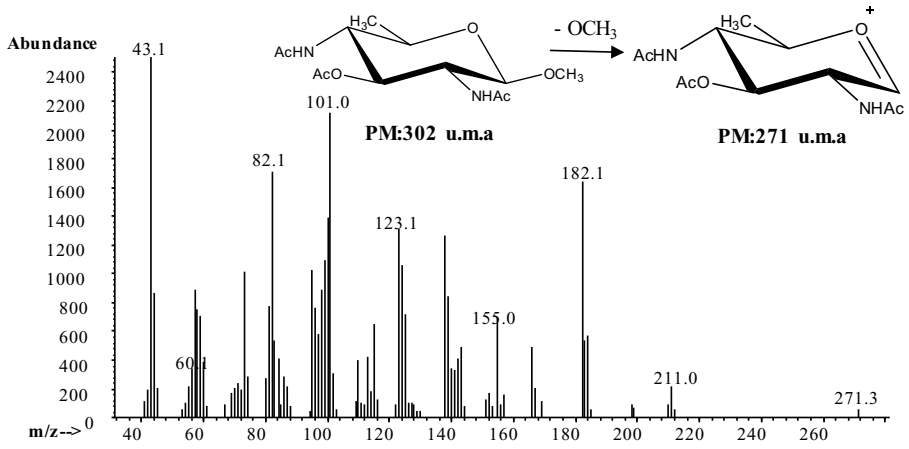


Fig. 3.30 EI-MS spectrum of the diNAcBac (RT 19.7 min). The fragment at m/z 271 is consistent with the oxonium ion of a six-deox-ysugar with two amino functions.

3.3.4 Conclusions

Moumouvirus australensis represents the first member of the linage B for which the sugars fibrils composition is determined. This is a preliminary analysis that disclosed the major saccharide components as GlcNAc, Qui2N and diNAcBac. So, *M. australensis* shared with Mimivirus and Megavirus the glucosamine, but in contrast with them, the glucosamine did not seem modified by the pyruvic acid, as suggested from the ^{1}H NMR analysis of the isolated glycans fibrils (Fig. 3.27). The pyruvic acid confers a negative charge to the polysaccharides that could play a role in the host interaction process, prompting the question if in *Moumouvirus australensis* there could be another type of modification (see chapter 4), disclosing at the later point the function of these sugars modifications. In addition, this virus presents the QuiNAc as Megavirus and further analysis are required to evaluate if also RhaNAc is present, since its biosynthesis is strongly related to that of QuiNAc as determined in Megavirus *chilensis* (see Chapter 2). However, *M. australensis* exhibits another rare 6-deoxy-aminosugar, the bacillosamine. Linkage analysis and NMR analysis are working in progress to elucidate the glycans structures associated to the *M. australensis* fibrils.

Does *Moumouvirus australensis* have the genes for the Glucosamine and the Quinovosamine? It is possible to think that in analogy with Mimivirus and Megavirus it has the genes to produce these sugars, but further investigations are required (see Chapter 4). Is it able to produce the bacillosamine?

Is this sugar composition unique of *M. australensis* or is it conserved along all the lineage B? It will be discussed in Chapter 4.

3.4 Collective discussion of Mimivirus, Megavirus and Moumouvirus results for fibril glycosylation

The first step to understand the innovative glycosylation system of the Megavirinae family (Gallot-Lavallée et al., 2017), was the elucidation of the fibrils glycans structures of *Mimivirus*, *Megavirus chilensis* and *Moumouvirus australensis* used as a prototypes.

In this thesis work was set up a protocol to isolate the glycans, detaching the fibrils from the viral capsid. It is interesting that the only action of the DTT 50 mM was enough to remove the fibrils in presence/absence of the heating. Indeed, it has been promoted the idea that the fibrils could be anchored to extended polypeptide loops of the major capsid protein (MCP)(Kuznetsov et al., 2010; Xiao et al., 2009) through di-sulfide bridges. In order to assess the basis of this hypothesis, a multiple alignment of the major capsid proteins of *Mimivirus* (L425), *Megavirus chilensis* (mg 464), *Moumouvirus australensis* (mc389) was done, revealing that there were conserved cysteine residues. In the major capsid protein of Mimivirus there were 4 cysteines, and three of them (C309, C351, C392) were conserved in Megavirus *chilensis* and only two (C351 and C392) in *M. australensis*, while the C306 was present only in *Mimivirus* (Fig. 3.31). Further investigations are required to confirm this hypothesis.

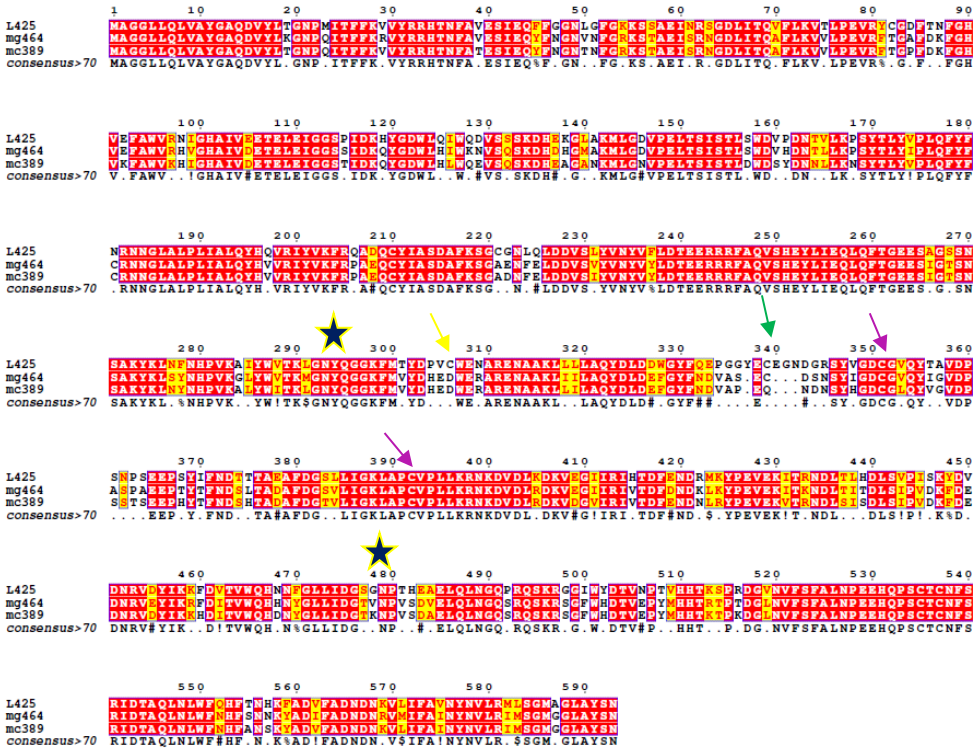


Fig. 3.31 Multiple alignment of the major capsid protein of Mimivirus L425 (YP_003986929.1), Megavirus *chilensis* mg464 (YP_004894515.1), Moumouvirus australensis mc389 (AVL94775.1). The stars indicate the loop of 190 aa of the major capsid proteins. The yellow arrow indicates the C306 present only in Mimivirus, the green arrow indicates the C309 that is conserved in Mimivirus and Megavirus. The pink arrows indicate the C351 and 392 that are conserved in the three. The multiple alignment was performed with Expresso T Coffee (Armougom et al., 2006) and the picture was done with ESPript (Gouet et al., 2003).

A combination of chemical and NMR analysis allowed to elucidate the sugar composition of *Mimivirus*, *Megavirus chilensis* and *Moumouvirus australensis* fibrils. The Mimivirus fibrils are heavily glycosylated with two distinct polysaccharides (Fig.): polysaccharide 1 (or poly_1, Fig. 3.8 A) has a linear repeating unit made of 3)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→, with GlcNAc further having at position 4,6 a pyruvic acid. Polysaccharide 2 instead (poly_2, Fig. 3.8 B) has a branched repeating unit, with 2)-α-L-Rha-(1→3)-β-D-GlcNAc-(1→ in the linear backbone, with Rha further branched at O-3 with a terminal β-D-2OMe-VioNAc. The *Megavirus chilensis* fibrils seems to be composed by more than one species, as

suggested from purification results. A tri-saccharide of RhaNAc was been identified (Fig. 3.20), despite it was not possible to understand how it was linked. In addition, other sugars compose the Megavirus fibrils, as GlcNAc linked in 3,4 with the pyruvic acid (Fig. 3.24) 3,4-GlcNAc and Qui2NAc. Unfortunately, it was not possible elucidate how these sugars were organized. One hypothesis is that, in contrast with Mimivirus, the fibrils could be heavily glycosylated with small oligosaccharides and monosaccharides, but further investigation are required.

A preliminary study revealed the sugar composition of *Moumovirus australensis* as GlcNAc, Qui2NAc and diBacNAc, despite the glycans structure is under investigation.

It is made clear that *Mimivirus*, *Megavirus chilensis* and *Moumovirus australensis* have a different fibrils sugars composition that could be related with the different length and thickness of the fibrils. This difference opened several questions:

1) Is the sugar fibrils composition conserved inside the same linage? (Chapter 4)
2) The only sugar conserved along the three linages is the GlcNAc. What is the role of the GlcNAc in the virus biology? (Chapter 4);
3) What are the implications of these different types of glycosylation?

The first two questions will be discussed in the next chapter (chapter 4).

About the last question, a consequence of the different fibrils glycosylation could be the different ability of these giant DNA viruses to be infected by the virophage, small double-stranded DNA viruses. The first virophage to be discovered was Sputnik in association with Mamavirus (lineage A) that infects Acanthamoeba cells (La Scola et al., 2008). Since that time, dozens of virophages have been characterized. It is evident that the virophages are able to attach the fibrils of the giant DNA viruses, despite not all the virophages are able to infect the viruses of the three linages. Indeed, Sputnik can infect all lineages, while Zamilon virophage (Gaia et al., 2014) can infect only the lineage B and C. A hypothesis could be that the different

composition in sugars impacts on the possibility of the giant viruses to be infected with the virophages, but further studies are necessary.

Chapter 4

Lineage specificities of the genes involved in the *Megavirinae* fibrils formation

The chemical characterization of the *Megavirinae* fibrils revealed a high diversity between the glycosylation patterns of the three lineages.

Indeed, the fibrils of Mimivirus, Megavirus chilensis and Moumouvirus australensis are decorated with different types of sugars: N-acetylglucosamine, a sugar common to all domain of the life, the very rare sugars such as viosamine (*Mimivirus*), N-acetylrhamnosamine (*Megavirus chiliensis*), N-acetylquinovosamine (*Megavirus chilensis* and *Moumouvius australensis*) and N,N'-diacetylbacillosamine (*Moumouvirus australensis*).The picture has become even more complex due to the presence of additional sugar modifications such as acetylation, methylation and pyruvylation. These findings raise two questions:

1) Does the *Megavirinae* have all the genes for sugar production and modification?
2) Is the fibrils glycosylation conserved within the same lineage?

To address these questions and to get a global view on the *Megavirinae* fibrils glycosylation, we scrutinized their genome in search of enzymes possibly involved in the synthesis of such glycans. We complemented this study by the in depth comparative analysis of the candidate proteins.

The general approach used to attribute a specific function to the genes was to compare translated sequences with the closest annotated protein sequences, most often of bacterial origin, using multiple alignment on the Expresso Server (Armougom et al., 2006).

We also used tblastn to assess the presence, and the level of conservation, of all the proteins possibly involved in the fibrils glycosylation pathway for all members of

Megavirinae from the 3 clades. We restricted the study to the ones for which we had the complete genome sequence. The results are presented as conservation heatmaps. The identification of the enzymes that could explain the *in vivo* data is a prerequisite to demonstrate that giant DNA viruses are autonomous for the glycosylation of the fibrils surrounding their capsids. Moreover, the understanding of the different glycosylation gene cluster could also shed light on the origin of the glycosylation machinery in the *Megavirinae*.

4.1 Extension of the nine-gene cluster of *Mimivirus*

It was previously reported that a nine genes cluster in *Mimivirus* (Piacente et al., 2012) genome was involved in glycans formation, as discussed in Chapter 2.
The complete elucidation of the glycan structures of *Mimivirus* fibrils makes the ground for proposing precise functions for the proteins encoded by the gene cluster. Yet, additional enzymes are needed to explain the structures of the polysaccharides. This information provided the starting point to scrutinize the genome in search of these missing genes.
For instance, the characterization of *Mimivirus* fibrils revealed that the GlcNAc had a pyruvate linked to O-4 and O-6, that the viosamine was acetylated on the amino function 4 and methylated on O-2, yet the genes encoding for the corresponding enzymes were not identified. The N-Terminal domain of L142 (inside the nine-gene cluster) was predicted as N-acetyltransferase and its function was demonstrated *in vitro* as discussed in Chapter 5. The gene encoding the L143 protein, also inside the cluster, was predicted as a polysaccharide pyruvyltransferase. It is the best candidate for the modification of the GlcNAc of *Mimivirus* poly-1 (Fig. 3.8). To explain the additional methylation, the Mimivirus genome was scrutinized in search of possible methyltransferases.

R132 (outside of the gene cluster) is a good candidate as it is absent from the Mimivirus M4 mutant devoid of fibrils. Otherwise, this gene is very well conserved in other *Mimiviruses* of the linage A, such as *Mimivirus Bombay*, *Niemeyer virus*, *Hirudovirus* strain Sangsue and *Mamavirus*.

It is worth noting that S-adenosylmethionine-dependent methyltransferase enzymes (MTs) are characterized by low sequence identity. Indeed, the residues involved in the binding of the S-adenosylmethionine (SAM) are poorly conserved and the region involved in substrate binding is highly variable due to the diversity of substrates of the enzymes. Despite this lack of sequence identity, all MTs exhibit a highly conserved fold made of a seven-stranded β sheet edged by three helices on each side (Martin and McMillan, 2002). The secondary structure prediction using psipred (Jones, 1999) supported that alternation of β-strands and helices in R132, from G-47 to K-221 (not shown). The closest homologs of R132 are in the bacterial world, its homolog in *Rizhobiales* shares 39% identity covering 77% of the entire query (WP11257168.1), in *Lelliottia* 35% identity for 87% over the entire query (WP10214719.1) and in *Pantoea dispersa* 35% identity for 85% of the entire query (WP107702876.1). The multiple-alignment of R132 and some of its orthologs in *Mimiviruses* of lineage A with the C-Terminal catalytic domain of MycE (residues 161-399), a SAM and metal-dependent methyltransferase responsible for the methylation in O-2 of the 6-deoxyallose of the mycinamicins (potent antibiotics from *Micromonospora griseorubida*)(Akey et al., 2011), confirmed the conservation of all the residues responsible for the catalytic activity and the metal dependency (Fig. 4.1). Interestingly, viosamine (4-amino-4,6-dideoxy-D-glucose) is very similar to the 6-deoxy-allose. Indeed, the 6-deoxy-allose is an epimer in C-3 of the six-deoxy-glucose and there is an additional amino function in C-4 in the viosamine. The 6-deoxyallose docks in the chair conformation with the 2- and 3-OH, coordinating the active site magnesium along with two aspartate carboxylates D275 and D304 (Akey et al., 2011). They correspond to D74 and D115, respectively in R132, suggesting

that a similar coordination could occur. In addition, the histidine residue (H278) involved in the deprotonation of the 2-OH is also conserved and correspond to H-118 (Fig.2.4.1).

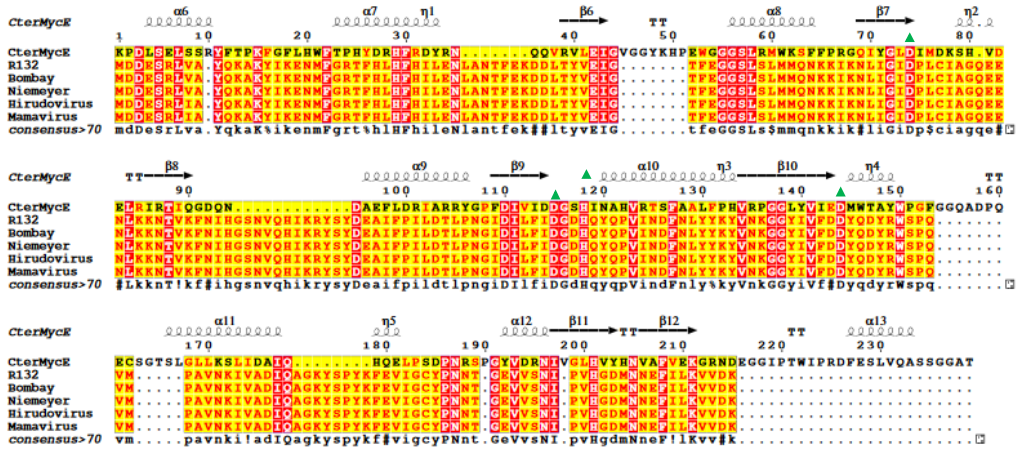


Fig. 4.1 Multiple alignment of the C-Ter of MycE (PDB 3SSN), R132 of *Mimivirus* (YP_003986624), its homolog in *Mimivirus Bombay* (AMZ02580.1), Niemeyer virus (ACR8363.1), *Hirudovirus* strain Sangsue (AHA45742.1). The green triangles indicate the residues involved in catalytic activity. The D75 and D115 coordinate the magnesium and H118 is responsible for the catalysis. The multiple alignment was performed with Expresso Server (Armougom et al., 2006) and the figure was done with ESPript (Gouet et al., 2003).

A model of R132 built by using Swiss model (Schwede et al., 2003) was superimposed on the C-terminal domain of MycE using Coot (Emsley et al., 2010), leading to a RMSD (root-mean-square deviation based on Cα superimposition) of 2.5 Å over 135 amino-acids, indicating that the average distance between the backbone of the superimposed proteins was good. In the model, the catalytic residues were properly positioned relative to the reference structure.

Finally, the level of expression of R132 is comparable to those of the nine genes in the cluster. These data strengthen the idea that the R132 gene product could be responsible for the methylation at O-2 of the N-acetylviosamine. *In vitro* validation will be required to definitively demonstrate the molecular function of R132.

We thus propose to extend the Mimivirus gene cluster to R131, a putative initiation translation factor, also absent in M4 (Boyer et al., 2011) mutant of *Mimivirus*, thus including in this extended cluster also R133 and R134, both encoding hypothetical proteins also lost in M4 (Fig. 4.2).

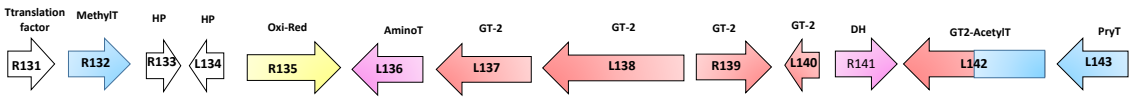


Fig. 4.2 Exstension of the *Mimivirus* cluster involved in glycans formation. The arrows indicate the coding strand. The cluster was extended from R135 to R131, including R132, that could be responsible of the transfer of a methyl group to the 2-OH of the N-acetylviosamine. R131 is a putative translation initiation factor. R133 and L134 encode hypothetical proteins. R135 is an oxido-reductase (in yellow). L136 and R141 are responsible for the UDP-D-viosamine production (in pink). L137, L138, L139 and L140 are putative glycosyltransferase (in red). The N-terminal domain of L142 catalyzes the transfer of an amino group to the 4-amino function of the viosamine (in blue), while the C-terminal domain is a putative glycosyltransferase (in red). L143 is a putative polysaccharide pyruvyltransferase (in blue).

The 13-gene cluster of *Mimivirus* is very well conserved along lineage A, as shown in the heatmap (Fig. 4.3). L780, outside of the gene-cluster, was included in this comparative analysis, because it is responsible along with R141 for the production of UDP-L-Rha. As expected, M4 does not cluster with the wild type Mimivirus due to the loss of most of the gene-cluster. Although M4 kept L780, it was possible to predict that it does not have a functional pathway for the UDP-L-Rha, due to the loss of R141. The major variability in the linage was observed for L142, which is less conserved from Mimivirus fauteuil to M4, as show from the values of the heatmap. This finding could be explained by the splitting of the gene L142 in two genes in M4 (L142a and L142b) and in *Hirudovirus sangsue* (S25 and S26).

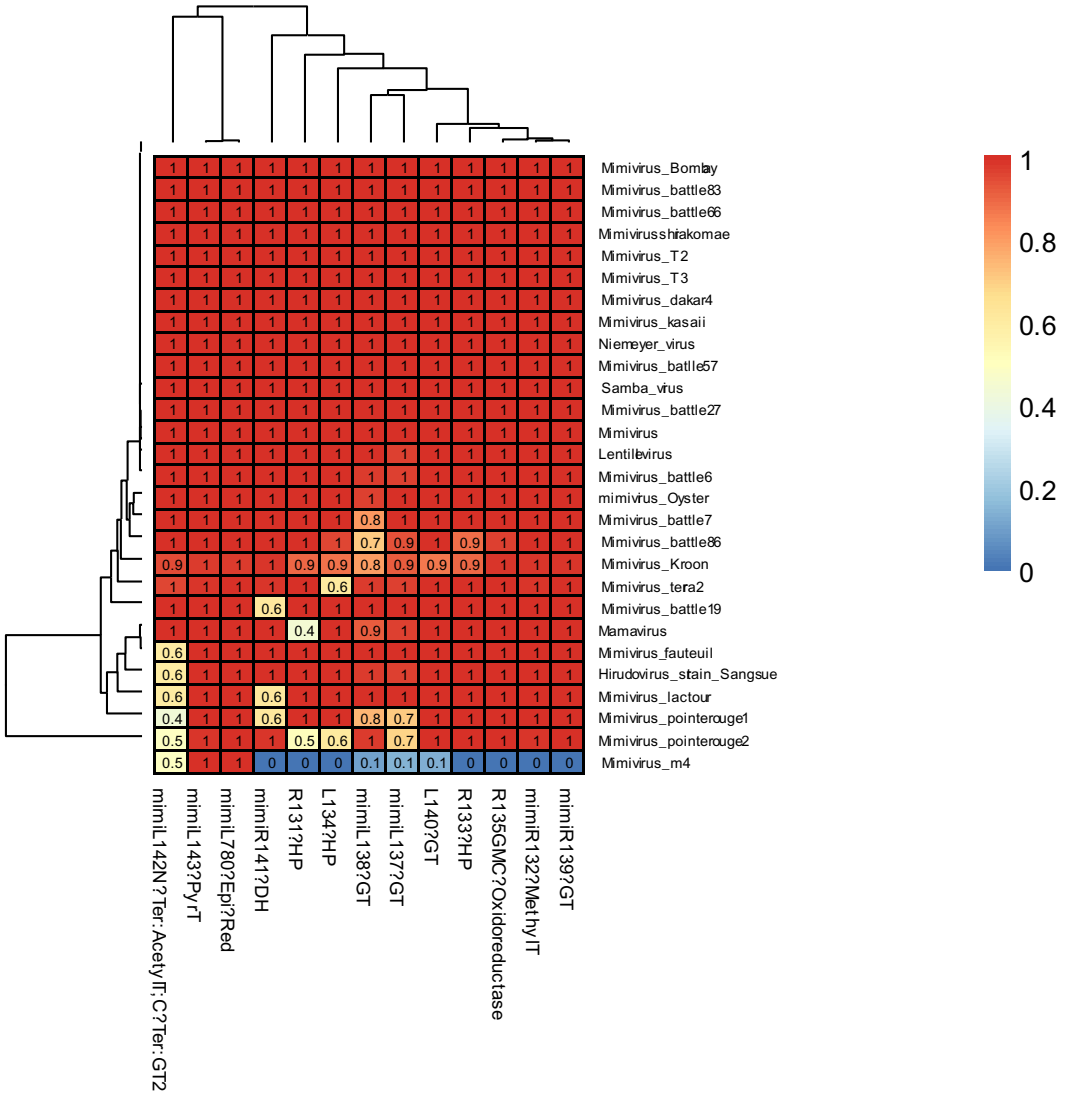


Fig. 4.3 Heatmap based on the level of conservation of the Mimivirus 13-gene cluster in lineage A of the *Mimiviridae*. The conservation score for each protein in the different genomes ranges from 1 to 0 with a color code going from red to blue.

Interestingly, L142 exhibits 41% identity for a query coverage of 40% with mc466 in *Moumouvirus australensis* (MG807320.1) and 38% and 39% of identity with *Tupanvirus soda lake* (KY523104.1) and *Tupanvirus deep ocean* (MF405918.1) for 42 and 41% of query coverage, respectively. *Moumouvirus australensis* and the two strains of *Tupanvirus* only share the acetyltransferase domain with L142, suggesting that some sugars could be acetylated on the amino function. R135, a GMC-oxidoreductase not involved in the glycosylation, is the only protein to be conserved through the three lineages. It is absent from *Tupanviruses*. The *Mimivirus* gene cluster is specific of linage A. The glycosyltransferases, L137, L138, L139, R140 and the C-terminal domain of L142 are specific of the lineage and C-L142 can be responsible of the transfer of the N-acetylviosamine to its acceptor (Chapter 5), while at list two of the other glycosyltransferases of the cluster can be responsible for the two different branching of the rhamnose in the Mimivirus fibrils polysaccharides (Fig. 3.8). Again, *in vitro* experiments are needed to address the specific function of the glycosyltransferases.

4.2 Moumouvirus australensis UDP-D-diNAcBac pathway

Moumouvirus australensis is the first member of linage B for which the sugar composition of the fibrils has been investigated. The chemical analysis revealed the presence of another uncommon acetamido sugar, the N,N'-diacetylbacillosamine (diNAcBac), that was not detected in *Mimivirus* and *Megavirus chiliensis*. The identification of bacillosamine prompted the question about the genes encoding the enzymes for its production. This sugar, as for the viosamine and the rhamnosamine is a very rare sugar present in pathogenic bacteria such as *Campilobacter jejuni*, *Neissera gonorrhoeae*, *Vibrio cholera* and *Pseudomonas reactans* (Morrison and

Imperiali, 2014). The role of this sugar is unknown, but it was proposed that it was not recognized by mammalian host (Perera et al., 2007).

The UDP-diNAcBac phatway has been characterized for *C. jejuni*, for which the catalytic activity and the structure of all enzymes, such as PglF(Riegert et al., 2017), PglE (Badger et al., 2005; Riegert et al., 2015) and PglD (Olivier and Imperiali, 2008), involved in the pathway has been determined. Using the available literature, three candidates genes were identified for *Moumouvirus australensis*: *mc465*, *mc466* and *mc467* (Fig. 4.4).

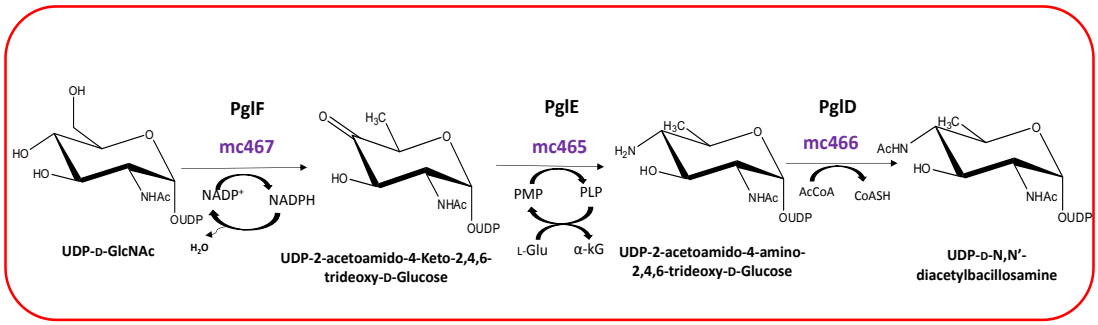


Fig. 4.4 Proposed UDP-diNAcBac synthesis pathway in *Moumouvirus australensis*. The biosynthesis of UDP-D-diNacBac starts from UDP-D-GlcNAc, which could be transformed in UDP-2-acetamido-4-keto-2,4,6-trideoxy-D-Glucose, by *mc467* (PglF in *C. jejuni*) a $NADP^+$ dependent dehydratase. Then there is a transamination reaction, that could be catalyzed by *mc465* (PglE), a PLP dependent aminotransferase, leading to the UDP-2-acetamido-4-amino-2,4,6-trideoxy-D-Glucose. The last step is the acetylation of the 4-amino function of the UDP-2-acetamido-4-amino-2,4,6-trideoxy-D-Glucose, that could be made by *mc466* (PglD) with the formation of UDP-D-N,N'-diacetylbacillosamine (UDP-diNacBac).

Mc467 (AVL94853.1) is a putative dehydratase, which could catalyze the first step of the bacillosamine pathway, namely the dehydratation of UDP-D-GlcNAc to UDP-4-keto-2,4,6-trideoxy-D-glucose (Fig. 4.4). Mc465 is predicted as pyridoxal phosphate aminotransferase that could be responsible of the transfer of an amino group from L-glutamate to the C-4 of the UDP-2-acetamido-4-keto-2,4,6-trideoxy-D-Glucose, leading to UDP-2-acetamido-4-amino-2,4,6-trideoxy-D-glucose. Mc466 was predicted to be a N-acetyltransferase that could catalyze the transfer of an acetyl

group from the AcCoA to the 4-amino function of the UDP-2-acetamido-4-amino-2,4,6-trideoxy-D-glucose with the production of UDP-D-N,N'-diacetylbacillosamine(Fig. 4.4).

Surprisingly, the closest homologs of the Mc467, Mc465 and Mc466 proteins were not identified in the other members of the B lineage, but in the two strains of *Tupanviruses* with whom they share 78 to 64% sequence identity over the entire sequences. Mc467 was the only one to have some homologues in lineage B, with two putative dehydratases in *Moumouvirus* (YP_007354434.1 and AGF85279.1) (69% and 67% sequence identity, respectively) and a putative dehydratase of *Moumouvirus Monve* (AEX62730.1) (53% of sequence identity). In the bacterial world, the best homologs were identified in different strains of *Bacillus*: for Mc467 and Mc466 in *Bacillus cereus* (WP_098337419.1; 33% sequence identity and WP_098710015.1, 45% identity, respectively) and *Bacillus toyonensis* (WP_097999810.1; 29% sequence identity and WP_087948883.1; 54% identity, respectively) with a query coverage of 100%, while the best hit for Mc465 was with WeeJ from *Ralstonia solanacearum* (CUV14353.1, 50% identity, query coverage 97%).

The multiple alignment of Mc467, Mc465 and Mc466 with PglF, PglE and PglD respectively revealed that all catalytic residues were conserved (Fig. 4.5, 4.6 and 4.7). Interestingly, the multiple alignment of Mc467 and its viral orthologs with the C-terminal catalytic domain of PglF, reveals that *Moumouvirus* gp464 protein present a mutation D150N. The mutagenesis of the corresponding D396 to N in PglF was performed and it induces the loss of the enzyme catalytic activity (Riegert et al., 2017). The *gp461* and *gp464* genes could correspond to a duplication event and one of the two proteins lost its catalytic activity. Finally, we observed the same mutation in mvR525 from *Moumouvirus monve*. Since this gene is not duplicated, this virus should not have the functional bacillosamine pathway.

From the *in silico* analyses of mc467, mc465 and m466, it appears that all the catalytic residues are conserved, suggesting that these enzymes are functional and

are responsible for the *Moumouvirus australensis* UDP-D-diNAcBac pathway. These data are strongly supported by the experimental evidence that N,N'-diacetylbacillosamine is a component of *Moumouvirus australensis* fibrils.

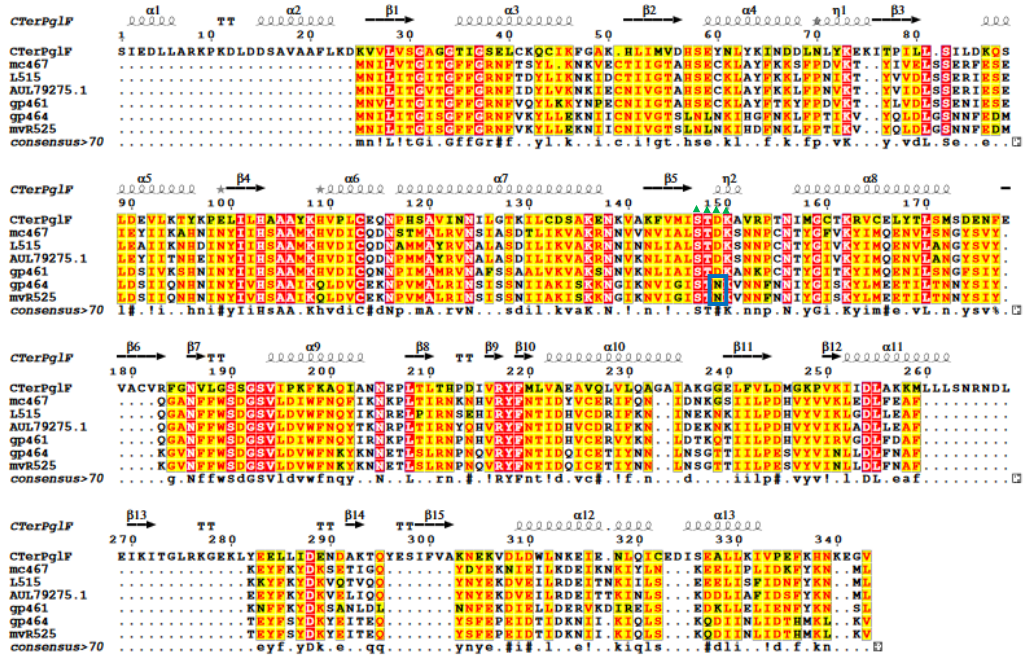


Fig. 4.5 Multiple alignment of the C-Terminal catalytic domain of PglF (PDB 5BJU), *Moumouvirus australensis* mc467 (AVL94853.1), *Tupanvirus* soda lake L515 (AMZ02580.1, 78% identity), *Tupanvirus* deep ocean (ACR8363.1, 78% identity), *Moumouvirus* gp 461 (YP_007354434.1) and gp464 (AGF85279.1), and *Moumouvirus monve*, mvR525 (AEX62730.1). The green triangles indicate the residues of the catalytic site. The blue rectangle indicates the mutation D150N in gp464 and mvR525. The multiple alignment was performed with Expresso Server (Armougom et al., 2006) and the picture was done with ESPript (Gouet et al., 2003).

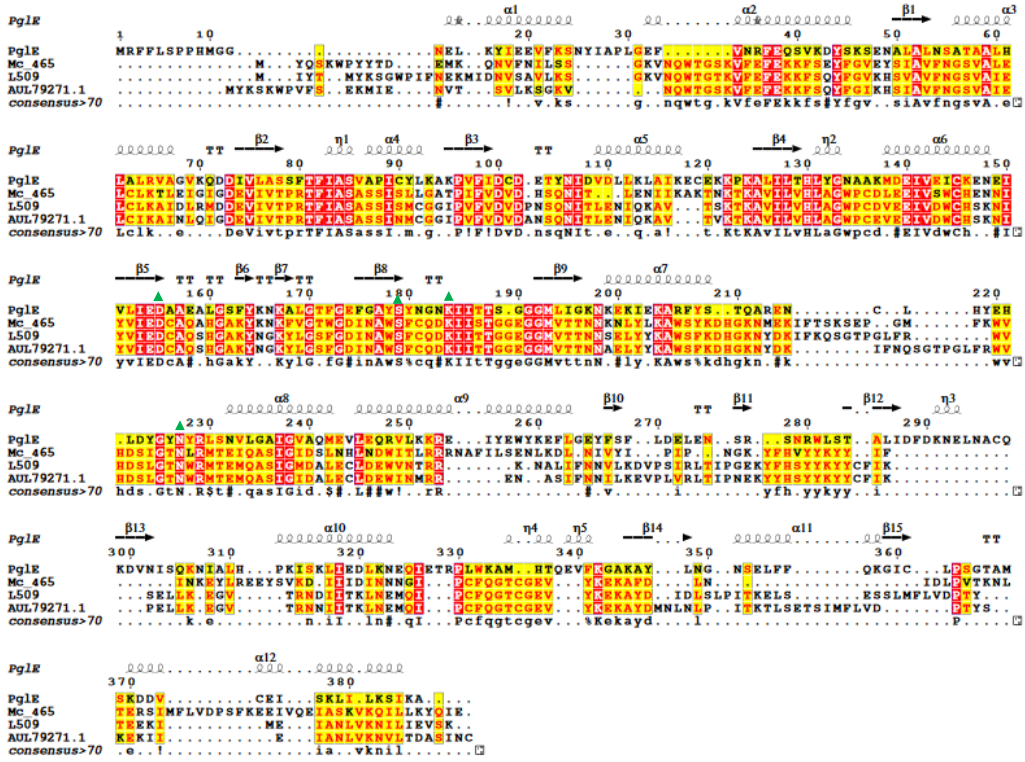


Fig. 4.6 Multiple alignment of the PglE (PDB 1O61), mc465 (AVL94851.1) of *Moumouvirus australensis*, its homolog in *Tupanvirus* soda lake L509 (AUL77977.1, 67% identity), *Tupanvirus* deep ocean (AUL79271.1, 67% identity). The green triangles mark the residues involved in the catalytic activity. The multiple alignment was performed with Expresso Server (Armougom et al., 2006) and the picture was done with ESPript (Gouet et al., 2003).

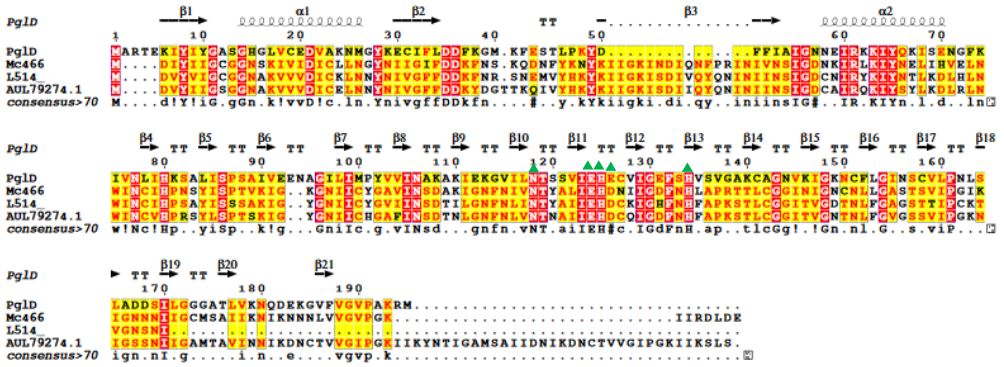


Fig. 4.7 Multiple alignment of PglD (PDB 3BSS), mc466 (AVL94852.1) of *Moumouvirus australensis*, its homolog in *Tupanvirus* soda lake L514 (AUL77982.1), *Tupanvirus* deep ocean (AUL79274.1). The green triangles indicate the residues involved in the catalytic activity(Olivier and Imperiali, 2008). The multiple alignment was produced using Expresso Server (Armougom et al., 2006) and the picture was done with ESPript (Gouet et al., 2003).

4.3 Identification of a fourteen-gene cluster in *Moumouvirus australensis*

Moumouvirus australensis fibrils are decorated by N-acetylglucosamine and N-acetylquinovosamine in addition to the N,N'-di-acetylbacillosamine.

Consequently, the genome was scrutinized in search of all the genes responsible for sugar production. The genes possibly encoding for the enzymes involved in UDP-D-GlcNAc production were not organized in a cluster. They were identified as *mc652*, *mc514* and *mc192*, based on the published data on *Mimivirus* (Piacente et al., 2014b). Indeed, the multiple alignment of the *Moumouvirus* protein sequences with that of *Mimivirus* revealed that all the catalytic residues were conserved (data not shown). However, a complex cluster composed of 14 genes was identified as encoding enzymes possibly involved in quinovosamine and bacillosamine production (Fig. 4.8). This cluster encompasses sugar nucleotide metabolism (*mc458*, *mc459*, *mc460*, *mc465*, *mc466*, *mc467*), a putative glycosyltransferase (*mc468*), sugar modifying enzymes (*mc457*, *mc461*), two papain-like (*mc462*, *mc464*) and a hypothetical

protein (*mc463*). At the boundaries of the cluster, there is a gene encoding for a helicase (*mc456*) and a thioredoxin-like enzyme (*mc469*).

Based on the published cluster identified in *Megavirus chilensis* (Piacente et al., 2014a), the genes involved in the quinovosamine production were easily identified as *mc458*, *mc459* and *mc460*. In contrast to *Mimivirus* and *Megavirus*, the gene encoding for the polysaccharide pyruvyltransferase was absent. Instead two genes involved in the sulfate metabolism were identified, *mc457* (a putative adenylyl-sulfate kinase) and *mc461* (a putative sulfotransferase), suggesting the presence of sugar modified by sulfate group. The possible candidates for the bacillosamine production were identified as *mc465*, *mc466* and *mc467*, as discussed in the previous paragraph.

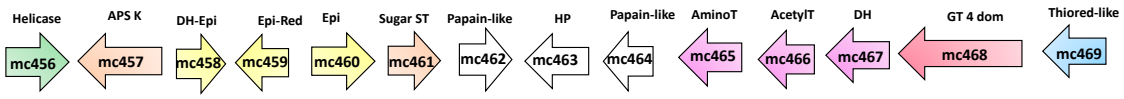


Fig. 4.8 *Moumouvirus australensis* fourteen-gene cluster involved in glycans production. The arrows indicate the coding strand. At the two edges of the cluster there are a helicase (*mc456*) and a thioredoxine-like enzyme (*mc469*). The orange arrows represent the two genes involved in sulfate metabolism (*mc457* and *mc461*). The yellow arrows indicate the three candidates responsible for the UDP-L-quinovosamine synthesis. The whith arrows represent two papain-like proteins and a hypothetical protein. The pink arrows indicate the three candidates for the UDP-D-N,N'-diacetylbacillosamine production.

Surprisingly, the *Moumouvirus australensis* gene cluster is unique among the *Megavirinae,* as shown from the heatmap (Fig. 4.9). It appears clear that the other members of the linage B do not have this cluster gene. Indeed, they only share the first enzyme (mc467) for the bacillosamine production, but not the others. In addition, the evidences of duplication gene event and mutation in mc467 as for *Moumouvirus* and *Moumouvius monve*, strongly suggest that this pathway is going to be completely lost from the lineage B. Astonishing, the new strains of *Tupanvirus*

exhibit all the genes for the bacillosamine production and the multiple alignment with the protein sequences of bacterial origin, disclosed that the catalytic site is conserved, as discussed previously. Yet, the two strains of *Tupanvirus* and *Moumouvirus australensis* share the enzyme mc457 that is a putative adenylyl-sulfate kinase (APKS). Despite, mc461 a putative sugar-sulfotransferase is absent in *Tupanvirus*, suggesting that maybe this type of modification can be absent, but further investigations are required. The presence of sugar modified with a sulfate group has to be demonstrated experimentally also for *Moumouvirus australensis*. *Moumouvirus australensis* is the unique member of the lineage B to share with the lineage C the biosynthetic pathway for the N-acetylquinovosamine, as show from the heatmap (Fig. 4.9.) and confirmed from the experimental data.

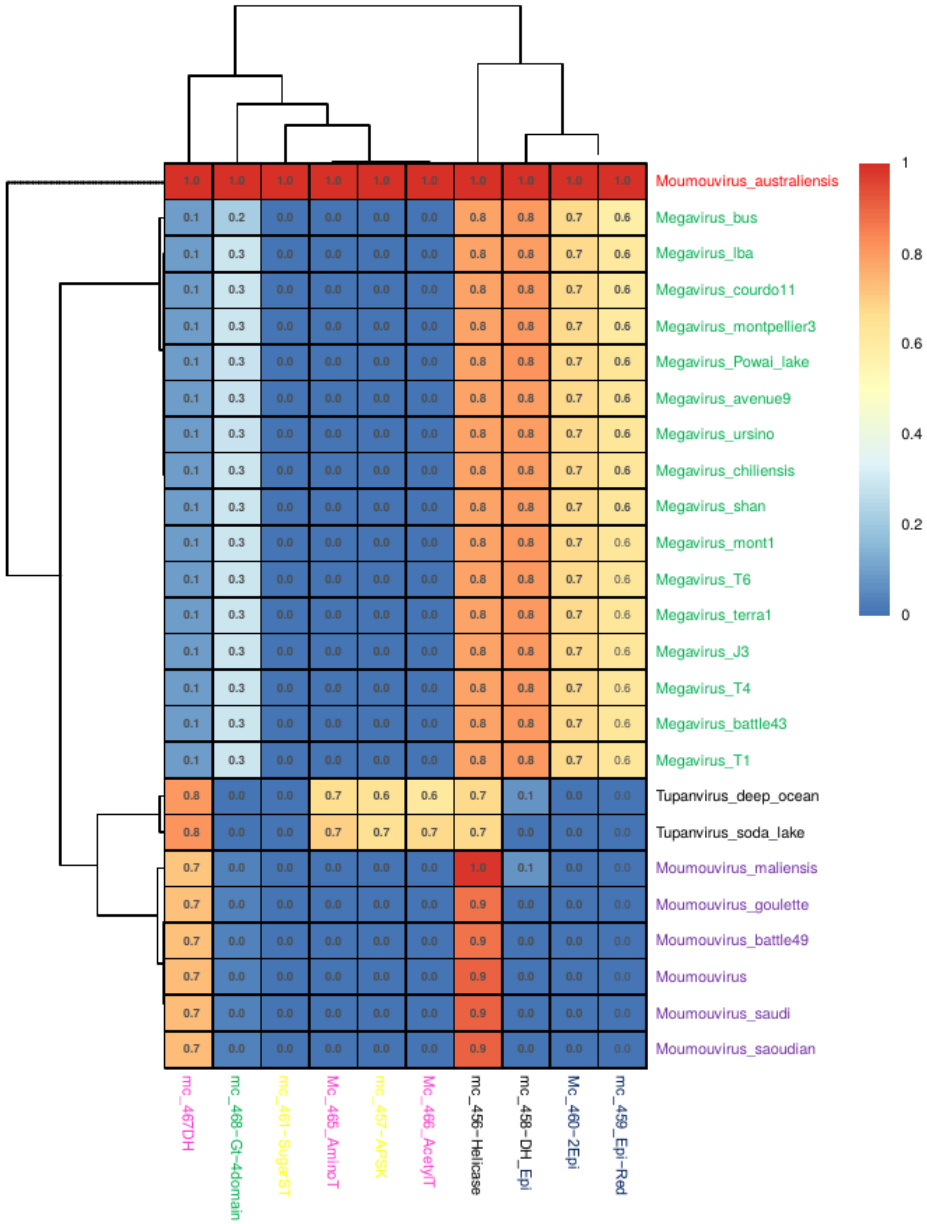


Fig. 4.9 Heat map of the conservation level of the 14-gene cluster of *Moumouvirus australensis* in the *Megavirinae* and *Tupanviruses*. The score of conservation for each protein in each genome ranges from 1 to 0 with a color code going from red to blue. Members of lineage C are in green, lineage B in purple and *Tupanvirus* in black. The enzymes involved in the UDP-L-Qui2NAc are in blue, while those for UDP-D-diNacBac are in pink. The glycosyltransferase mc468is in green and. the enzymes of the sulfate metabolism in yellow. The helicase mc465 that is used as control of the level of the conservation is in black.

Inside the gene cluster there is a glycosyltransferase, mc468 presenting four GT-domains and no homologues in lineage B. Interestingly, there is a low level of conservation with mg539 presenting 3 GT-domains. A preliminary analysis revealed that mc468 and mg539 share 40% identity for a protein sequence coverage of 59%, suggesting that at least one GT-domain could be shared between *Moumouvirus australensis* and *Megavirus chiliensis*. Further *in vitro* studies will be required to identify the GT-domain they have in common and its precise molecular function. The elucidation of the *Megavirus chiliensis* and *Moumouvirus australensis* glycans composition and structure is the prerequisite to address the specific function of all GT domains. Interestingly, the gene cluster of *Moumouvirus australensis* is not conserved in the B lineage suggesting there could be differences in glycan composition for members of a same lineage.

Finally, according to the heatmap (Fig. 4.9) there was a common ancestor between *Moumouvirus australensis* and members of lineage C. *Moumouvirus australensis* appears to have diverged prior the divergence between the B and C lineages and inside the B lineage there is a clear subgrouping between the *Moumouviruses* and the *Tupanviruses*. Definitively, *Moumouvirus australensis* glycosylation cluster is unique.

4.4The N-acetylglucosamine is conserved along all the Mimivirinae family and Tupanvirus

From the fibrils sugar composition analysis, the N-acetylglucosamine is the only sugar well conserved between *Mimivirus*, *Megavirus* chilensis and *Moumouvirus australensis*. In addition, GlcNAc was modified by pyruvic acid as evidenced by the NMR analysis of *Mimivirus* and *Megavirus chilensis*, while it seems that this modification did not occur in *Moumouvirus australensis*. *In vitro* studies revealed

the presence of a functional pathway for the production of UDP-D-GlcNAc in *Mimivirus*. Instead, the pyruvilation of the glucosamine was evidenced from the compositional analysis of the natural fibrils, therefore L143 was proposed to be responsible for the attachment of pyruvic acid as a ketal. Consequently, the Mimivirus proteins involved in glucosamine production and modification were used to evaluate the conservation of this pathway among the *Megavirinae* and *Tupanviruses.* This analysis revealed that the GlcNAc is very well conserved, as shown by the heatmap (Fig. 4.10).

Interestingly, M4, the mutant of *Mimivirus* devoid of fibrils, exhibits all the genes involved in GlcNAc production, including its modification with pyruvic acid, leading to two different hypotheses:

1) M4 has a non-functional pathway for UDP-D-GlcNAc;
2) M4 is able to produce this sugar and to modify it, even in absence of the fibrils

The only way to get a clear answer would be to investigate the M4 viral particles composition.

Looking at the level of conservation of L143, it appears that the modification of the GlcNAc with the pyruvic acid is less conserved among the three lineages. Indeed, the pyruvilation is conserved in lineages A (in blue) and C (in green), while it is absent in lineage B (in purple) and in *Tupanviruses* (in black). In addition, inside lineage C, three strains of *Megavirus* (*Megavirus montpellier 3*, *Megavirus courdo 11*, *Megavirus bus*) present a low conservation of L143 (values of 0.2 against 0.4 for the other strains, Fig. 4.10.).

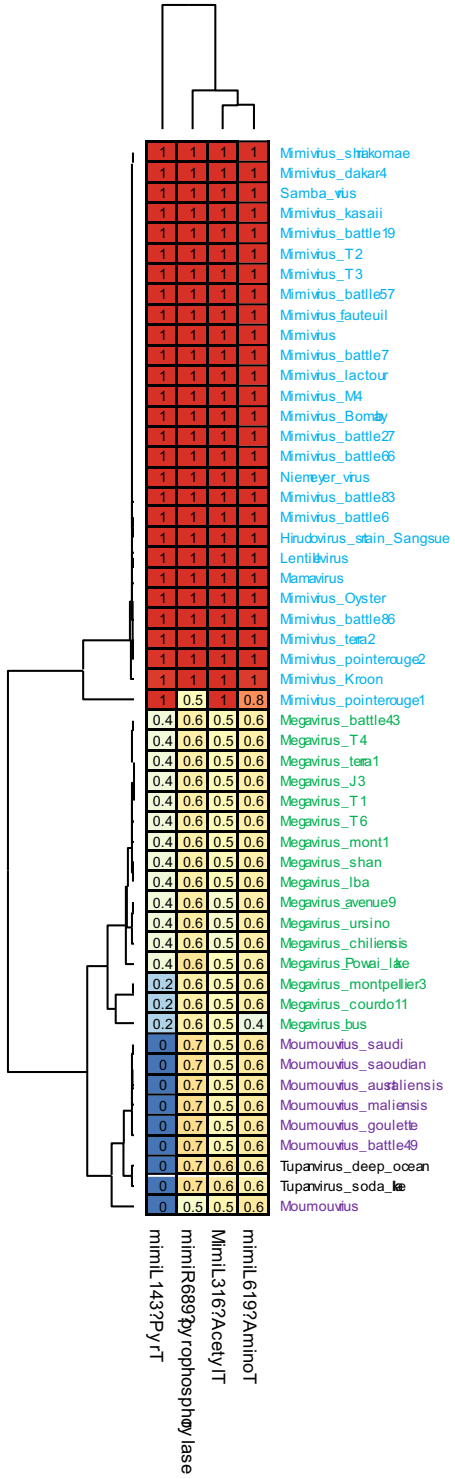


Fig. 4.10 Heat map of the conservation level of all the enzymes involved in the UDP-D-GlcNAc production and modification in the *Megavirinae* family and *Tupanvirus*. The score of each protein in each genome is show from the value ranging from 1 to 0 and the color, going from red to blue. In blue the lineage A, in green the lineage C, in purple the lineage B and Tupanvirus in black.

The pyruviltransferase of *Mimivirus* (L143) exhibits 46% of identity with that of *Megavirus chiliensis* (mg538) along the entire protein sequence. This supports the experimental data that revealed a different type of linkage for the pyruvic acid between Mimivirus (pyruvic acid linked in O-4 and O-6) and Megavirus (pyruvic acid linked in O-3 and O-4). Surprisingly, the heatmap tree revealed that *Moumouvirus*, the prototype of the linage B, does not cluster with the other *Moumouviruses* for the UDP-D-GlcNAc pathway. Similarly, *Mimivirus pointerouge1* does not cluster with the other *Mimiviruses*. These findings could be explained by the lower level of conservation of the pyrophosphorylase R689 compared to the other members of the same lineage (value of 0.5 for Mimivirus pointerouge1 instead of 1 and 0.5 for *Moumouvirus* instead of 0.7).

Even if N-acetylglucosamine is present in the amoeba host, all giant DNA viruses encode their own proteins for this sugar. The N-acetylglucosamine is precursor of several six-deoxy-aminosugar, as the N-acetylrhamnosamine, the N-acetylquinovosamine and the N,N'-di-acetylbacillosamine, which compose the fibrils. This finding strongly suggests that the GlcNAc biosynthesis allowed the viruses to have their own supply of nucleotide sugars, decreasing a lot their dependency on their host.

4.5 Conclusion and discussion

The experimental data highlighted a very high diversity in the fibrils glycosylation of the *Megavirinae*. I demonstrated that this diversity is the consequence of very divergent glycosylation genes cluster. Indeed, Mimivirus has a 13-gene cluster that is specific of lineage A. This cluster was previously identified as a nine-gene cluster, and in this thesis, it was extended to be in accordance with the glycan structures observed *in vivo*. In addition, the complete elucidation of fibrils glycans structures allowed attributing specific functions to some of the genes present in the original

nine-gene cluster. For instance, L143 is a polysaccharide pyruviltransferase that could be responsible of the glucosamine modification of the poly_1 connected to Mimivirus fibrils (Fig. 3.8). Yet, the N-terminal domain of L142 was identified as an N-acetyltransferase that could be involved in the acetylation of the four amino function of the viosamine, while R132 was predicted as a putative methyltransferase able to transfer a methyl group to the O-2 of the viosamine.

A 14-gene cluster involved in fibrils glycosylation was identified in *Moumouvirus australensis*, representing the first glycosylation cluster to be identified in lineage B. In contrast to the *Mimivirus* gene cluster, this cluster is not conserved along the B lineage, but is unique of *Moumouvirus australensis*. Interestingly, *Moumouvirus australensis* shares with the C lineage the UDP-L-quinovosamine pathway and with the two strains of *Tupanviruses* the biosynthesis of the UDP-D-N,N'-diacetylbacillosamine. The only sugar very well conserved among the *Megavirinae* and *Tupanviruses* is the N-acetylglucosamine. Consequentially, the genome was scrutinized in search of putative glycosyltransferase well conserved among the *Megavirinae* that could be responsible for the GlcNAc polymerization. Preliminary results identified *Mimivirus* L373 and L192 as good candidates, but *in vitro* experiments will be needed to confirm these data. The production of this sugar, also found in the amoeba host, could represent a strategy to decrease the virus dependency from the host resources during the infection process, that finally leads to the production of about ~1000 viral particle in each amoeba cell.

To conclude, the fibrils glycosylation is lineage specific and *Moumouvirus australensis* represents an exception in the B lineage because it retained the ability to assemble fibrils carrying more complex types of sugars compared to the other viruses.

Finally, the identification of all these genes involved in sugar production, sugar modification and glycosyltransferase, strongly supports the concept that *Megavirinae* are autonomous for their fibrils glycosylation.

Chapter 5

Validation in vitro of Mimivirus L142 function

A combination of the carbohydrate chemistry with bioinformatic is the winning strategies to discover the viral glycan related genes (Chapter 4), for which the function must be validate in vitro. In this thesis, the function of the gene L142 was investigated in order to complete the pathway of the viosamine, a rare sugar composing the Mimivirus fibrils. Preliminary bioinformatic analysis revealed that L142 could encode for a bifunctional enzyme. Indeed, the N-terminal domain was attributed as N-acetyltransferase and the C-terminal as a putative glycosyltransferase. The investigation of the activity and structure of L142 is a prerequisite to understand how the viosamine acetylation occur and how the viosamine is transferred to its acceptor in the Mimivirus glycans fibrils.

5.1 The Rare Sugar *N*-acetylated Viosamine is a Major Component of Mimivirus Fibers

This work has been published on JBC (doi: 10.1074/jbc.M117.783217) and the full article is in the appendix A. Here is reported a summary of this work.

The rare sugar viosamine, identified only in some bacteria species, is always acylated on amino group. In order to verify this status in Mimivirus glycans, we adopted a dual strategy. We scrutinized the 9-gene cluster localized at 5'end of Mimivirus genome, possibly devoted to glycan production, in search of a candidate for Vio acetylation and identified the N-terminal domain of L142 protein as a possible acyltransferase. Second, we performed GC-MS and NMR analysis of Mimivirus

glycans. Bioinformatic analysis of the N-terminal region of L142 revealed the presence of hexapeptide repeat motif, typical of the left-handed β-helix (LβH domain), that is found in many acyltransferases. The N-terminal domain of L142 was expressed as GST-fusion protein and the pure protein without tag was used for the enzymatic activity test. The purified N-L142 was incubated with UDP-Vio and Acetyl-CoA and the product formation was monitored by anion exchange HPLC. The product of N-L142 was purified by anion exchange HPLC and solid phase extraction (SPE) for ESI and NMR analysis. ESI-MS revealed the presence of a main ion at 590 m/z, consistent with the expected mass of UDP-D-N-acetylviosamine (UDP-D-VioNAc). The NMR analysis, using both 2D homo- and heteronuclear NMR sequences, allowed to identify the UDP-D-VioNAc. This data obtained in vitro, was confirmed by analyzing the sugar composition of the isolated Mimivirus fibrils. The GC-MS analysis of partially methylated alditol acetates (PMAA) revealed the presence of the VioNAc and in addition indicated that is a terminal sugar. A complete NMR analysis on the isolated fibrils confirmed these results. A combination of the in vitro and in vivo studies demonstrated that Mimivirus L142 gene product acetylates the 4-amino group of UDP-D-Vio. To our knowledge it represents the first virally encoded sugar-N-acetyltransferase.

5.2 Preliminary study of the C-terminal domain of L142

The C-terminal domain of L142 (C-L142) is predicted to be a glycosyltransferase of type 2 (GT-2) in the Carbohydrate-Active enZYmes database (CAZY). The GT-2 can transfer the nucleotide sugar to the acceptor with inversion of the configuration of the anomeric carbon. This finding is in agreement with the experimental data, that disclosed the presence of a β-viosamine in the Mimimivirus polysaccharide, and with the in vitro data that proved the formation of UDP-α-viosamine.

Using BLAST search, best hits of C-terminal-domain of L142 correspond to bacterial proteins in loci responsible for glycan formation. This finding strongly suggested that C-L142 catalyzes VioNAc transfer onto Mimivirus glycans. The only way to prove the function of the C-terminal domain of L142 is the determination of its activity and its structure. Accordingly, the C-terminal domain was express and purified, while the biochemical activity and the structure are under investigation.

5.2.1 Cloning of the the gene corresponding to the C-terminal domain of L142

The portion of *L142* gene corresponding to the C-terminal domain was amplified from Mimivirus DNA by PCR reaction, using standard conditions. The band of the gene C-*L142* was at the MW expected 0,759 Kb, compared to the DNA Ladder (Fig. 5.1), confirming that only the gene of interest was amplified.

Fig. 5.1 Agarose gel 0.8%.
A=DNA Ladder;
B= amplification of C-L142 (0.759 Kb).

The PCR product was inserted by homolog recombination in pETDuet-PB-V7 vector to express the C-terminal domain as 6His-Sumo-tag fusion protein to make the protein more soluble (SUMO) and to facilitate subsequent purification steps (His-Tag). The usage of the 6His-tag is one of the most commonly used method for obtaining fusion proteins that can be readily purified by affinity chromatography (Bornhorst and Falke, 2000). It has been demonstrated that SUMO (small ubiquitin-like modifier protein) as an N-terminal fusion partner enhances functional protein production in prokaryotic and eukaryotic expression systems, improving protein stability and solubility (Peroutka III et al., 2011).

The plasmid with the gene of interest was used to transform Stellar Competent cells (E. coli HST08 strain), which provide high transformation efficiency. The

transformed cells (bacterial colonies) were screened by PCR to evaluate the uptake of the gene. The positive colonies were grown O.N and then the plasmid was extracted, using standard procedures. The plasmid was sequenced to verify the absence of mutation in the gene. Verified the integrity of the gene, the plasmid was used for the protein expression.

5.2.2 Expression and purification of the C-terminal domain of L142

Protein expression was conducted in a mutant strain of *E. coli*, called C41, that is frequently used to overcome the toxicity associated with overexpressing recombinant proteins, using the bacteriophage T7 RNA polymerase expression system (Dumon-Seignovert et al., 2004). The expressed protein was purified on Ni-NTA column. Elution was performed by increasing the imidazole concentration of the buffer (TrisHCl 50 mM, NaCl 300 mM, pH 8.5). The fractions eluted at different imidazole concentration were checked by SDS PAGE (Fig. 5.2).

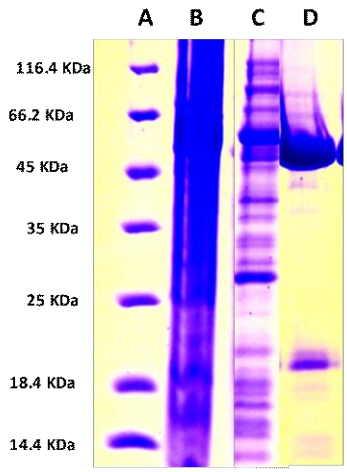


Fig. 5.2 SDS PAGE on the C-Ter-L142 purified on Ni-NTA column (gel 12.5%). A= protein ladder; B= e. coli protein eluted in absence of imidazole; C= protein of E. coli eluted at 25 mM imidazole; C-Ter-L142 eluted at 150 mM imidazole.

The *E. coli* proteins do not have affinity for the Ni-NTA column and were eluted in absence of imidazole (Fig. 5.2 B). A wash at low concentration of imidazole (25 mM) allowed the complete removal of the proteins of *E. coli* (Fig. 5.2 C). The C-terminal of L142 was eluted with 150 mM imidazole (Fig. 5.2 D).

The band at 18.4 KDa could correspond to a little amount of His-SUMO tag that was autocleaved.

The in-house-produced His-tagged human rhinovirus 3C protease was added to the eluted

protein in order to remove the 6His-SUMO-tag from the C-ter domain of L142.

A second run of affinity purification was performed to separate the cleaved C-terminal L142 protein from the uncleaved protein and from the His-tagged protease, which were retained on the Ni–NTA column. Each fraction was checked by SDS PAGE (Fig. 5.3). This analysis revealed that the protein was perfectly cleaved, because it appeared at the expected MW (30.42 KDa). The purified protein was desalted on HiTrap^{TM} 5 ml desalting column on an AKTA explorer 10S FPLC system, using as buffer TrisHCl 10 mM, NaCl 100 mM at pH 8.5. Then an aliquot of the protein was concentrated at 1 mg/ml protein concentration and analyzed by gel filtration chromatography. The remaining protein was further concentrated to 5 mg/ml and it was used for preliminary crystallography and activity test.

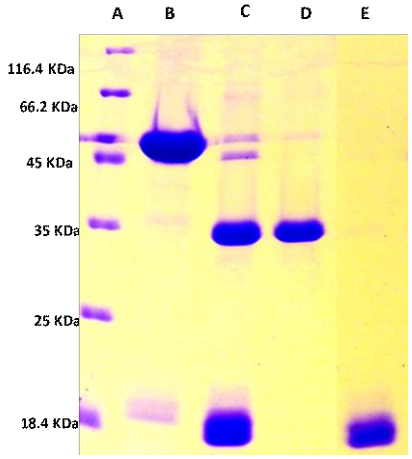


Fig. 5.3 SDS PAGE on the C-Ter-L142 purified on Ni-NTA column afteer clevage (gel 12.5%). A= protein ladder; The band at 45 kDa of the ladder is double due to the migration; B= C-Ter-L142 before clevage; C= C-Ter-L142 clevead before purification; D= C-Ter-L142 clevead eluted in absence of imidazole; E.= 6His-SUMO-tag eluted at 150 mM imidazole.

5.2.3 Gel filtration of the C-terminal domain of L142

To evaluate the oligomeric state of the C-Terminal domain of L142, a gel filtration was performeded on a small aliquot of the protein (1 mg/ml). The chromatographic profile revealed the presence of two peaks, indicating that the C-terminal domain was present in two forms under the used conditions (see materials and methods) (Fig. 5.4). The fractions corresponding to the two peaks were checked on polyacrylamide gel, confirming that both peak were the protein of interest (does not shown).

Comparison between the elution volume of the two peaks with that of proteins of known molecular weight, disclosed that the first peak (Elution time ~ 51 min) was the dimeric form of C-L142, while the second peak (Elution time ~58 min) was C-L142 in the monomeric form (Tab. 10). The not perfect correspondence between the MW and the oligomeric state, could be explained by a not globular shape of the C-Ter of L142 (Tab 10).

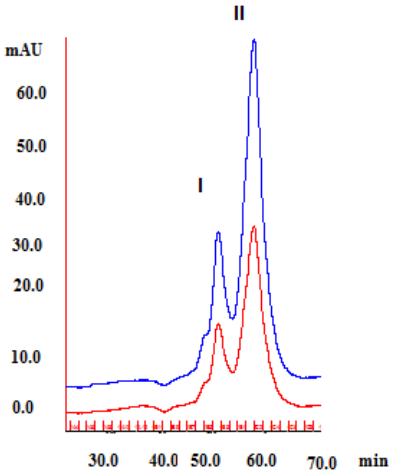


Fig. 2.43 Gel filtration on the C-L142. The presence of two peaks (I and II) indicated that the protein was present in two different oligomeric state. In blu absorbance at 280 nm; in red absorbance at 260 nm.

Proteins	MW	Kav
Chimotripsinogeno	25000	0,70
Ovalbumin	43000	0,58
Albumin	75000	0,51
Aldolase	158000	0,40
Peak I C-L142	64800	0,57
Peak II C-L142	20200	0,74

Tab. 10 Molecular weight of the proteins used to calculate the molecular weight of the two peaks of the C-terminal ofer L142. Proteins with molecular weight from 25000 Da to 158000 Da were inject on the SuperdexTM 200 10/300 GL column and were used to construct a calibration curve (MW in function of the Kav, which is the proportion of pores available to the molecule; for molecule totally excluded Kav=0, while for totally included Kav=1) with the following straight-line equation: -2E-06x+0.6996. The molecular weight of the two peaks was determined by solving the straight-line equation. The MW of the C-Ter-L142 is 30.42 KDa, meaning that the first peak corresponds to a dimeric form and the second peak to a monomeric form. The not perfect correspondence of the MW can be related to ta not globular shape of the protein.

5.3 Perspectives

The activity determination of the C-L142 protein is complex, due to the specificity of the glycosyltransferase for the donor (UDP-viosamine or UDP-N-acetylviosamine) and the acceptor (Rhamnose linked to the Glucosamine in the Mimivirus polysaccharide). The first approach will be to perform STD NMR experiments to evaluate the affinity of the protein for the UDP-viosamine and UDP-N-acetylviosamine to establish if the viosamine is transferred to the acceptor prior or after acetylation. Usually, GT-2 transferases are cation dependent, however for C-L142, it was not possible foresee if it depended on a cation and from which type of cation, due to its low homology with any well-known GT-2. Consequentially, future work will take two directions. First, STD NMR experiments will be performed with or without magnesium, while manganese cannot be used because of its paramagnetic

properties. STD-NMR will be used to understand which form of viosamine (free-amino or N-acetylated) is preferred from C-L142.

Second, I will perform the reaction with C-L142, UDP-viosamine (with or without N-acetyl) and any of the two possible cations (Mg and Mn) with rhamnose as possible acceptor and I will analysis via HPLC the formation of the glycosylation product.

Due to the low homology of the C-terminal domain of L142 with other glycosyltransferases, it is not possible to build a model for this protein. Consequentially, crystallographic test will be performed to find the appropriate crystallization condition. Then the protein will be expressed as seleniomethyl protein to facilitate the resolution of the crystal structure.

Along with this work, the silencing of L142 (full length protein) will be tested by using an approach under development at IGS, with the purpose to evaluate the role of this putative GT in the fibrils assemblage.

The success of the silencing will be evaluated by analyzing the glycans of the fibrils, by the consolidated chemical and NMR approach.

Finally, the full-length protein will be expressed and purified to determine its complete structure.

Methods

Section III

Material and methods

Chapter 6

6.1 Production and purification of viruses

Mimivirus, *Megavirus chilensis* and *Moumoumovirus australensis* were produced and purified using the following procedure:

1) *A. castellanii* (Douglas) Neff (American Type Culture Collection 30010TM) cells were cultured in 20 flask of 175 cm^2 with 20 ml of PPYG (1 mL 2% (wt/vol) proteose peptone, 0.1% yeast extract, 100 µM glucose, 4 mM $MgSO_4$, 0.4 mM $CaCl_2$, 50 µM $Fe(NH_4)_2(SO4)_2$, 2.5 mM Na2HPO4, 2.5 mM KH_2PO_4, pH 6.5) medium, supplemented with antibiotics (ampicilline 100 µg/mL and penicillin–streptomycin (Gibco) 100 µg/mL,) at 32°C. ;
2) The cells were cultured until 100% of confluence and infected with the virus at multiplicity of infection (M.O.I) of 0.25 meaning 1 virus for 4 cells. The viruses belong to IGS of Marseille, which was responsible of their discovery. Then the cells were incubated at 32°C for two days;
3) The infection was followed by light microscopy, as discussed in Chapter 3. After two days, most of the cells were lysed and the viral particles were released in the medium;
4) To remove the cellular debris, the medium was centrifuged at 500 g for 10 min, 20°C;
5) The supernatant, containing the viral particles, was centrifuged at 6800 g for 45 min, then the pellet was washed with pure water and centrifuged at 6800 g, 45 min, 20°C;
6) The viral particles were resuspended in CsCl 1.2 density, laid down on a gradual CsCl gradient of 1.3/1.4/1.5 densities and centrifuged 20 h at 100000 g;
7) The white band, corresponding to the viral particles, was recovered by a syringe and washed 3 times with pure water;
8) The purified viral particles were checked by light microscopy (ZEISS);

9) The concentration of the viral production was mesured with the bio photometer (Eppendorf) at OD 600 nm.

For each virus was done 6 productions in order to have about $3x10^{12}$ viral particles.

6.2 Fibrils isolation

For all the viruses the fibrils were detached from the viral capsid using the following procedure. ~1.5×10^{11} viral particles (~2,5 ml of viral suspension in water) were treated with 50 mM DTT (20 ml) and the reaction was conducted at 100°C for 2 h under stirring (100 rpm).

The supernatant containing the fibrils was separated from the viral particles by centrifugation (10 min., 4650 g, room temperature), the solid was washed twice with water (1 ml) and the supernatants pooled. The supernatant was dialyzed against water (3500 Da) to remove the DTT. An aliquot of the supernatant and of the viral particle was used for TEM analysis and the remaining part was freeze-dried. The yields are reported in Tab. 11.

For Mimivirus the fibrils were removed also through an incubation with DTT 2h at room temperature and not at 100°C.

Virus	Treatment	Fibrils (mg)
Mimivirus	50 mM DTT 2h 100°C	10
Mimivirus	50 mM DTT 2h RT	5
Megavirus	50 mM DTT 2h 100 °C	8
M. australensis	50 mM DTT 2h 100°C	13

Tab.11 Yields of the fibrils obtained after the DTT treatment in different conditions.

6.3 Transmission electron microscopy

The untreated virus, the fibrils and the defibered virus obtained after DTT treatment in different conditions, were analyzed by TEM. All the samples were fixed in glutaraldehyde 2.5% in water for 1h at room temperature. After the fixing the samples were centrifuged at 5000 g for 10 min and the pellet were washed twice with water. In order to highlight the structure of the fibrils, was performed the negative staining plus methyl cellulose (M6385 Sigma), using as contrast uranil acetate 2% in water. Then all the samples were observed on TEM TECNAI.

6.4 Sugar composition of the fibrils

Monosaccharide composition analysis (acetylated methyl glycoside), substitution pattern (partially methylated and acetylated alditol) were performed as reported (De Castro et al. 2010) (De Castro et al., 2010). GC-MS analyses were performed with an Agilent instrument (GC instrument Agilent 6850 coupled to MS Agilent 5973), equipped with a SPB-5 capillary column (Supelco, 30 m × 0.25 i.d., flow rate, 0.8 mL min−1) and 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 temperature program used for the analyses was the following: 150°C for 5 min, 150 → 280°C at 3°C/min, 300°C for 5 min.

6.5 NMR analysis of fibrils glycans

For NMR analysis, 1D and 2D NMR spectra of the different samples, were recorded on a Bruker 600 DRX equipped with a CryoProbe™ at different temperatures (329 or 310 K), according to the sample. Acetone was used as internal standard (^{1}H 2.225

ppm, ^{13}C 31.45 ppm) and 2D spectra (COSY, TOCSY, NOESY, HSQC and HMBC) were acquired by using Bruker software (TopSpin 2.0). Homonuclear experiments were recorded using 512 FIDs of 2048 complex with a different number of scans per FID (ranging between 25 and 40 according to the sample), mixing times of 100 and 200 ms were used for TOCSY and NOESY spectra, respectively. HSQC and HMBC spectra were acquired with 512 FIDs of 2048 complex point, accumulating a different number of scans, according to the sample (ranging between 50 and 100). Spectra were processed and analyzed using a Bruker TopSpin 3 program.

NMR spectra of the purified fractions of Mimivirus were recorded at 310 K and the accurate integration of methyl signals used as reporter groups along with the minimization of the residual water signal was achieved by analyzing the mono-dimensional DOSY spectrum, which was set with little and big DELTA respectively to 1200000 and 0.100000001, the variable gradient was set to 1-100%. The experiments were recorded using 32768 FID and 208 scans.

6.6 Separation of Mimivirus polysaccharides and Megavirus chilensis fibrils glycans by ion exchange chromatography

Independently from the protease treatment of the fibrils, the carbohydrate material (3-10 mg) was purified by ion exchange chromatography using Q-Sepharose fastflow as adsorbent (0.1 ml of gel for 5-10mg of fibrils) and fractions were collected by increasing stepwise NaCl concentration (10 mM, 100 mM, 200 mM, 400 mM, 700 mM and 1M). Each solution was applied for 5 column volumes and eluates with the same ionic strength were desalted on a Biogel P10 column (Flow: 12ml/h) by using mq H_2O as eluent and monitoring the eluate with an on-line refractive index detector (K-2310 Knauer). The carbohydrate material was identified by ^1H NMR analysis.

For Mimivirus, DOSY spectra were recorded to better appreciate the intensity of the anomeric signal close to residual solvent signal and to run an accurate integration of the methyl signals at 1.47 ppm (the reporter group of poly_1) and at 1.24 ppm (inclusive of both poly_1 and poly_2). The proportion between the two species was calculated with the following formula:

$$\frac{poly\_1}{poly\_1}; \frac{poly_2 = [(poly_1 + poly\_2) - poly_1]/2}{poly\_1}$$

For Megavirus the observation of a variation between the signals at 1.47 ppm e 1.3 ppm in the fraction eluted at different NaCl concentration was an indication of the presence of more than one specie. Despite, it was not possible to evaluate the proportion between the different species, as for Mimivirus, due to the lack of information about its polysaccharide structures.

Each protease digestion was performed treating the material (the fibrils or fibrils already digested once) with proteinase K (Sigma P6556-100 MG, CAS 39450-01-6), using 0.5 mg of proteinase for 10 mg of sample in the Digestion Buffer (10 mM Tris, 5 mM NaCl, 1 mM $MgCl_2$, pH 7,5), at 55°C O.N. Then the samples were dialyzed against water (3500 Da) and dried. Yields for the different purification of Mimivirus are reported in Table 3, while those for Megavirus chilensis purification are reported in Tab. 12.

$[NaCl]_{mM}$	Proteinase K digestion (mg)
Starting material	3
100	0.60
200	0.45
400	0.70

Tab. 12. Yields of Megavirus chilensis digested fibrils (3 mg) purified by ion exchange chromatography. Fractions eluted at NaCl concentration higher than 400 mM did not contain carbohydrate material.

6.7 Separation of Mimivirus polysaccharides by RP-HPLC of untreated fibrils

The fibrils (5 mg) were purified by reverse phase C18 HPLC (RP-18, 5 µm, 25 cm x 4,6 mm). For this analysis, 400 µl aliquots of the sample (2 mg/ml solution) were injected onto a reverse phase column (SUPELCO 58858) and eluted with 70% Acetonitrile whit 50 mM ammonium bicarbonate (Solvent B) and 50 mM ammonium bicarbonate (Solvent A) following the gradient showed in Tab. 13. The eluates were monitored by UV absorbance at 206 nm. The peaks of the HPLC were analyzed by ^{1}H NMR.

Time	% Solvent B	Flow
0	20	0.800
15	100	0.800
20	100	0.800
21	20	0.800
25	20	0.800

Tab. 13 RP-HPLC gradient used to elute the polysaccharide material. The Solvent B is 70% acetonitrile with 50 mM ammonium bicarbonate and the solvent A is 50 mM ammonium bicarbonate.

6.8 Molecular weight determination of Mimivirus fibrils polysaccharides

The molecular weight of both polysaccharides, was determined by SEC-HPLC (TSK gel G5000 PWXL, 30 cm x 7.8 mm ID). In order to disclose the PM of the poly_1 (eluted in the fraction 400 mM NaCl of double digested fibrils) and poly_2 (eluted in the fraction 100 mM NaCl of digested fibrils), 50 µl (1 mg/ml solution) of each polysaccharide was injected on the TSK column and eluted with 100% of 50

mM ammonium bicarbonate (Flow 0.8 ml/min). The eluate was monitored by refractive index. The column was calibrated with dextrans of known molecular weight (5 KDa, 50 KDa, 150 KDa, 410 KDa, 610 KDa), the data fitted in a linear regression and the molecular weight of poly_1 and poly_2 evaluated accordingly (Section II, Tab. 4).

6.9 Protein Electrophoresis of Mimivirus

SDS PAGE was performed on Mimivirus fibrils (after DTT 50 mM 100°C), on the defibered virus and on the total virus, as control.

The total virus and the defibered virus were lysed with lysis Buffer 1X (40 mM Tris pH 7.5, 60 mM DTT, 2% SDS and antiprotease).

The protein content of Mimivirus fibrils was concentrated using Amicon Ultra centrifugal filter units MWCO 10 kDa to the 10% of the initial volume. The concentration of all samples was determined by Pierce^{TM} 660 nm Protein Assay (Thermo Fisher Scientific) (Tab. 14).

Sample	Concentration µg/ml
Mimivirus total	1759
Defibered virus	1168
Fibrils	1356

Tab. 14 The concentration of the protein content of the Mimivirus total, the defiberd virus (appropriately lysed) and the fibrils was performed by Pierce^{TM} 660 nm Protein Assay.

12 µg of proteins were resuspended in Laemmli sample buffer (2% SDS, 10% glycerol, 5% 2-mercatoethanol, 60 mM Tis-Cl pH 6.8, 0.01% bromophenol blue) in ratio 1:1, heated at 95°C for 5 min and loaded on Polyacrylamide gel 12, 5%.
The proteins were visualized by two different staining: Coomassie blue (GelCode Blue Stain Reagent; Pierce, Thermo Scientific) and periodic acid-Schiff method (Glycoprotein detection kit; Sigma-Aldrich). Horseradish peroxidase (5 µg) was used as a positive control for glycoprotein detection. To have a better comparison between the two staining, on the same gel each sample was loaded twice to avoid differences related to migration.

6.10 Comparative proteomic analysis of Mimivirus

A comparative proteomic analysis was performed on the isolated fibrils, the defibered virus and the total virus (as control) in collaboration with the Biosciences and Biotechnology Institute of Grenoble (BIG). The proteomic analysis was performed through "classical" nanoLC-MS/MS analyses.
The proteomic data were interpreted using ACD tool (Claverie and Ta, 2018), a software developed at IGS (URL: www.igs.cnrs-mrs.fr/acdtool/), which allows the comparison between the different fractions (see Section II, Tab: 5 and 6).

6.11 Depolymerization of the fibrils glycans of Megavirus chilensis

$1.5x10^{11}$ viral particles (dried) underwent solvolysis with TFMS ≥ 99% (Perepelov et al., 2000) (347817 Sigma), under dry condition, at 4°C for 4 h under stirring. The reaction was blocked by neutralizing the sample with ammonia 33% until a pH 7. Then the sample was centrifuged at R.T. for 15 min, at 5000 rpm. The precipitate was washed with water twice and supernatants pooled. The supernatant was dried with the rotavapor, washed with 5 ml of mq water and dried again. This sample was

resuspended in 2 ml of mq water and loaded on Biogel P2 column (Flow: 10 ml/h) by using mq H_2O as eluent and monitoring the eluate with an on-line refractive index detector (K-2310 Knauer). The ^1H NMR analysis identified the carbohydrate material in four fractions, yields are reported in Tab. 15.

Fraction	Yield (mg)
A	0.80
B	0.76
C	1.31
D	4.26

Tab. 15. Yields of the different fractions of the Biogel P2 on the Megavirus viral particles treated with TFMS at 4°C, 4h.

6.12 Reverse phase HPLC on the carbohydrate material obtained after TFMS treatment on Megavirus chilensis viral particles

The fractions C and D of the size exclusion chromatography on the viral particles treated with TFMS (1.3 and 4.2 mg respectively) were purified by reverse phase C18 HPLC (RP-18, 5 μm, 250 mm x 4.6 mm). For this analysis, 300 μl aliquots of the sample (1 mg/0.5 ml solution) were injected onto a reverse phase column (Phenomenex 4602) and eluted with pure water. The eluates were monitored by UV absorbance at 206 nm. The peaks of the HPLC were analyzed by ^1H NMR. The yields are reported in Tab. 16.

Starting sample	Peak HPLC	Yield (mg)
Fraction C	I	0.12
	II	0.5
	III	0.3
Fraction D	I	1.12
	II	0.3
	III	0.6
	IV	0.3
	V	0.15
	VI	0.20
	VII	0.23

Tab. 16 Yields of the different fractions of the HPLC on the Biogel P2 fractions C and D (see tab. 15).

6.13 Multiple alignment of sequences

All the proteins sequences reported in this thesis were recovered by NCBI (https://www.ncbi.nlm.nih.gov/) and the accession number of each sequence is reported in Tab. 17. All the multiple alignments were done using Expresso Server (Armougom et al., 2006) (http://tcoffee.crg.cat/apps/tcoffee/do:expresso) and the pictures were done by ESPript (http://espript.ibcp.fr/ESPript/ESPript/) (Gouet et al., 2003).

Source	Protein	Accession number
Mimivirus	L425	YP_003986929.1
	R132	YP_003986624
Mimivirus Bombay	R132	AMZ02580.1
Niemeyer virus	R132	ACR8363.1
Hirudovirus strain Sangsue	R132	AHA45742.1
Megavirus chilensis	Mg464	YP_004894515.1
Moumououvirus australensis	Mc389	AVL94775.1
	Mc467	AVL94853.1
	Mc465	AVL94851.1
	Mc466	AVL94852.1
Tupanvirus soda lake	L515	AMZ02580.1
	L509	AUL77977.1
	L514	AUL77982.1
Tupanvirus deep ocean	L515	ACR8363.1
	L509	AUL79271.1
	L514	AUL79274.1
Moumouvirus	Gp461	YP_007354434.1
	Gp464	AGF85279.1
Moumouvirus monve	Mv525	AEX62730.1
C. jejeuni	PglF	PDB-5BUJ
C. jejeuni	PglE	PDB-1O61
C. jejeuni	PglD	PDB-3BSS

Tab.17 The accession number of proteins sequence used for the multiple alignment.

6.14 Heat map

The heat map was built using the program AnnaMat, a private software developed at IGS. The construction of the heatmap was based on the following steps:

1) tblastn of N proteins (Proteine file) against M genomes (genomeFile), meaning that each protein sequence (query) is compared to the six-frame translations of the nucleotide genomes sequences;
2) recovery for each protein of the best scores in each genome, with normalization by the size of the protein (score/autoscore);
3) generation of the matrix (N best scores) X (M genomes) and plot of the heatmap (library pheatmap de R in R, with default parameters).

The complete viral geomes and protein sequences used for the Heat map are aviable to the website NCBI.

6.15 Cloning of the gene corresponding to the C-terminal of L142

The portion of L142 gene corresponding to C-terminal domain was amplified from Mimivirus DNA and cloned by homologous recombination using an In-Fusion HD Cloning Kit (Clontech) according to the manufacturer's recommendations.

For PCR amplification a forward primer and a reverse primer, containing a sequence specific for the plasmid (with the restriction site included) and a sequence specific for the gene to amplify were built with the InFusion® Primer design tool (http://bioinfo.clontech.com/infusion/convertPcrPrimersInit.do).

50 ng of DNA was used for the PCR reaction, following the High-Fidelity DNA polymerase (BioLabs) conditions.

The PCR product was inserted by recombination in an in–house-modified pETDuet expression vector (Novagen), called pETDuet-PB-V7. This vector was modified to express the C terminal domain as 6His-Sumo-tag fusion protein.

The gene L142 corresponding to the C-terminal domain was cloned by homologous recombination, in pETSumoDuet-PB-V7 vector, digested by BamHI and NotI, using an In-Fusion HD Cloning Kit (Clontech).

The In-Fusion PCR product was used to transform Stellar Competent cells, using the protocol PT5055-2 (Clontech). After the transformation step, the colonies, were screened by PCR, to verify the uptake of the plasmid with the gene of interest. The positive colonies were grown O.N at 37°C, in LB medium plus ampicillin and the day after the plasmid was extracted using PureLink® Quick Plasmid DNA Miniprep Kits. In order to confirm that this plasmid had the gene of interest, the sample was sequenced by Eurofins Genomics (https://www.eurofinsgenomics.eu/). Once verified the sequence, the plasmid was used for the expression of the protein in a suitable strain of *E. coli*.

6.16 Expression and purification of C-terminal of L142

The expression was performed in *E. coli* C41 transfected cells grown at 310 K in 2YT (Difco). Induction was performed at 290 K using 0.2 mM IPTG, when OD_{600} reached ~0.6. The pellet from a 0.51 overnight culture was resupended in 40 ml 50 mM tris-Hcl, 300 mM NaCl pH 8.5 (buffer A) containing 0,1% triton x-100, 5% glycerol, 0.01 %(w/v) DNAse, 0.01 % (w/v) lysozyme and antiprotease. The total protein was extracted by sonication: three cycles of 3 min. The protein was recovered by 45 min of centrifugation, at 13000g, 4°C. The surnatant, containing the soluble proteins, was separated from the pellet (bacteria component). The surnatant was filtrated and purified on Ni-NTA column, on the bench.

The cleared lysate was applied onto a 1 ml HisPur Ni-NTA column (Pierce) charged whit Ni and equilibrated whit buffer A (TrisHCl 50 mM, NaCl 300 mM, pH 8.5). The column was washed with ten column volumes of buffer A, the colum volume of buffer A containing 25 mM imidazole, ten column volume containing 50 mM imidazole, the column volume containing 150 mM imidazole and ten volume buffer A, 1 M imidazole. Each fraction was checked in polyacrylamide gel 12.5%.

The in-house-produced His-tagged human rhinovirus 3C protease was added to the pooled elution fractions followed by dialysis overnight at 277 K against 50 mM Tris–HCl buffer pH 8.5, 300 mM NaCl, 1 mM DTT.

A second run of affinity purification was performed to separate the cleaved C-terminal L142 protein from uncleaved protein and His-tagged protease, which were retained on the Ni–NTA column. Each fraction was checked on polyacrylamide gel 12.5%. The fraction corresponding to the protein was desalted on HiTrap™ 5 ml desalting column on an AKTA explorer 10S FPLC system (GE Healtheare) at flow rate of 2 ml/min. The column was washed with 2 column volume of the buffer ofTrisHCl 10 mM, NaCl 100 mM, pH 8.5 and the elution was done with six column volume of the same buffer and monitored by UV absorbance at 280 nm. Then the protein was concentrated using Amicon Ultra centrifugal filter units MWCO 10 kDa, until have a final concentration of 5 mg/ ml. The protein was stored at – 80°C.

6.17 Gel filtration of the C-terminal of L142

A small aliquot of the C-Ter domain of of L142 (1 mg/ml) was loaded on Superdex™ 200 10/300 GL column (Column volume: 23.5 ml), on an AKTA explorer 10S FPLC system (GE Healtheare) at flow rate of 0.3 ml/min. The column was washed with one volume column of the buffer (Tris HCl 10 mM at pH 8.5, NaCl 100 mM) and the elution was done with 1.5 volume column of the same buffer. The elution was

monitored by UV absorbance at 280 nm. All the fractions were checked on polyacrylamide gel 12.5 %.

Comparison between the elution volume of the two peaks, corresponding to the two state of the protein, with that of protein at molecular weight know, allowed to determine the two oligomeric state of the C-terminal of L142 protein (Section II, Tab. 10).

References

Abergel, C., Rudinger-Thirion, J., Giegé, R., and Claverie, J.-M. (2007). Virus-encoded aminoacyl-tRNA synthetases: structural and functional characterization of mimivirus TyrRS and MetRS. J. Virol. *81*, 12406–12417.

Abergel, C., Legendre, M., and Claverie, J.-M. (2015). The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. FEMS Microbiol. Rev. *39*, 779–796.

Abrahão, J., Silva, L., Silva, L.S., Khalil, J.Y.B., Rodrigues, R., Arantes, T., Assis, F., Boratto, P., Andrade, M., Kroon, E.G., et al. (2018). Tailed giant Tupanvirus possesses the most complete translational apparatus of the known virosphere. Nat. Commun. *9*, 749.

Aherfi, S., Andreani, J., Baptiste, E., Oumessoum, A., Dornas, F.P., Andrade, A.C. dos S.P., Chabriere, E., Abrahao, J., Levasseur, A., Raoult, D., et al. (2018). A Large Open Pangenome and a Small Core Genome for Giant Pandoraviruses. Front. Microbiol. *9*.

Akey, D.L., Li, S., Konwerski, J.R., Confer, L.A., Bernard, S.M., Anzai, Y., Kato, F., Sherman, D.H., and Smith, J.L. (2011). A New Structural Form in the SAM/Metal-Dependent O-Methyltransferase Family: MycE from the Mycinamicin Biosynthetic Pathway. J. Mol. Biol. *413*, 438–450.

Andrade, A.C. dos S.P., Rodrigues, R.A.L., Oliveira, G.P., Andrade, K.R., Bonjardim, C.A., La Scola, B., Kroon, E.G., and Abrahão, J.S. (2017). Filling Knowledge Gaps for Mimivirus Entry, Uncoating, and Morphogenesis. J. Virol. *91*.

Andreani, J., Aherfi, S., Bou Khalil, J.Y., Di Pinto, F., Bitam, I., Raoult, D., Colson, P., and La Scola, B. (2016). Cedratvirus, a Double-Cork Structured Giant Virus, is a Distant Relative of Pithoviruses. Viruses *8*.

Andreani, J., Khalil, J.Y.B., Baptiste, E., Hasni, I., Michelle, C., Raoult, D., Levasseur, A., and La Scola, B. (2018). Orpheovirus IHUMI-LCC2: A New Virus among the Giant Viruses. Front. Microbiol. *8*.

Antwerpen, M.H., Georgi, E., Zoeller, L., Woelfel, R., Stoecker, K., and Scheid, P. (2015). Whole-genome sequencing of a pandoravirus isolated from keratitis-inducing acanthamoeba. Genome Announc. *3*.

Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V., and Notredame, C. (2006). Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. *34*, W604-608.

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., and Claverie, J.-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc. Natl. Acad. Sci. U. S. A. *108*, 17486–17491.

Azza, S., Cambillau, C., Raoult, D., and Suzan-Monti, M. (2009). Revised Mimivirus major capsid protein sequence reveals intron-containing gene structure and extra domain. BMC Mol. Biol. *10*, 39.

Badger, J., Sauder, J.M., Adams, J.M., Antonysamy, S., Bain, K., Bergseid, M.G., Buchanan, S.G., Buchanan, M.D., Batiyenko, Y., Christopher, J.A., et al. (2005). Structural analysis of a set of proteins resulting from a bacterial genomics project. Proteins Struct. Funct. Bioinforma. *60*, 787–796.

Bagdonaite, I., and Wandall, H.H. (2018). Global aspects of viral glycosylation. Glycobiology *28*, 443–467.

Bajrai, L.H., de Assis, F.L., Azhar, E.I., Jardot, P., Robert, C., Abrahão, J., Raoult, D., and La Scola, B. (2016). Saudi Moumouvirus, the First Group B Mimivirus Isolated from Asia. Front. Microbiol. *7*, 2029.

Bock, K., and Pedersen, C. (1983). Carbon-13 Nuclear Magnetic Resonance Spectroscopy of Monosaccharides. In Advances in Carbohydrate Chemistry and Biochemistry, (Elsevier), pp. 27–66.

Bornhorst, J.A., and Falke, J.J. (2000). Purification of proteins using polyhistidine affinity tags. Methods Enzymol. *326*, 245–254.

Boughalmi, M., Saadi, H., Pagnier, I., Colson, P., Fournous, G., Raoult, D., and La Scola, B. (2013). High-throughput isolation of giant viruses of the Mimiviridae and Marseilleviridae families in the Tunisian environment. Environ. Microbiol. *15*, 2000–2007.

Boyer, M., Azza, S., Barrassi, L., Klose, T., Campocasso, A., Pagnier, I., Fournous, G., Borg, A., Robert, C., Zhang, X., et al. (2011). Mimivirus shows dramatic genome reduction after intraamoebal culture. Proc. Natl. Acad. Sci. U. S. A. *108*, 10296–10301.

Carter, G.C., Law, M., Hollinshead, M., and Smith, G.L. (2005). Entry of the vaccinia virus intracellular mature virion and its interactions with glycosaminoglycans. J. Gen. Virol. *86*, 1279–1290.

Cherrier, M.V., Kostyuchenko, V.A., Xiao, C., Bowman, V.D., Battisti, A.J., Yan, X., Chipman, P.R., Baker, T.S., Van Etten, J.L., and Rossmann, M.G. (2009). An icosahedral algal virus has a complex unique vertex decorated by a spike. Proc. Natl. Acad. Sci. U. S. A. *106*, 11085–11089.

Ciucanu, I., and Kerek, F. (1984). A simple and rapid method for the permethylation of carbohydrates. Carbohydr. Res. *131*, 209–217.

Claverie, J.-M., and Abergel, C. (2010). Mimivirus: the emerging paradox of quasi-autonomous viruses. Trends Genet. TIG *26*, 431–437.

Claverie, J.-M., and Abergel, C. (2018). Mimiviridae: An Expanding Family of Highly Diverse Large dsDNA Viruses Infecting a Wide Phylogenetic Range of Aquatic Eukaryotes. Viruses *10*, 506.

Claverie, J.-M., and Ta, T.N. (2018). ACDtool: a web-server for the generic analysis of large data sets of counts. Bioinforma. Oxf. Engl.

Colson, P., Fournous, G., Diene, S.M., and Raoult, D. (2013a). Codon usage, amino acid usage, transfer RNA and amino-acyl-tRNA synthetases in Mimiviruses. Intervirology *56*, 364–375.

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., et al. (2013b). “Megavirales”, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch. Virol. *158*, 2517–2521.

De Castro, C., Carannante, A., Lanzetta, R., Nunziata, R., Piscopo, V., and Parrilli, M. (2004). Elucidation of two O-chain structures from the lipopolysaccharide fraction of Agrobacterium tumefaciens F/1. Carbohydr. Res. *339*, 2451–2455.

De Castro, C., Parrilli, M., Holst, O., and Molinaro, A. (2010). Microbe-Associated Molecular Patterns in Innate Immunity. In Methods in Enzymology, (Elsevier), pp. 89–115.

De Castro, C., Molinaro, A., Piacente, F., Gurnon, J.R., Sturiale, L., Palmigiano, A., Lanzetta, R., Parrilli, M., Garozzo, D., Tonetti, M.G., et al. (2013). Structure of N-linked oligosaccharides attached to chlorovirus PBCV-1 major capsid protein reveals unusual class of complex N-glycans. Proc. Natl. Acad. Sci. U. S. A. *110*, 13956–13960.

Diesend, J., Kruse, J., Hagedorn, M., and Hammann, C. (2017). Amoebae, Giant Viruses, and Virophages Make Up a Complex, Multilayered Threesome. Front. Cell. Infect. Microbiol. *7*, 527.

Dumon-Seignovert, L., Cariot, G., and Vuillard, L. (2004). The toxicity of recombinant proteins in Escherichia coli: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3). Protein Expr. Purif. *37*, 203–206.

Emsley, P., Lohkamp, B., Scott, W.G., and Cowtan, K. (2010). Features and development of *Coot*. Acta Crystallogr. D Biol. Crystallogr. *66*, 486–501.

Fischer, M.G., Allen, M.J., Wilson, W.H., and Suttle, C.A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. Proc. Natl. Acad. Sci. U. S. A. *107*, 19508–19513.

Forterre, P. (2017). Viruses in the 21st Century: From the Curiosity-Driven Discovery of Giant Viruses to New Concepts and Definition of Life. Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am. *65*, S74–S79.

Gaia, M., Benamar, S., Boughalmi, M., Pagnier, I., Croce, O., Colson, P., Raoult, D., and La Scola, B. (2014). Zamilon, a Novel Virophage with Mimiviridae Host Specificity. PLoS ONE *9*, e94923.

Gallot-Lavallée, L., Blanc, G., and Claverie, J.-M. (2017). Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. J. Virol. *91*.

Garegg, P.J., Jansson, P.-E., Lindberg, B., Lindh, F., Lönngren, J., Kvarnström, I., and Nimmich, W. (1980). Configuration of the acetal carbon atom of pyruvic acid acetals in some bacterial polysaccharides. Carbohydr. Res. *78*, 127–132.

Gouet, P., Robert, X., and Courcelle, E. (2003). ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins. Nucleic Acids Res. *31*, 3320–3323.

Iyer, L.M., Aravind, L., and Koonin, E.V. (2001). Common Origin of Four Diverse Families of Large Eukaryotic DNA Viruses. J. Virol. *75*, 11720–11734.

Jeffers, S.A., Sanders, D.A., and Sanchez, A. (2002). Covalent Modifications of the Ebola Virus Glycoprotein. J. Virol. *76*, 12463–12472.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne. J. Mol. Biol. *292*, 195–202.

Klose, T., Herbst, D.A., Zhu, H., Max, J.P., Kenttämaa, H.I., and Rossmann, M.G. (2015). A Mimivirus Enzyme that Participates in Viral Entry. Structure *23*, 1058–1065.

Knirel', I.A., and Kochetkov, N.K. (1994). [Structure of lipopolysaccharides from gram-negative bacteria. III. Structure of O-specific polysaccharides]. Biokhimiia Mosc. Russ. *59*, 1784–1851.

Korn, E.D., and Weisman, R.A. (1967). Phagocytosis of latex beads by Acanthamoeba. II. Electron microscopic study of the initial events. J. Cell Biol. *34*, 219–227.

Kuznetsov, Y.G., and McPherson, A. (2011). Nano-fibers produced by viral infection of amoeba visualized by atomic force microscopy. Biopolymers *95*, 234–239.

Kuznetsov, Y.G., Gurnon, J.R., Van Etten, J.L., and McPherson, A. (2005). Atomic force microscopy investigation of a chlorella virus, PBCV-1. J. Struct. Biol. *149*, 256–263.

Kuznetsov, Y.G., Xiao, C., Sun, S., Raoult, D., Rossmann, M., and McPherson, A. (2010). Atomic force microscopy investigation of the giant mimivirus. Virology *404*, 127–137.

Kuznetsov, Y.G., Klose, T., Rossmann, M., and McPherson, A. (2013a). Morphogenesis of Mimivirus and Its Viral Factories: an Atomic Force Microscopy Study of Infected Cells. J. Virol. *87*, 11200–11213.

Kuznetsov, Y.G., Klose, T., Rossmann, M., and McPherson, A. (2013b). Morphogenesis of Mimivirus and Its Viral Factories: an Atomic Force Microscopy Study of Infected Cells. J. Virol. *87*, 11200–11213.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E., et al. (2008). The virophage as a unique parasite of the giant mimivirus. Nature *455*, 100–104.

La Scola, B., Campocasso, A., N'Dong, R., Fournous, G., Barrassi, L., Flaudrops, C., and Raoult, D. (2010). Tentative Characterization of New Environmental Giant Viruses by MALDI-TOF Mass Spectrometry. Intervirology *53*, 344–353.

Legendre, M., Audic, S., Poirot, O., Hingamp, P., Seltzer, V., Byrne, D., Lartigue, A., Lescot, M., Bernadac, A., Poulain, J., et al. (2010). mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. Genome Res. *20*, 664–674.

Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., et al. (2014). Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc. Natl. Acad. Sci. U. S. A. *111*, 4274–4279.

Legendre, M., Lartigue, A., Bertaux, L., Jeudy, S., Bartoli, J., Lescot, M., Alempic, J.-M., Ramus, C., Bruley, C., Labadie, K., et al. (2015). In-depth study of *Mollivirus sibericum* , a new 30,000-y-old giant virus infecting *Acanthamoeba*. Proc. Natl. Acad. Sci. *112*, E5327–E5335.

Legendre, M., Fabre, E., Poirot, O., Jeudy, S., Lartigue, A., Alempic, J.-M., Beucher, L., Philippe, N., Bertaux, L., Christo-Foroux, E., et al. (2018). Diversity and evolution of the emerging Pandoraviridae family. Nat. Commun. *9*, 2285.

Levasseur, A., Andreani, J., Delerce, J., Bou Khalil, J., Robert, C., La Scola, B., and Raoult, D. (2016). Comparison of a Modern and Fossil *Pithovirus* Reveals Its Genetic Conservation and Evolution. Genome Biol. Evol. *8*, 2333–2339.

Lönngren, J., and Svensson, S. (1974). Mass Spectrometry in Structural Analysis of Natural Carbohydrates. In Advances in Carbohydrate Chemistry and Biochemistry, (Elsevier), pp. 41–106.

Luther, K.B., Hülsmeier, A.J., Schegg, B., Deuber, S.A., Raoult, D., and Hennet, T. (2011). Mimivirus collagen is modified by bifunctional lysyl hydroxylase and glycosyltransferase enzyme. J. Biol. Chem. *286*, 43701–43709.

Lwoff, A. (1957). The Concept of Virus. Microbiology *17*, 239–253.

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.

Martin, J.L., and McMillan, F.M. (2002). SAM (dependent) I AM: the S-adenosylmethionine-dependent methyltransferase fold. Curr. Opin. Struct. Biol. *12*, 783–793.

Moniruzzaman, M., LeCleir, G.R., Brown, C.M., Gobler, C.J., Bidle, K.D., Wilson, W.H., and Wilhelm, S.W. (2014). Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. Virology *466–467*, 60–70.

Moreira, D., and Brochier-Armanet, C. (2008). Giant viruses, giant chimeras: The multiple evolutionary histories of Mimivirus genes. BMC Evol. Biol. *8*, 12.

Morrison, M.J., and Imperiali, B. (2014). The Renaissance of Bacillosamine and Its Derivatives: Pathway Characterization and Implications in Pathogenicity. Biochemistry *53*, 624–638.

Mutsafi, Y., Zauberman, N., Sabanay, I., and Minsky, A. (2010). Vaccinia-like cytoplasmic replication of the giant Mimivirus. Proc. Natl. Acad. Sci. U. S. A. *107*, 5978–5982.

Mutsafi, Y., Shimoni, E., Shimon, A., and Minsky, A. (2013). Membrane Assembly during the Infection Cycle of the Giant Mimivirus. PLoS Pathog. *9*, e1003367.

Mutsafi, Y., Fridmann-Sirkis, Y., Milrot, E., Hevroni, L., and Minsky, A. (2014). Infection cycles of large DNA viruses: Emerging themes and underlying questions. Virology *466–467*, 3–14.

Nandhagopal, N., Simpson, A.A., Gurnon, J.R., Yan, X., Baker, T.S., Graves, M.V., Van Etten, J.L., and Rossmann, M.G. (2002). The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. Proc. Natl. Acad. Sci. U. S. A. *99*, 14758–14763.

Netherton, C.L., and Wileman, T. (2011). Virus factories, double membrane vesicles and viroplasm generated in animal cells. Curr. Opin. Virol. *1*, 381–387.

Okamoto, K., Miyazaki, N., Song, C., Maia, F.R.N.C., Reddy, H.K.N., Abergel, C., Claverie, J.-M., Hajdu, J., Svenda, M., and Murata, K. (2017). Structural variability and complexity of the giant Pithovirus sibericum particle revealed by high-voltage electron cryo-tomography and energy-filtered electron cryo-microscopy. Sci. Rep. *7*.

Olivier, N.B., and Imperiali, B. (2008). Crystal Structure and Catalytic Mechanism of PglD from *Campylobacter jejuni*. J. Biol. Chem. *283*, 27937–27946.

Pabst, M., Chang, M., Stadlmann, J., and Altmann, F. (2012). Glycan profiles of the 27 N-glycosylation sites of the HIV envelope protein CN54gp140. Biol. Chem. *393*.

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

Perepelov, A.V., Senchenkova, S.N., Shashkov, A.S., Komandrova, N.A., Tomshich, S.V., Shevchenko, L.S., Knirel, Y.A., and Kochetkov, N.K. (2000). First application of triflic acid for selective cleavage of glycosidic linkages in structural studies of a bacterial polysaccharide from Pseudoalteromonas sp. KMM 634 †. J. Chem. Soc. Perkin 1 363–366.

Perera, V.N., Nachamkin, I., Ung, H., Patterson, J.H., McConville, M.J., Coloe, P.J., and Fry, B.N. (2007). Molecular mimicry in *Campylobacter jejuni* : role of the lipo-oligosaccharide core oligosaccharide in inducing anti-ganglioside antibodies. FEMS Immunol. Med. Microbiol. *50*, 27–36.

Peroutka III, R.J., Orcutt, S.J., Strickler, J.E., and Butt, T.R. (2011). SUMO Fusion Technology for Enhanced Protein Expression and Purification in Prokaryotes and Eukaryotes. In Heterologous Gene Expression in E.Coli, T.C. Evans, and M.-Q. Xu, eds. (Totowa, NJ: Humana Press), pp. 15–30.

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., et al. (2013). Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. Science *341*, 281–286.

Piacente, F., Marin, M., Molinaro, A., De Castro, C., Seltzer, V., Salis, A., Damonte, G., Bernardi, C., Claverie, J.-M., Abergel, C., et al. (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.

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

Piacente, F., De Castro, C., Jeudy, S., Molinaro, A., Salis, A., Damonte, G., Bernardi, C., Abergel, C., and Tonetti, M.G. (2014b). Giant Virus *Megavirus chilensis* Encodes the Biosynthetic Pathway for Uncommon Acetamido Sugars. J. Biol. Chem. *289*, 24428–24439.

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.

Raoult, D. (2004). The 1.2-Megabase Genome Sequence of Mimivirus. Science *306*, 1344–1350.

Renesto, P., Abergel, C., Decloquement, P., Moinier, D., Azza, S., Ogata, H., Fourquet, P., Gorvel, J.-P., and Claverie, J.-M. (2006). Mimivirus Giant Particles Incorporate a Large Fraction of Anonymous and Unique Gene Products. J. Virol. *80*, 11678–11685.

Riegert, A.S., Young, N.M., Watson, D.C., Thoden, J.B., and Holden, H.M. (2015). Structure of the external aldimine form of PglE, an aminotransferase required for *N* , *N* ’-diacetylbacillosamine biosynthesis: Structure of the Aminotransferase PglE. Protein Sci. *24*, 1609–1616.

Riegert, A.S., Thoden, J.B., Schoenhofen, I.C., Watson, D.C., Young, N.M., Tipton, P.A., and Holden, H.M. (2017). Structural and Biochemical Investigation of PglF from *Campylobacter jejuni* Reveals a New Mechanism for a Member of the Short Chain Dehydrogenase/Reductase Superfamily. Biochemistry *56*, 6030–6040.

Rodrigues, R.A.L., dos Santos Silva, L.K., Dornas, F.P., de Oliveira, D.B., Magalhães, T.F.F., Santos, D.A., Costa, A.O., de Macêdo Farias, L., Magalhães, P.P., Bonjardim, C.A., et al. (2015). Mimivirus Fibrils Are Important for Viral Attachment to the Microbial World by a Diverse Glycoside Interaction Repertoire. J. Virol. *89*, 11812–11819.

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack, K.E., Noordeloos, A.A.M., Brussaard, C.P.D., et al. (2013). Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proc. Natl. Acad. Sci. *110*, 10800–10805.

Schulz, F., Yutin, N., Ivanova, N.N., Ortega, D.R., Lee, T.K., Vierheilig, J., Daims, H., Horn, M., Wagner, M., Jensen, G.J., et al. (2017). Giant viruses with an expanded complement of translation system components. Science *356*, 82–85.

Schvarcz, C.R., and Steward, G.F. (2018). A giant virus infecting green algae encodes key fermentation genes. Virology *518*, 423–433.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. (2003). SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. *31*, 3381–3385.

Scola, B.L. (2003). A Giant Virus in Amoebae. Science *299*, 2033–2033.

Seibert, M.M., Ekeberg, T., Maia, F.R.N.C., Svenda, M., Andreasson, J., Jönsson, O., Odić, D., Iwan, B., Rocker, A., Westphal, D., et al. (2011). Single mimivirus particles intercepted and imaged with an X-ray laser. Nature *470*, 78–81.

Sharma, V., Colson, P., Pontarotti, P., and Raoult, D. (2016). Mimivirus inaugurated in the 21st century the beginning of a reclassification of viruses. Curr. Opin. Microbiol. *31*, 16–24.

Suzan-Monti, M., Scola, B.L., Barrassi, L., Espinosa, L., and Raoult, D. (2007). Ultrastructural Characterization of the Giant Volcano-like Virus Factory of Acanthamoeba polyphaga Mimivirus. PLoS ONE *2*, e328.

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*, 21559–21565.

Toukach, P.V. (2011). Bacterial Carbohydrate Structure Database 3: Principles and Realization. J. Chem. Inf. Model. *51*, 159–170.

Van Etten, J.L., and Meints, R.H. (1999). Giant Viruses Infecting Algae. Annu. Rev. Microbiol. *53*, 447–494.

Van Etten, J.L., Meints, R.H., Kuczmarski, D., Burbank, D.E., and Lee, K. (1982). Viruses of symbiotic Chlorella-like algae isolated from Paramecium bursaria and Hydra viridis. Proc. Natl. Acad. Sci. *79*, 3867–3871.

Wilson, W.H., Gilg, I.C., Duarte, A., and Ogata, H. (2014). Development of DNA mismatch repair gene, MutS, as a diagnostic marker for detection and phylogenetic analysis of algal Megaviruses. Virology *466–467*, 123–128.

Xiao, C., Chipman, P.R., Battisti, A.J., Bowman, V.D., Renesto, P., Raoult, D., and Rossmann, M.G. (2005). Cryo-electron Microscopy of the Giant Mimivirus. J. Mol. Biol. *353*, 493–496.

Xiao, C., Kuznetsov, Y.G., Sun, S., Hafenstein, S.L., Kostyuchenko, V.A., Chipman, P.R., Suzan-Monti, M., Raoult, D., McPherson, A., and Rossmann, M.G. (2009). Structural Studies of the Giant Mimivirus. PLoS Biol. *7*, e1000092.

Xiao, C., Fischer, M.G., Bolotaulo, D.M., Ulloa-Rondeau, N., Avila, G.A., and Suttle, C.A. (2017). Cryo-EM reconstruction of the Cafeteria roenbergensis virus capsid suggests novel assembly pathway for giant viruses. Sci. Rep. *7*.

Yan, X., Chipman, P.R., Castberg, T., Bratbak, G., and Baker, T.S. (2005). The Marine Algal Virus PpV01 Has an Icosahedral Capsid with T=219 Quasisymmetry. J. Virol. *79*, 9236–9243.

Yan, X., Yu, Z., Zhang, P., Battisti, A.J., Holdaway, H.A., Chipman, P.R., Bajaj, C., Bergoin, M., Rossmann, M.G., and Baker, T.S. (2009). The Capsid Proteins of a Large, Icosahedral dsDNA Virus. J. Mol. Biol. *385*, 1287–1299.

Yutin, N., Wolf, Y.I., Raoult, D., and Koonin, E.V. (2009). Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. Virol. J. *6*, 223.

Zauberman, N., Mutsafi, Y., Halevy, D.B., Shimoni, E., Klein, E., Xiao, C., Sun, S., and Minsky, A. (2008). Distinct DNA Exit and Packaging Portals in the Virus Acanthamoeba polyphaga mimivirus. PLoS Biol. *6*, e114.

Zhang, X., Xiang, Y., Dunigan, D.D., Klose, T., Chipman, P.R., Van Etten, J.L., and Rossmann, M.G. (2011). Three-dimensional structure and function of the Paramecium bursaria chlorella virus capsid. Proc. Natl. Acad. Sci. *108*, 14837–14842.

Appendix A

The rare sugar N-acetylated viosamine is a major component of Mimivirus fibers

Francesco Piacente[1], Cristina De Castro[2], Sandra Jeudy[3], Matteo Gaglianone[1], Maria Elena Laugieri[1], Anna Notaro[4,3], Annalisa Salis[1], Gianluca Damonte[1], Chantal Abergel[3] and Michela G. Tonetti[1*]

[1]From the Department of Experimental Medicine and Center of Excellence for Biomedical Research, University of Genova, Genova, Italy; [2]Department of Agricultural Sciences, University of Napoli, Italy; [3]Aix-Marseille Univ, Centre National de la Recherche Scientifique, Information Génomique et Structurale, -UMR 7256, IMM FR3479, 13288 Marseille Cedex 9, France;
[4]Department of Chemical Sciences, University of Napoli, Italy

Running title: *A Mimivirus sugar acetyltransferase*

*To whom correspondences should be addressed: Michela Tonetti, Department of Experimental Medicine, University of Genova, Viale Benedetto XV, 1 - 16132 Genova, Italy, Telephone: +39-010-3538151; FAX: +39-010-35338162; E-mail: tonetti@unige.it;

Keywords: Mimivirus, viosamine, acetylCoA, glycosylation, mass spectrometry, NMR, structural model

ABSTRACT

The giant virus Mimivirus encodes an autonomous glycosylation system, which is thought to be responsible for the formation of complex and unusual glycans composing the fibers surrounding its icosahedral capsid, including the dideoxyhexose viosamine. Previous studies have identified a gene cluster in the virus genome, encoding enzymes involved in nucleotide-sugar production and glycan formation, but the functional characterization of these enzymes and the full identification of the glycans found in viral fibers remains incomplete. As viosamine is typically found in acylated forms, we suspected that one of the genes might encode an acyltransferase, providing directions to our functional annotations. Bioinformatic analyses indicated that the L142 protein contains a N-terminal acyltransferase domain and a predicted C-terminal glycosyltransferase. Sequence analysis of the structural model of the L142 N-terminal domain indicated significant homology with some characterized sugar acetyltransferases that modify the C-4 amino group in the bacillosamine or perosamine

biosynthetic pathways. Using mass spectrometry and NMR analyses, we confirmed that L142 N-terminal domain is a sugar acetyltransferase, catalyzing the transfer of an acetyl moiety from acetyl-CoA to the C-4 amino group of UDP-D-viosamine. Loss of activity for a H136A mutant was consistent with the predicted role for this residue in the transfer mechanism. The presence of acetylated viosamine *in vivo* has also been confirmed on the glycosylated viral fibers, using GC-MS and NMR. This study represents the first report of a virally-encoded sugar acetyltransferase.

INTRODUCTION

Previous reports have provided evidences that the genome of some giant and large DNA viruses encode autonomous glycosylation systems (1, 2). These machineries include glycosyltransferases and the enzymes required to produce their nucleotide-sugar substrates. Interestingly, for the few cases reported so far (3-5), novel and atypical glycans have been described. Pathways for monosaccharide and complex glycan synthesis have been characterized in Chlorella viruses (Phycodnaviridae family) and in some members of the Mimiviridae family (6-11). However, inspection of the genomes of newly identified giant viruses also revealed that genes associated with glycosylation pathways are also found in other members of the Nucleo-Cytoplasmic Large DNA Viruses (12).

Mimivirus is the first identified member of the growing family of Mimiviridae (13). Its 1.2 Mbp genome encodes ~1000 proteins and the pseudo-icosahedral virions of 400 nm diameter are covered by a dense array of 150 nm-long, highly glycosylated fibers (14). The micrometer size of the viral particles together with the glycosylated fibers mimic the bacteria on which the Mimivirus natural host, *Acanthamoeba castellanii*, feeds. In a previous study, data from compositional analysis of Mimivirus glycans preparations revealed the presence of various monosaccharides, including *N*-acetylglucosamine, glucose, rhamnose and the very rare 4-amino-4,6-dideoxyhexose viosamine along with its methylated derivative (6). The genes encoding the putative enzymes of the UDP-D-Viosamine biosynthesis pathway have been identified in a 9-gene cluster localized at the 5' end of Mimivirus genome, possibly devoted to glycan production (6). Specifically, the first enzyme, R141, is an UDP-D-glucose 4,6 dehydratase that catalyzes the formation of a UDP-4-keto-6-deoxy intermediate, also common to the UDP-L-rhamnose pathway (8). A pyridoxal phosphate (PLP)-dependent amino transferase, L136, then transfers an amino group from glutamate to the 4-keto group, leading to UDP-D-Vio production (6).

Viosamine has only been identified in some bacterial species. In *Pseudomonas syringae* it is a component of the flagellin-associated trisaccharide required for virulence and in this case the 4-amino group is acylated and a 2-O-methyl group is also present (15,16). This unusual sugar, 2-O-methyl-4-(3-hydroxy-3-methylbutanamido)-4,6-dideoxy-D-glucose, is also termed anthrose, due to its occurrence in *Bacillus anthracis* exosporium (17). Modified Vio has also been reported in few other bacterial organisms (18-25).

Since Vio is always acylated on the amino group, we undertook a study to verify its status in Mimivirus, by adopting a dual strategy. First, we scrutinized the cluster in search of a candidate for Vio acetylation and identified the N-terminal domain of L142 protein as a possible acyltransferase. Second, we performed GC-MS and NMR analyses of Mimivirus glycans. We now report the characterization of the N-terminal domain of L142 (N-L142) as an acetylCoA-dependent enzyme able to modify the 4-amino group of UDP-D-Vio. GC-MS and NMR analyses of Mimivirus glycans then confirmed that Vio is completely acetylated *in vivo*. To our knowledge, this represents the first report of a virally encoded sugar-N-acetyltransferase.

RESULTS

Sequence and structure analysis — L142 gene product is a bifunctional protein with a N-terminal domain with a predicted left-handed β-helix (LβH) superfamily fold and a C-terminal one, predicted as glycosyltransferase because of sequence homology. L142 is part of the 9-gene cluster (6) encoding two enzymes involved in UDP-D-Vio production (L136 and R141), the R135 protein already identified as a component of the outer fibers (26,27) and several putative glycosytransferases, some of them displaying two domains possibly derived from gene duplication events, as shown in Figure 1. In particular, the C-terminal domain of L-142 has strong homology with the N-terminal regions of L137 and L138 (Figure 1), both predicted to be type 2 glycosyltransferases (GT-2) (26). C-L142 is also predicted to belong to GT-2 family in the CAZy database (28).

Bioinformatic analysis of the N-terminal region of L142 (N-L142) revealed the presence of a hexapeptide repeat motif, typical of the left-handed β helix (LβH) domain, initially described in UDP-N-acetylglucosamine 3-O-acyltransferase (LpxA) (29). The LβH domain is found in many acyltransferases and contributes to the subunit interface, promoting trimerization, as well as active site formation. As already observed for L136, the enzyme which transfers the C-4 amino group, no close homologue of N-L142 was found in other viral genomes, excluding Mimivirus close relatives. Out of the Mimiviridae family, BLAST best hits corresponded to uncharacterized proteins from different bacterial species (with 30 to 37% identity over 196

amino acids) and from environmental sequences. The sequence conservation of N-L142 with known sugar acetyltransferases is reported in Table 1.

Interestingly, N-L142 did not show homology with enzymes known to modify Vio, such as VioB and AntD (15,22). Low homology was found with QdtC, which modifies the 3-amino group in the synthesis of dTDP-3-acetamido-3,6-dideoxy-□-D-glucose (30). On the other hand, a significant homology was observed with the well characterized PglD from *Campylobacter jejuni,* as well as with the C-terminal ATD domains of *Neisseria gonorreae* PglB and *Acinetobacter baumannii* WeeI, all catalyzing an acetyl transfer to the C-4 amino group for UDP-D-BacNAc$_2$ (31,32). Among the characterized sugar acetyltransferases, the best hit was with *Caulobacter crescentus* PerB catalyzing the last step of GDP-D-N-acetyl-perosamine synthesis (33). The structural alignment of N-L142 with PglD, PglB, WeeI and PerB is reported in Figure 2, which also highlights the residues involved in catalysis and substrate recognition (31-33).

Structural comparison of the N-L142 Phyre2 model with *N. gonorreae* PglB confirmed that it possesses all the necessary determinants for AcetylCoA and UDP-D-Vio recognition, as well as the catalytic residues involved in acetyl transfer and those involved in trimer formation. Interestingly, the structural comparison of the N-L142 model (Figure 3), based on PglB (4M99), with the PerB (4EA8) structure in complex with Acetyl CoA and GDP-D-N-acetylperosamine highlighted a steric hindrance due to an aspartate residue (D232 in 4M99, D40 in L142 and A39 in 4EA8) in the nucleotide binding site, preventing the accommodation of GDP but allowing a UDP to be properly positioned in the cavity. D55 in PerB is also replaced by T247 in PglB and I55 in N-L142. The overall loop is in a closer state than in the GDP bound structure. Moreover, while in the PglB structure an asparagine residue is known to interact with the acetyl bound to the C-2 amino group in bacillosamine (N162), in the N-L142 model it is another strand of the LβH that provides a specific residue (R155), able to engage a H-bond with the hydroxyl group in C-2 of the viosamine. This suggests that viosamine methylation in C-2 is the last step of the sugar modification, probably after the transfer of VioNAc on its acceptor.

Purification of Recombinant N-L142 and enzymatic activity — Wild-type (WT) N-142, H136A and H145A mutants, expressed as GST- fusion proteins, were purified to homogeneity using affinity purification. WT and mutants proteins were soluble after proteolytic release from GST; comparable amounts, 2 to 4 mg/L of bacterial culture, could be obtained for WT and mutant proteins.

UDP-D-Vio was produced as described (6) and used as substrate in the enzymatic activity assays. UDP-D-Vio (Figure 4, peak A) was incubated with purified WT N-L142 and acetyl-

CoA; product formation was monitored by anion exchange HPLC (Figure 4, peak B). Enzymatic activity of H136A mutant incubated in identical condition resulted in a decrease by three orders of magnitude of the enzymatic activity (3992 ± 554 µmole/min/mg protein for the WT versus 2.1 ± 1.5 µmole/min/mg protein for the H136A mutant tested within 48 hours from purification). While the WT protein remained stable, the H136A mutant displayed about 20-30% decrease of its initial activity after a week of storage. Increase in absorbance at 332 nm, suggestive of protein aggregation, or presence of a precipitate were not observed for the WT or H136A proteins. Even if the mutant protein lost activity at a faster rate than the WT enzyme, the three orders of magnitude difference observed immediately after purification is likely due to the mutation of the catalytic His136 rather than to a massive perturbation of the protein fold. This conclusion is also in agreement with the proposed role of the active site His in the catalytic mechanism for the PglD and PerB enzymes (32). No significant changes in the enzymatic activity were observed for the H145A mutant, in agreement with previous reports for PglD and PerB, where mutation of the corresponding His caused only minor effects in the catalytic activity (32).

Structural Characterization of L142 Product — The L142 product was purified using anion exchange HPLC and solid phase extraction (SPE), as described (6), for further ESI-MS and NMR analyses. ESI-MS revealed the presence of a main ion at 590 m/z, consistent with the expected mass of UDP-D-N-acetylviosamine (UDP-D-VioNAc). (Figure 5). A minor peak of 611.9 m/z, was consistent with the sodium adduct of UDP-D-VioNAc.

The purified product was also analysed by NMR. Identification of the UDP-D-VioNAc signals was possible by using both 2D homo- and heteronuclear NMR sequences (Table 1). The area at 6.0-5.5 ppm (Figure 6A) contained H-5 of the uracil moiety and two anomeric signals, the one at 5.99 ppm belonged to the expected ribofuranose unit of the UDP. Our analysis focused on that second one, which had to be related to the Vio residue. The proton at 5.57 ppm correlated with a carbon at 96.5 ppm (HSQC spectrum in Figure 6B), a chemical shift similar to that previously reported for the non N-acetylated form of UDP-D-Vio (6). Similarly to that, H-1 of N-acetylVio appeared as a double doublet, due to the coupling with two NMR active nuclei, phosphorous ($^{3}J_{H1,P}$ 6.7 Hz) and H-2 ($^{3}J_{H1,H2}$ 3.3 ppm). Identification of all the ring proton resonances was accomplished by COSY spectrum interpretation (Figure 6C). H-2 was at 3.61 ppm, partially overlapped to H-4. Nevertheless it was possible to evaluate the multiplicity of both proton signals: H-2 was a double triplet (Figure 6, inset) due to the coupling to H-1 ($^{3}J_{H2,H1}$ 3.3 Hz), to phosphorous ($^{4}J_{H2,P}$ 3.3 Hz) and H-3 ($^{3}J_{H2,H3}$ 9.8 Hz). Both H-3 (3.76 ppm) and H-4 (3.62 ppm) appeared as triplet (^{3}J 10.0 Hz) meaning that, as H-2 and H-5, they were axial

substituents of the pyranose ring of the sugar. This information confirmed the *gluco* stereochemistry of the sugar while the carbon chemical shift of C-4 (57.6 ppm) together with H-6/C-6 (1.17/18.2 ppm) values confirmed that it was Vio. Importantly, compared to the non-acetylated Vio, H-4 was at lower field (3.62 versus 3.02 ppm) indicating that the amino function was acetylated, as confirmed by the occurrence of an acetyl group in the proton/carbon spectrum at $^{1}H/^{13}C$ 2.03/23.3 ppm, in a one to one ratio with the methyl group of the 6-deoxy position. HMBC spectrum (Figure 6B) disclosed that the methyl of this acetyl and H-4 correlated with the same carbonyl group (175.6 ppm) demonstrating that Vio was N-acetylated.

VioNAc from Mimivirus *surface glycans*— The sugar composition of Mimivirus glycans was already investigated by GC-MS in a previous study (6). After the alditol acetate derivatization, the major components of viral glycans were rhamnose, glucose, N-acetylglucosamine and Vio. Analysis of the fragmentation spectrum of an unknown peak revealed the presence of methyl-Vio. However, the presence of acetylation on the C-4 amino group could not be detected using this method. Thus, Vio acylation on the surface of purified Mimivirus particles was analysed by partially methylated alditol acetate (PMAA) method, followed by GC-MS analysis. The presence of a single peak containing one methyl group on Vio C-4 amino group, as evidenced by GC-MS fragmentation spectrum, is consistent with complete N-acylation of this monosaccharide in Mimivirus glycan (Figure 7). In addition, PMAA analysis indicated that Vio is terminal and not further elongated by other sugars (Figure 7).

The presence of Vio and the identity of the N-linked substituent were confirmed by NMR analyses (Figure 8, Table 1) of the glycans extracted from Mimivirus. Inspection of the HSQC spectrum (Figure 8C) disclosed a complex pattern of anomeric signals (^{1}H range 5.1-4.5 ppm), a crowded carbinolic area (^{1}H range 4.4-3.1 ppm), two main acetyl signals (~ 2.0 ppm) and a group of methyl signals (not shown) characteristic of 6-deoxysugars at ~ 1.3 ppm. All these signals suffered of low resolution and recording the spectra at high temperature did not improve their overall quality. However, identification of the Vio unit was achieved, even though it was not possible to determine at which residue it was further connected.

Vio anomeric signal was at ($^{1}H/^{13}C$) 4.59/105.3 ppm, values diagnostic of a residue β configured at the anomeric centre. TOCSY spectrum (Figure 8B) disclosed that this signal correlated with other five protons (3.64, 3.56, 3.54, 3.22 and 1.24 ppm): this correlation pattern along with the presence of a methyl group at 1.24 ppm, identified this unit as Vio, labelled with **V**. Combination of COSY and TOCSY spectra, established the sequence of the different protons, while HSQC identified the corresponding carbon chemical shifts. H-2 (3.22 ppm) correlated with a carbon at low field (84.3 ppm) because methylated at the corresponding hydroxyl

function (O-CH_3 at $^1H/^{13}C$ 3.63/60.9 ppm), as confirmed by the $V_{2,2OMe}$ and $V_{2OMe,2}$ cross peaks in the HMBC spectrum (Figure 8C). H-2 enabled H-3 identification which in turn led to H-4; H-5 was found almost coincident with H-3, as suggested from NOESY spectrum (Figure 8A) which had one intense cross peak at ~ 3.55 ppm embracing both H-3 and H-5; H-5 correlated further with H-6. Inspection of long range correlations from H-6 (1.24 ppm) identified both C-4 and C-5 (57.8 and 72.0 ppm, respectively) so that C-3 value was finally selected a confirmed and its value indicated by the H-4/C-3 correlation in the HMBC spectrum (Figure 8C). Of note, H-4 chemical shift (3.64 ppm), was similar to that of UDP-D-VioNAc (3.62 ppm) and not to that of UDP-D-Vio (3.02 ppm), in agreement with the amino function acylation. Accordingly, both H-4 and the methyl of the acetyl at 2.07 ppm had a long-range correlation with a carbonyl at 175.1 ppm, disclosing that the acyl of the amino group was an acetyl.

Indeed, in the polysaccharide fibers, Vio has a β-anomeric linkage, is methylated at position 2, is acetylated at the amino function and has no other substituent; it occupies a terminal position in agreement with PMAA analysis. HMBC and NOESY spectra identified the density of a monosaccharide linked with Vio (see at the cross of the dotted lines in Figure 8), but poor spectra resolution hampered the elucidation of the nature of this residue.

DISCUSSION

Previous studies showed that the unusual sugar Vio is a component of Mimivirus glycans and that it is mainly contained in the long fibers that surround the capsid (5,6). Here we demonstrate that the 4-amino group of UDP-D-Vio is acetylated by the N-terminal domain of L142 gene product. The *in vitro* data obtained with the recombinant enzyme are consistent with the finding that Vio is also N-acetylated *in vivo* in Mimivirus glycans. Indeed, GC-MS and NMR data clearly indicate that, in fibers, Vio is terminal and it is both N-acetylated on C-4 and methylated on C-2.

Viosamine, acetylated on the C-4 amino group, often also methylated on C-2 is restricted to some bacterial species, pathogenic to both vertebrates and plants. In *P. syringae*, dTDP-VioNAc is produced by a set of enzymes contained in a gene cluster named "Vio island", which includes VioA, which transfers the amino group to the 4-keto group of dTDP-4-keto-6-deoxy-D-glucose and VioB that acetylates the 4-amino group (15). Other enzymes in the cluster further convert dTDP-D-VioNAc to dTDP-D-N-(3-hydroxy-1-oxobutyl)Vio and its 2-methylated derivative. Modified Vio was found in the flagellin-associated glycans and disruption of its biosynthetic pathway impairs motility and virulence on host tobacco leaves (16). Similar "Vio islands" were also identified *in P. aeruginosa* PAK (15,34) and in *B. anthracis* (17). In this

latter organism, modified Vio is a component of the exosporium pentasaccharide (17,29). The biosynthetic genes for dTDP-VioNAc were also identified in *E. coli* O7 and *Shigella dysenteriae* type 7 (22). However, in most cases the acetyl group is further modified into more complex moieties and simple acetylation has been rarely reported.

Bioinformatic analysis of N-L142 sequence showed that, surprisingly, it has no homology to known enzymes involved in Vio N-acetylation or acylation. Similarly, the first enzyme of the Mimivirus Vio pathway, L136, has very low homology with the corresponding VioA of *P. syringae* or *E. coli*. On the other hand, a significant homology was found with enzymes that catalyze the acetylation of the 4-amino group of UDP-D-BacNAc (31,33). $BacNAc_2$ is an essential component of bacterial N-linked and O-linked glycans, where it represents the first sugar attached to the protein. Moreover, good homology was found with *C. crescentus* PerB, responsible for GDP-D-perosamine acetylation (33). This finding was also confirmed by comparison of N-L142 structural model with *N. gonorreae* PglB and *C. crescentus* PerB (Figure 3). N-L142 shows a typical LβH superfamily fold, typical of this type of acyltransferases. Chantigian *et al*., starting from X-ray structures and site-directed mutagenesis analyses, have proposed the presence of two different classes of LβH enzymes able to acylate nucleotide-sugars, based on substrate binding orientations and reaction mechanisms (35). *N. gonorrae* PglB, *C. jejuni* PglD and *C. crescentus* PerB belongs to Class I, since they use a conserved histidine in the active site as a catalytic base. On the other hand, for class II enzymes, comprising QdtC and AntD, a substrate-assisted catalytic mechanism has been proposed (35). To confirm that N-L142 also belongs to class I acetyltransferases, we have performed site-directed mutagenesis of the proposed catalytic His residue. Indeed, the H136A mutant activity, assayed immediately after purification, was three orders of magnitude lower when compared to the WT. This finding matches what described for PglB and PglD His mutants, although the decrease of the catalytic efficiency of these other enzymes ranged from four to six orders of magnitude. Upon storage, the N-L142 H136A mutant activity decreased faster than the WT, being 70-80% of the original value after one week of storage. On the other hand, protein aggregation or precipitation could not be detected suggesting that H136 has a minor role in perturbing the structure of the protein. Accordingly, our data and the comparison of the N-L142 structural model with the published homologous structures clearly pinpointed a catalytic role for this His, as well as the other molecular determinants responsible for substrate specificity.

The C-terminal part of the L142 protein displays a glycosyltrasferase GT-2 fold, suggesting that this domain can be responsible for the attachment of the acetylated Vio on the Mimivirus fibers. Interestingly, NMR analysis has revealed that VioNAc is bound via a β-anomeric linkage, thus

indicating that the involved enzyme behaves as an inverting transferase. Several GT-2 enzymes are contained in the same gene cluster as L142, and are probably derived from gene duplication and fusion events. The origin of this cluster is not clear, since it contains genes related to both prokaryotes (i.e. L136 and L142) and eukaryotes (R141). However, since these putative GTs are only found in Mimiviridae members of group A, it is likely that they are involved in the transfer of the monosaccharides that are uniquely produced by these viruses, i.e. rhamnose and Vio (6,7).

The mechanisms of production of the complex carbohydrates of Mimivirus and other large DNA viruses are largely unknown. Besides Chlorella viruses, which revealed the presence of novel and unique structures (3), glycans from other viral families still await characterization. Several evidences have already suggested that glycosylation occurs in the cytosol, in the so-called "viral factories", but information about the organization of the glycosylation machinery in these factories are still lacking, as well as on the origin of the enzymes involved in these processes. Identification and characterization of the enzymes encoded by the viral genomes will help shed light on these issues.

EXPERIMENTAL PROCEDURES

L142 Sequence and Structural Analyses —The most similar homologues of N-L142 and C-L142 were identified using the BLAST tool on the NCBI server, using the "nr" and the "env-nr" databases. CAZy database was also used for glycosyltransferase identification. The N-L142 sequence was submitted to the Phyre server (36), which returned a model of the L142 N-terminal domain (Leu8 to Ile206) based on the PglB structure from *N. gonorrhoeae* (100% confidence, 4M99) (31). We used the 4M99 structure to model the Acetyl CoA cofactor in the N-terminal domain of the L142 binding site and the PerB structure in complex with Acetyl CoA and GDP-D-N-acetylperosamine (4EA8) to define the nucleotide acetylated sugar binding site (33). A model of UDP-D-VioNAc manually built from the GDP-D-N-acetylperosamine and UDP-D-BacNAc$_2$ (3BSS, 31) was fitted in the nucleotide sugar binding pocket of N-L142. We used the molecular visualization program VMD (37) to compare the four structures (PglB, PerB, PglD and WeeI) and define the amino-acids involved in ligand binding.

Expression and Purification of Recombinant L142 Proteins— The N-terminal acyltransferase domain of L142 (N-L142; aminoacid 1-213) was expressed as a recombinant glutathione GST-fusion protein in *E. coli* strain BL21 (DE3) (New England Biolabs) using the pGEX-6-P1 vector (GE Healthcare). The PCR amplified sequence corresponding to bases 1-639 of *L142* ORF, was digested with BamHI-HF® and XhoI restriction enzymes (New England Biolabs) and ligated in

the plasmid vector. Site-directed mutagenesis was performed using Quickchange (Agilent), following the recommended protocols. Primers were designed using Quickchange primer design program. Sequencing of WT and mutants was performed by TibMolbiol (Genova, Italy). Protein expression, purification and proteolytic cleavage were performed as described previously (8). Proteins were concentrated to about 1 mg/ml using Amicon Ultra-4 10K (Millipore) and stored in PBS at 4°C. They were analysed by UV absorbance from 210 to 340 nm and concentration was determined using ε_{280} = 15930 M^{-1} cm^{-1} (38). Purity was determined by SDS-PAGE.

Enzymatic Assays — Mimivirus N-L142 enzymatic activity was assayed on the UDP-D-Vio, produced using recombinant Mimivirus R141 and L136 (6,8). UDP-D-Vio was incubated with acetyl-CoA in presence of N-L142 in PBS, pH 7.3 at 25 °C. Reactions with different concentrations of substrates were performed. The reactions were stopped by heat inactivation for 3 min at 80 °C at different time points of incubation and the solutions were clarified by microfiltration (Millipore). After clarification, the reaction mixtures were analysed by HPLC, as previously described (7). Specific activity was determined using the reduction of UDP-D-Vio peak area and expressed as mean ± SD of µmoles converted per min/mg protein, using 0.2 mM UDP-D-Vio and 0.4 mM acetyl-CoA as substrates. Analyses were performed in duplicate at different time points after enzyme purification from two independent protein preparations.

Structural Characterization of L142 Product — To confirm the acyltransferase activity of Mimivirus N-L142, its product was purified as previously described (6) and analyzed by ESI-MS and by NMR. ESI-MS analysis was performed in direct infusion analysis at 5 µl/min on an Agilent 1100 series LC/MSD ion trap XCT instrument (Agilent Technologies, Palo Alto, CA, USA). Nucleotide sugars were diluted up to 10 pmol/ml in a water:acetonitrile (50:50) solution containing 0.1% formic acid. Spectra were acquired in negative ion mode in the mass range of the expected *m/z* ratios, as described (6).

For NMR analysis, 1D and 2D NMR spectra were recorded on a Bruker 600 DRX equipped with a CryoProbe™ on a solution of 500 µl of D_2O, at 25°C. DQF-COSY experiment was performed using data sets of 2048 × 512 points (39,40); the data matrix was zero-filled in both dimensions to give a matrix of 4K × 2K points and was resolution enhanced in both dimensions by a cosine-bell function before Fourier transformation. Coupling constants were determined on a first order basis from high-resolution 1D spectra. HSQC and HMBC spectra were measured in the ^{1}H-detected mode via single quantum coherence with proton decoupling in the ^{13}C domain, using data sets of 2048 × 512 points. Experiments were carried out in the phase-

sensitive mode (39) and the data matrix was extended to 4096 × 2048 points using forward linear prediction extrapolation.

Analysis of VioNAc presence on Mimivirus *glycan*— The presence of acylation on the C-4 amino group of Vio in Mimivirus glycans was verified using the PMAA derivatization technique (41). Approximately 0.5-1 x 10^{11} whole Mimivirus particles were lyophilized overnight to remove any trace of water which could inhibit the following permethylation reaction. The dry viral particles were suspended in DMSO (1 ml) with powdered NaOH. Then methyl iodide (0.5 ml) was added and the sample stirred for 45 minutes at room temperature. The reaction was quenched by dropwise addition of 5% acetic acid aqueous solution. To purify the permethylated particles, 2 ml of chloroform were added and the reaction mixture made up to 5 ml with ultrapure water. The sample was thoroughly mixed and allowed to settle into two layers. The upper aqueous layer was removed and discarded. The chloroform layer was washed several times with ultrapure water and dried under a gentle stream of nitrogen.

The dry permethylated particles were hydrolysed with TFA and further processed to obtain the alditol acetate derivatives, as previously described (7). The PMAA were extracted from the solid crust at the bottom of the vial 4 times with 0.5 ml of dichloromethane and collected in a new vial. The samples, evaporated and suspended in a small volume of dichloromethane, were analysed by GC-MS as previously described (7).

Isolation of Mimivirus *polysaccharides and NMR analysis*— Mimivirus suspension (2 ml, ~ 2.4 x 10^{11} particles) were stirred with 6 ml of 0.5 M DTT solution at 100°C for 1 h to promote the complete removal of the fibrils. The slurry was centrifuged (8000 rpm, 10 min) and the solid washed twice with ultrapure water. Supernatants were pooled and dialyzed, yielding to a crude polysaccharide preparation (32 mg), an aliquot (5 mg) was analyzed *via* NMR without further purification. NMR conditions were the same as described for UDP-D-VioNAc, except temperature that was set to 56°C, and TOCSY and NOESY spectra were recorded, by using 100 and 200 ms mixing time, respectively. Spectra were calibrated using the methyl group of acetone ($^{1}H/^{13}C$ 2.225/31.45 ppm) added as internal standard.

Acknowledgements: This work was partially supported by FRA (Fondo per la Ricerca di Ateneo), University of Genova, and French National Research Agency ANR-14-CE14-0023-01. AN is recipient of a UIF Vinci program 2015 PhD fellowship (contract C3_90).

Conflict of interest: Authors declare no conflicts of interest with the content of this article.

Authors contributions: FP, MG and MEL performed protein cloning and expression, mutagenesis, enzyme analysis, samples preparation and purification for further characterization. FP and CA performed bioinformatic analyses. AS and GLD analyzed samples by ESI-MS and GC-MS. CDC and AN performed NMR analysis. CA and SJ developed and analyzed the L142 structural model. MT conceived and coordinated the study and wrote the paper. All authors analyzed the results and approved the final version of the manuscript.

REFERENCES

1. Piacente, F., Gaglianone, M., Laugieri, M. E., and Tonetti, M. G. (2015) The Autonomous Glycosylation of Large DNA Viruses. *Int J Mol Sci* **16**, 29315-29328

2. Van Etten, J. L., Gurnon, J. R., Yanai-Balser, G. M., Dunigan, 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

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

4. 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 Engl* **55**, 654-658

5. Hulsmeier, A. J., and Hennet, T. (2014) O-Linked glycosylation in Acanthamoeba polyphaga mimivirus. *Glycobiology* **24**, 703-714

6. 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

7. Piacente, F., De Castro, C., Jeudy, S., Molinaro, A., Salis, A., Damonte, G., Bernardi, C., Abergel, C., and Tonetti, M. G. (2014) Giant virus Megavirus chilensis encodes the biosynthetic pathway for uncommon acetamido sugars. *J Biol Chem* **289**, 24428-24439

8. Parakkottil Chothi, M., Duncan, G. A., Armirotti, A., Abergel, C., Gurnon, J. R., Van Etten, J. L., Bernardi, C., Damonte, G., and Tonetti, M. (2010) Identification of an L-rhamnose synthetic pathway in two nucleocytoplasmic large DNA viruses. *J Virol* **84**, 8829-8838

9. 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

10.Luther, K. B., Hulsmeier, A. J., Schegg, B., Deuber, S. A., Raoult, D., and Hennet, T. (2011) Mimivirus collagen is modified by bifunctional lysyl hydroxylase and glycosyltransferase enzyme. *J Biol Chem* **286**, 43701-43709

11.Rommel, A. J., Hulsmeier, A. J., Jurt, S., and Hennet, T. (2016) Giant mimivirus R707 encodes a glycogenin paralog polymerizing glucose through alpha and beta glycosidic linkages. *Biochem J*

12.Iyer, L. M., Balaji, S., Koonin, E. V., and Aravind, L. (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. *Virus Res* **117**, 156-184

13.Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J. M. (2004) The 1.2-megabase genome sequence of Mimivirus. *Science* **306**, 1344-1350

14.Legendre, M., Santini, S., Rico, A., Abergel, C., and Claverie, J. M. (2011) Breaking the 1000-gene barrier for Mimivirus using ultra-deep genome and transcriptome sequencing. *Virol J* **8**, 99

15.Yamamoto, M., Ohnishi-Kameyama, M., Nguyen, C. L., Taguchi, F., Chiku, K., Ishii, T., Ono, H., Yoshida, M., and Ichinose, Y. (2011) Identification of Genes Involved in the Glycosylation of Modified Viosamine of Flagellins in Pseudomonas syringae by Mass Spectrometry. *Genes (Basel)* **2**, 788-803

16.Taguchi, F., Yamamoto, M., Ohnishi-Kameyama, M., Iwaki, M., Yoshida, M., Ishii, T., Konishi, T., and Ichinose, Y. (2010) Defects in flagellin glycosylation affect the virulence of Pseudomonas syringae pv. tabaci 6605. *Microbiology* **156**, 72-80

17.Dong, S., McPherson, S. A., Wang, Y., Li, M., Wang, P., Turnbough, C. L., Jr., and Pritchard, D. G. (2010) Characterization of the enzymes encoded by the anthrose biosynthetic operon of Bacillus anthracis. *J Bacteriol* **192**, 5053-5062

18.Arbatsky, N. P., Kondakova, A. N., Shashkov, A. S., Drutskaya, M. S., Belousov, P. V., Nedospasov, S. A., Petrova, M. A., and Knirel, Y. A. (2010) Structure of the O-antigen of Acinetobacter lwoffii EK30A; identification of d-homoserine, a novel non-sugar component of bacterial polysaccharides. *Org Biomol Chem* **8**, 3571-3577

19.Kondakova, A. N., Linder, B., Fudala, R., Senchenkova, S. N., Moll, H., Shashkov, A. S., Kaca, W., Zahringer, U., and Knirel, Y. A. (2004) New structures of the O-specific polysaccharides of proteus. 4. Polysaccharides containing unusual acidic N-acyl derivatives of 4-amino-4,6-dideoxy-D-glucose. *Biochemistry (Mosc)* **69**, 1034-1043

20.Liu, B., Chen, M., Perepelov, A. V., Liu, J., Ovchinnikova, O. G., Zhou, D., Feng, L., Rozalski, A., Knirel, Y. A., and Wang, L. (2012) Genetic analysis of the O-antigen of Providencia alcalifaciens O30 and biochemical characterization of a formyltransferase involved in the synthesis of a Qui4N derivative. *Glycobiology* **22**, 1236-1244

21.Perepelov, A. V., Wang, Q., Senchenkova, S. N., Shevelev, S. D., Shashkov, A. S., Feng, L., Knirel, Y. A., and Wang, L. (2008) Structure and characterization of the gene cluster of the O-antigen of Escherichia coli O49 containing 4,6-dideoxy-4-[(S)-3-hydroxybutanoylamino]-D-glucose. *Biochemistry (Mosc)* **73**, 406-410

22.Wang, Y., Xu, Y., Perepelov, A. V., Qi, Y., Knirel, Y. A., Wang, L., and Feng, L. (2007) Biochemical characterization of dTDP-D-Qui4N and dTDP-D-Qui4NAc biosynthetic pathways in Shigella dysenteriae type 7 and Escherichia coli O7. *J Bacteriol* **189**, 8626-8635

23.Ovchinnikova, O. G., Valueva, O. A., Kocharova, N. A., Arbatsky, N. P., Maszewska, A., Zablotni, A., Shashkov, A. S., Rozalski, A., and Knirel, Y. A. (2013) Structure of the O-polysaccharide of Providencia alcalifaciens O35 containing an N-[(S)-1-carboxyethyl]-L-alanine (alanopine) derivative of 4-amino-4,6-dideoxyglucose. *Carbohydr Res* **375**, 73-78

24.Kondakova, A. N., Kirsheva, N. A., Shashkov, A. S., Shaikhutdinova, R. Z., Arabtsky, N. P., Ivanov, S. A., Anisimov, A. P., and Knirel, Y. A. (2011) Low structural diversity of the O-polysaccharides of Photorhabdus asymbiotica subspp. asymbiotica and australis and their similarity to the O-polysaccharides of taxonomically remote bacteria including Francisella tularensis. *Carbohydr Res* **346**, 1951-1955

25.Vinogradov, E., Nossova, L., Korenevsky, A., and Beveridge, T. J. (2005) The structure of the capsular polysaccharide of Shewanella oneidensis strain MR-4. *Carbohydr Res* **340**, 1750-1753

26.Boyer, M., Azza, S., Barrassi, L., Klose, T., Campocasso, A., Pagnier, I., Fournous, G., Borg, A., Robert, C., Zhang, X., Desnues, C., Henrissat, B., Rossmann, M. G., La Scola, B., and Raoult, D. (2011) Mimivirus shows dramatic genome reduction after intraamoebal culture. *Proc Natl Acad Sci U S A* **108**, 10296-10301

27.Klose, T., Herbst, D. A., Zhu, H., Max, J. P., Kenttamaa, H. I., and Rossmann, M. G. (2015) A Mimivirus Enzyme that Participates in Viral Entry. *Structure* **23**, 1058-1065

29 Lombard, V., Golaconda Ramulu, H., Drula, E,, Coutinho, P.M., and Henrissat B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490-D495

28.Raetz, C. R., and Roderick, S. L. (1995) A left-handed parallel beta helix in the structure of UDP-N-acetylglucosamine acyltransferase. *Science* **270**, 997-1000

29.Kubiak, R. L., and Holden, H. M. (2012) Structural studies of AntD: an N-Acyltransferase involved in the biosynthesis of D-Anthrose. *Biochemistry* **51**, 867-878

30.Thoden, J. B., Cook, P. D., Schaffer, C., Messner, P., and Holden, H. M. (2009) Structural and functional studies of QdtC: an N-acetyltransferase required for the biosynthesis of dTDP-3-acetamido-3,6-dideoxy-alpha-D-glucose. *Biochemistry* **48**, 2699-2709

31.Morrison, M. J., and Imperiali, B. (2013) Biochemical analysis and structure determination of bacterial acetyltransferases responsible for the biosynthesis of UDP-N,N'-diacetylbacillosamine. *J Biol Chem* **288**, 32248-32260

32.Olivier, N. B., and Imperiali, B. (2008) Crystal structure and catalytic mechanism of PglD from Campylobacter jejuni. *J Biol Chem* **283**, 27937-27946

33.Thoden, J. B., Reinhardt, L. A., Cook, P. D., Menden, P., Cleland, W. W., and Holden, H. M. (2012) Catalytic mechanism of perosamine N-acetyltransferase revealed by high-resolution X-ray crystallographic studies and kinetic analyses. *Biochemistry* **51**, 3433-3444

34.Arora, S. K., Bangera, M., Lory, S., and Ramphal, R. (2001) A genomic island in Pseudomonas aeruginosa carries the determinants of flagellin glycosylation. *Proc Natl Acad Sci U S A* **98**, 9342-9347

35.Chantigian, D. P., Thoden, J. B., and Holden, H. M. (2013) Structural and biochemical characterization of a bifunctional ketoisomerase/N-acetyltransferase from Shewanella denitrificans. *Biochemistry* **52**, 8374-8385

36.Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**, 845-858

37.Humphrey, W., Dalke, A., and Schulten, K. (1996) VMD: visual molecular dynamics. *J Mol Graph* **14**, 33-38, 27-38

38.Gasteiger E., H. C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.;. (2005) Protein Identification and Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* (Walker, J. M. ed.), Humana Press. pp 571-607

39.States DJ, H. R., Ruben DJ. (1982) A Two-Dimensional Nuclear Overhauser Experiment with Pure Absorption Phase in Four Quadrants. *J Magn Reson* **48**, 8

40.Rance, M., Sorensen, O. W., Bodenhausen, G., Wagner, G., Ernst, R. R., and Wuthrich, K. (1983) Improved spectral resolution in cosy 1H NMR spectra of proteins via double quantum filtering. *Biochem Biophys Res Commun* **117**, 479-485

41.Ciucanu, I. (2006) Per-O-methylation reaction for structural analysis of carbohydrates by mass spectrometry. *Anal Chim Acta* **576**, 147-155

42. Marchler-Bauer, A. (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res*. **43**, 222-226

FOOTNOTES

The abbreviations used are: Vio, viosamine; VioNAc, N-acetylviosamine; GC-MS, gas chromatography-mass spectrometry; NMR, nuclear magnetic resonance; GST, glutathione S-transferase, ESI-MS, electrospray ionization- mass spectrometry; DQF-COSY, double quantum filtered phase-sensitive homonuclear correlation spectroscopy; HSQC, heteronuclear single-quantum correlation spectroscopy; HMBC, heteronuclear multiple-bond correlation spectroscopy; NOESY, nuclear Overhauser effect spectroscopy; PMAA, partially methylated alditol acetate.

FIGURE LEGENDS

FIGURE 1. Organization of the L142 gene cluster. (A) Besides L142, the 9-gene cluster contains the two enzymes for UDP-D-Vio production (R141 and L136), a putative pyruvoyltransferase (L143), a structural protein of the outer fibers (R135) and putative glycosyltransferases (L137, L138, R139 and L140). (B) Organization of the glycosyltransferase domains. AT: acetyltransferase domain. GT2: glycosyltransferase domain.

FIGURE 2. Structural alignment of N-L142. C-terminal ATD domain of *Neisseria gonorreae* PglB (4M99), *Caulobacter crescentus* PerB (4EA8), *Acinetobacter baumannii* WeeI (4M9C) and *Campylobacter jejuni* PglD (3BSS). Residues involved in trimer interface and in acetyl-coA binding site are marked in red (based on cd03360, 42, and our analyses of the model). The conserved catalytic histidine is marked by a blue asterisk and the position corresponding to viosamine C-2-O binding in L142 is marked by a red star (R155 in bold), while the PglD N162 residue involved in C-2 acetylbacillosamine binding is marked by a green star.

FIGURE 3. Ribbon representation of three orientations (ABC) of the N-L142 homotrimer model. Each monomer has a different color for clarity. Acetyl-CoA (ACO) and UDP-D-N-acetylviosamine (UDV) are represented as CDKs colored by atom types. The R155 residue making H-bonds with the viosamine C-2O in each monomer is represented as yellow colored CDK and the catalytic H136 as green colored CDK.

FIGURE 4. Anion exchange HPLC analysis of nucleotide-sugars. UDP-D-Vio (peak A) was incubated with the recombinant L142 N-terminal domain in the presence of acetyl-CoA at 25 °C. The progressive formation of a new compound (peak B) was observed.

FIGURE 5. ESI-MS spectrum of the purified N-L142 product. The purified product of N-L142 was analysed by ESI-MS. The main ion of 590 m/z corresponds to the calculated UDP-D-VioNAc molecular mass. Ion 612 m/z represents the sodium adduct. Other peaks are consistent with residual Acetyl-CoA and CoA-SH and their sodium adducts, co-purified with the UDP-D-VioNAc.

FIGURE 6. NMR analysis of purified UDP-D-VioNAc. (A) (600 MHz, 25°C, D_2O) full proton spectrum recorded for UDP-D-VioNAc, the inset details H-2 and H-4 signals and the grey lines dissect the contribution given by each proton to the complex signal. Only signals belonging to

ribose (R) and Vio (V) are labelled, other signals belong to residual NADPH or impurities. (B) (600 MHz, 25°C, D_2O) expansion of COSY spectrum of UDP-D-VioNAc. Only densities belonging to ribose (R) and viosamine (V) cross peak densities are attributed, other signals belong to residual NADPH or impurities. (C) (600 MHz, 25°C, D_2O) expansion of COSY spectrum of UDP-D-VioNAc. Only cross peak densities belonging to ribose (R) and viosamine (V) are attributed, other signals belong to residual NADPH.

FIGURE 7. Identification of *N*-acylated Vio in Mimivirus by PMAA GC-MS analysis. The viral particles were derivatized by PMAA technique and then analysed by GC-MS. Chromatogram of VioNAc was obtained by extraction of m/z 117 ion current.Peak identification was performed by comparing UDP-D-VioNAc standard analysis and fragmentation spectra (A) Zoom view of the m/z 117 extracted chromatogram of Mimivirus. (B) Fragmentation spectrum of VioNAc

FIGURE 8. NMR of crude polysaccharide from Mimivirus fibrils. (600 MHz, 56°C, D_2O) 2D spectra recorded for the crude polysaccharide from Mimivirus fibrils. A) Expansion of NOESY spectrum showing correlation from Vio (V) anomeric proton. B) Expansion of TOCSY spectrum showing correlation from Vio anomeric proton. C) Overlap of HSQC (black/dark grey) and HMBC (pale grey) spectra. D) Expansion of HMBC spectrum detailing the long range correlations with the carbonyl group. Vio densities are labelled with V, dotted lines point to the HSQC density of the residue that has Vio linked. * inter-residual NOE density between Vio and the other monosaccharide. ** NOESY artefact due to minimal spin diffusion.

TABLE 1. Sequence identity of N-L142 with characterized sugar acetyltransferases

	Protein	Product	Identitity
BAH58344	VioB *Pseudomonas syringae*	dTDP-D-VioNAc (15)	18%
Q9XCW3	VioB *Escherichia coli*	dTDP-D-VioNAc (22)	-
3VBJ_A	AntD *Bacillus cereus*	dTDP-D-anthrose (29)	-
AAR85517	QdtC *Thermoanaerobacterium thermosaccharolyticum*	dTDP-3-acetamido-3,6-dideoxy-D-glucose (30)	25%
4M99	PglB *Neissera gonorreae*	UDP-D-BacNAc$_2$ (31)	30%
3BSS	PglD *Campylobacter jejuni*	UDP-D-BacNAc$_2$ (32)	30%
4M9C	WeeI *Acinetobacter baumannii*	UDP-D-BacNAc$_2$ (32)	24%
4EA8	PerB *Caulobacter crescentus*	GDP-D-N-acetyl-perosamine (33)	32%

TABLE 2. ^{1}H and ^{13}C chemical shifts of UDP-β-VioNAc as UDP precursor and in the polysaccharide

		UDP—VioNAc					
Residue	Nucleus	**1**	**2**	**3**	**4**	**5**	**6**
VioNAc	^{1}H	5.57	3.61	3.76	3.62	4.04	1.18
	^{13}C	96.4	73.3	71.8	57.6	69.3	18.2
Ac	^{1}H	--	2.03				
	^{13}C	175.6	23.3				
Ribose	^{1}H	5.99	4.38	4.37	4.29	4.24-4.20	--
	^{13}C	89.5	74.9	70.8	84.5	66.2	--
Uracil	^{1}H	--	--	--	--	5.97	7.96
	^{13}C	--	152.9	--	167.5	103.7	142.7
		VioNAc (polysaccharide)					
β-VioNAc	^{1}H	4.59	3.22	3.56	3.64	3.54	1.24
	^{13}C	105.3	84.3	73.9	57.8	72.0	18.3
OMe	^{1}H	3.63					
	^{13}C	60.9					
Ac	^{1}H	--	2.07				
	^{13}C	175.1	23.5				

* Spectra calibration used the methyl group of acetone ($^{1}H/^{13}C$ 2.225/31.45 ppm) added as internal standard.

FIGURE 1

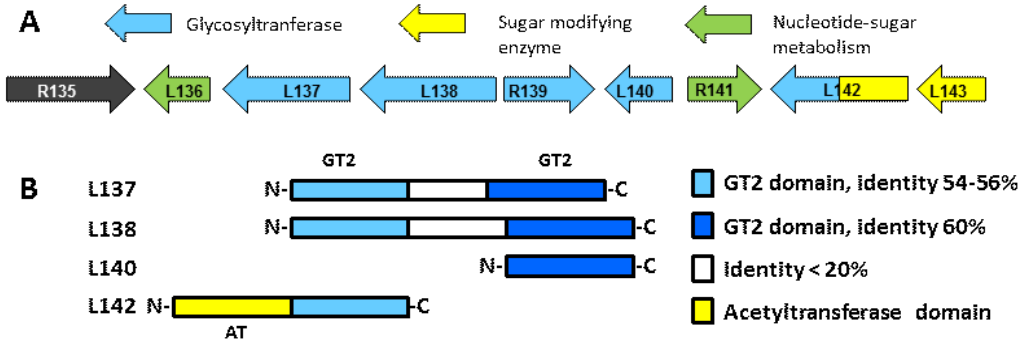

A
Glycosyltranferase
Sugar modifying enzyme
Nucleotide-sugar metabolism
R135
L136
L137
L138
R139
L140
R141
L142
L143
B
GT2
GT2
L137
L138
L140
L142
N-
-C
AT
GT2 domain, identity 54-56%
GT2 domain, identity 60%
Identity < 20%
Acetyltransferase domain

FIGURE 2

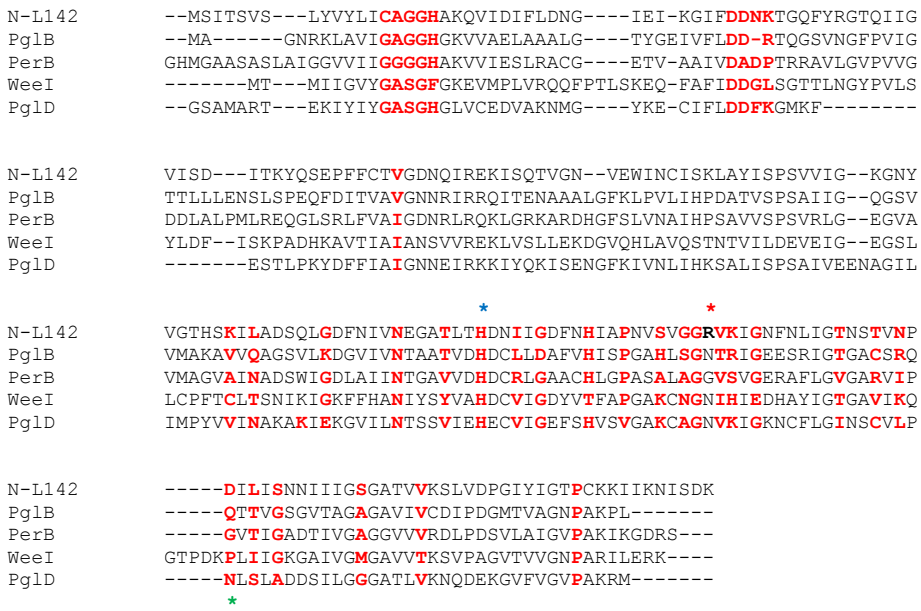

N-L142 --MSITSVS---LYVYLICAGGHAKQVIDIFLDNG----IEI-KGIFDDNKTGQFYRGTQIIG
PglB --MA------GNRKLAVIGAGGHGKVVAELAAALG----TYGEIVFLDD-RTQGSVNGFPVIG
PerB GHMGAASASLAIGGVVIIGGGGHAKVVIESLRACG----ETV-AAIVDADPTRRAVLGVPVVG
WeeI -------MT---MIIGVYGASGFGKEVMPLVRQQFPTLSKEQ-FAFIDDGLSGTTLNGYPVLS
PglD --GSAMART---EKIYIYGASGHGLVCEDVAKNMG----YKE-CIFLDDFKGMKF--------
N-L142 VISD---ITKYQSEPFFCTVGDNQIREKISQTVGN--VEWINCISKLAYISPSVVIG--KGNY
PglB TTLLLENSLSPEQFDITVAVGNNRIRRQITENAAALGFKLPVLIHPDATVSPSAIIG--QGSV
PerB DDLALPMLREQGLSRLFVAIGDNRLRQKLGRKARDHGFSLVNAIHPSAVVSPSVRLG--EGVA
WeeI YLDF--ISKPADHKAVTIAIANSVVREKLVSLLEKDGVQHLAVQSTNTVILDEVEIG--EGSL
PglD -------ESTLPKYDFFIAIGNNEIRKKIYQKISENGFKIVNLIHKSALISPSAIVEENAGIL
N-L142 VGTHSKILADSQLGDFNIVNEGATLTHDNIIGDFNHIAPNVSVGGRVKIGNFNLIGTNSTVNP
PglB VMAKAVVQAGSVLKDGVIVNTAATVDHDCLLDAFVHISPGAHLSGNTRIGEESRIGTGACSRQ
PerB VMAGVAINADSWIGDLAIINTGAVVDHDCRLGAACHLGPASALAGGVSVGERAFLGVGARVIP
WeeI LCPFTCLTSNIKIGKFFHANIYSYVAHDCVIGDYVTFAPGAKCNGNIHIEDHAYIGTGAVIKQ
PglD IMPYVVINAKAKIEKGVILNTSSVIEHECVIGEFSHVSVGAKCAGNVKIGKNCFLGINSCVLP
N-L142 -----DILISNNIIIGSGATVVKSLVDPGIYIGTPCKKIIKNISDK
PglB -----QTTVGSGVTAGAGAVIVCDIPDGMTVAGNPAKPL-------
PerB -----GVTIGADTIVGAGGVVVRDLPDSVLAIGVPAKIKGDRS---
WeeI GTPDKPLIIGKGAIVGMGAVVTKSVPAGVTVVGNPARILERK----
PglD -----NLSLADDSILGGGATLVKNQDEKGVFVGVPAKRM-------

FIGURE 3

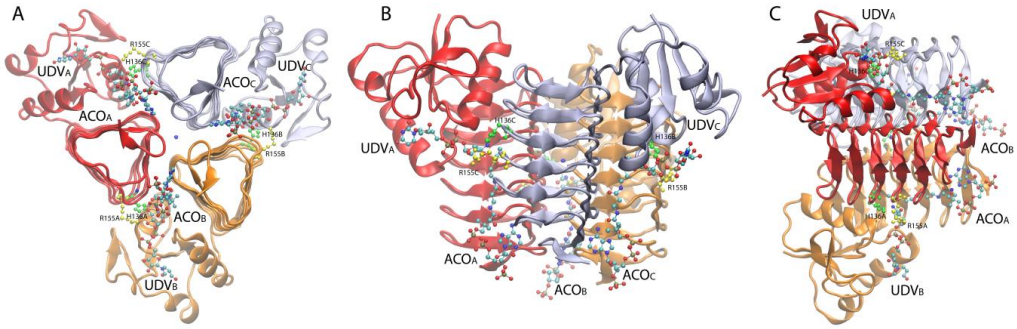

A
UDVA
ACOC
UDVC
ACOA
ACOB
UDVB
B
UDVA
UDVC
ACOA
ACOB
ACOC
C
UDVA
ACOB
ACOA
UDVB

FIGURE 4

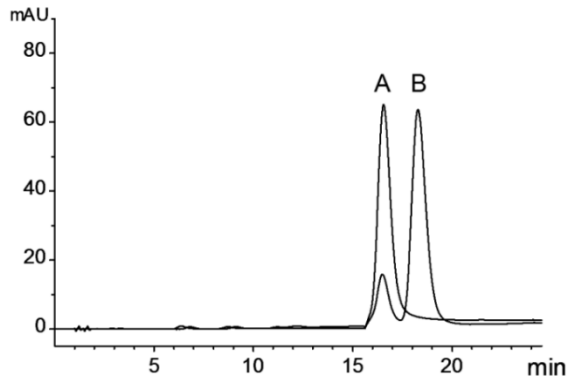


FIGURE 5

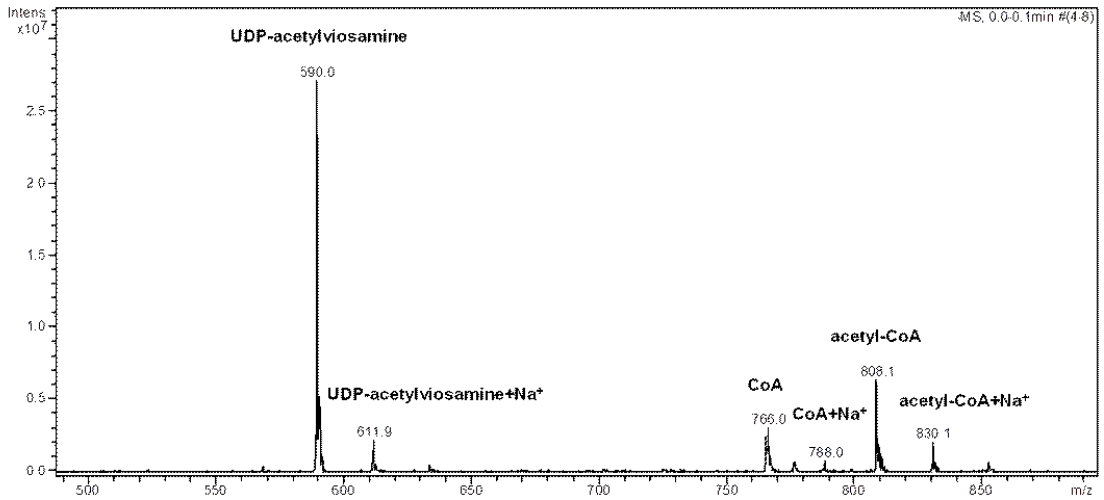

Intens
x10^7
MS, 0.0-0.1min #(4-8)
UDP-acetylviosamine
590.0
UDP-acetylviosamine+Na+
611.9
CoA
766.0
CoA+Na+
788.0
acetyl-CoA
808.1
acetyl-CoA+Na+
830.1
2.5
2.0
1.5
1.0
0.5
0.0
500
550
600
650
700
750
800
850
m/z

FIGURE 6

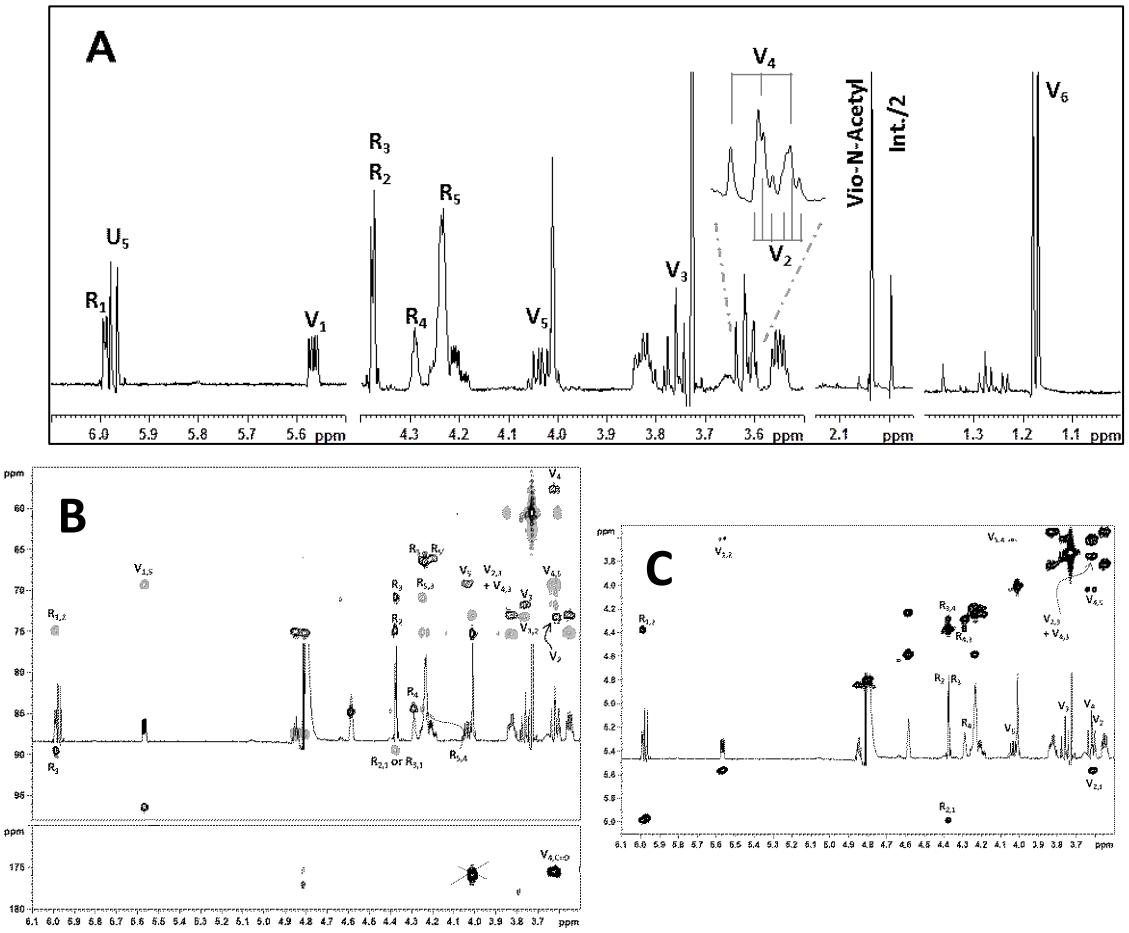
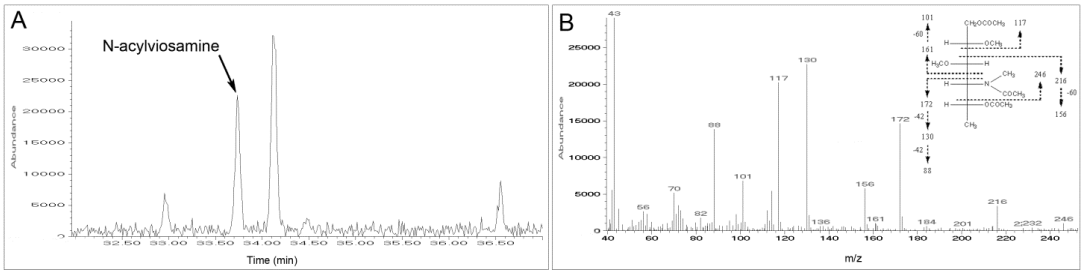
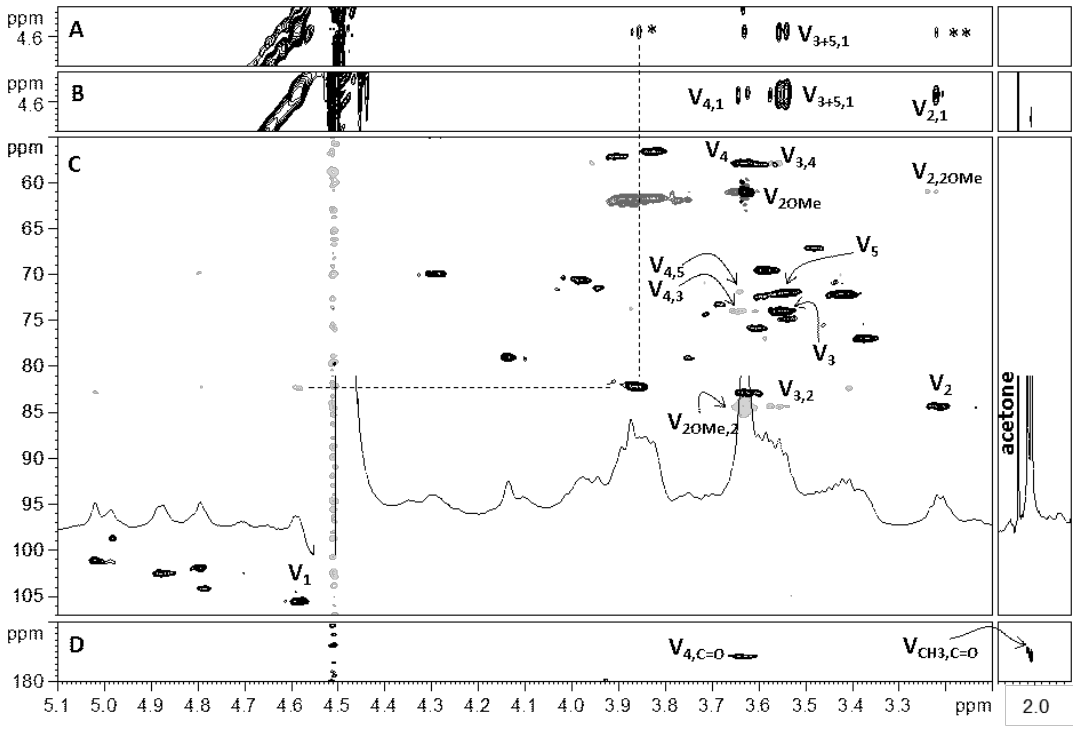


FIGURE 7



.

FIGURE 8


A
B
C
D
ppm
4.6
60
65
70
75
80
85
90
95
100
105
180
$V_{3+5,1}$
$V_{4,1}$
$V_{2,1}$
V_4
$V_{3,4}$
V_{2OMe}
$V_{2,2OMe}$
V_5
$V_{4,5}$
$V_{4,3}$
V_3
$V_{3,2}$
V_2
$V_{2OMe,2}$
V_1
acetone
$V_{4,C=O}$
$V_{CH3,C=O}$
5.1 5.0 4.9 4.8 4.7 4.6 4.5 4.4 4.3 4.2 4.1 4.0 3.9 3.8 3.7 3.6 3.5 3.4 3.3
ppm
2.0

Appendix B

PhD Course Activity Summary

1) Attended Courses:

- ✓ Tecniche di estrazione solido-liquido impiegate nella preparazione del campione per l'analisi chimica e nella produzione di estratti per usi industriali; Naviglio D.; 8 h,; University of Naples Federico II;
- ✓ Glicoscienza; Parrilli M., Bedini E.; 8 h; University of Naples Federico II;
- ✓ Produzione Ricombinante di proteine Naturali e Mutanti; Duilio A; from; 8 h; University of Naples Federico II;
- ✓ Glycomics (a complimentary course of the master degree in Chemical
- ✓ Sciences); De Castro C.; 48 hours; University of Naples Federico II;
- ✓ "Ecotoxicology: principles and main applications"; Pagano G.; 8 h; University of NaplesFederico II;
- ✓ For non French speakers: Scientific Publication: 2. Writing an article; 8 h; Aix-Marseille University;
- ✓ PF24, a course to become a teacher, doing the following exams: pedagogia scolastica (6 CFU), psicologia dell'insegnamento (6 CFU), antropologia culturale (6 CFU), didattica della biologia (6 CFU);
- ✓ Training in prevention, fire risk, biological risk, chimic risk, 1h 20 min of training, Aix-Marseille University.

2) Attended Seminars:

Title	Speaker	Place	Date
Innovative approaches for polysaccharide-based vaccines	Francesco Berti	Department of Chemical Sciences, University of Naples, Federico II	20/11/2015
Multimodal approaches for molecular imaging	L. Menichetti and M.Chiariello	Department of Chemical Sciences, University of Naples, Federico II	5/02/2016

Basics of detergents formulations and challenges	Giulia Bianchetti	Department of Chemical Sciences, University of Naples, Federico II	16/03/2016
EPDM back to basics	Martin Van Duin	Department of Chemical Sciences, University of Naples, Federico II	29/04/2016
The impotance of catalysis in the synthesis of active pharmaceutical ingredients	Michelangelo Scalone	Department of Chemical Sciences, University of Naples, Federico II	25/11/2016
Discovery novel hyperthermophilic Carbohydrate Active enZymes for plant biomass degradation	Andrea Strazzulli	Department of Biology, University of Naples, Federico II	8/03/2017
CRISPR tools to study and fight pathogenic bacteria	David Bikard	IMM, Aix Marseille University	15/09/17
Immunomodulatory antimicrobial peptides:biology and applications	Peter Haasgmann	Department of Biology, University of Naples, Federico II	15/02/2018

3) Visiting periods in institutions different from University of Naples "Federico II"

In the frame of the Italy-France Vinci project 2015, I spent an abroad period to IGS, Aix-Marseille University each year with the following dates:

First year: 12/07/2016 to 12/10/2015 (3 months)

Second year: 5/05/2017-5/11/2017 (6 months)

Third year: 26/04/2017 to 25/10/2017 (6 months)

4) Publications:

1. Kenyon JJ, Kasimova AA, Notaro A, Arbatsky NP, Speciale I, Shashkov AS, De Castro C, Hall RM, Knirel YA; (2017); Acinetobacter baumannii K13 and K73 capsular polysaccharides differ only in K-unit side branches of novel non-2-ulosonic acids: di-N-acetylated forms of either acinetaminic acid or 8-epiacinetaminic acid; Carbohydrate Research; Vol. 452; 149-155.

2. Kenyon J.; Notaro A.; Hsu Li Y.; , De Castro C., Hall R. M.; (2017); 5,7-Di-N-acetyl-8-epiacinetaminic acid: A new non-2-ulosonic acid found in the K73 capsule produced by an Acinetobacter baumannii isolate from Singapore; Scientific Reports; Vol. 7; 11357; | DOI:10.1038/s41598-017-11166-4.

3. Piacente F., De Castro C.; Jeudy S., Gaglianone M., Laugieri ME., Notaro A., Salis A., Damonte G., Abergel C., Tonetti MG., (2017); The rare sugar N-acetylated viosamine is a major component of Mimivirus fibers.; The Journal of Biological Chemistry.; Vol. 292; 7385-739.

5) Grant

Bando Star 2017 Linea 2 (Mobilità giovani ricercatori)

Grant of 5200£ for abroad period of 6 months at Aix-Marseille University to complete the study of Mimivirus fibrils.

6) Attended congresses/workshops/summer schools/contribution:

- ✓ Americal society of virology 2018 (14-18 July 2018, Maryland), oral presentation; "Understanding of glycans of Mimivirus fibrils".
- ✓ Junior Scientists Microbiology Meeting of Marseille (17 et 18 Mai 2018, Marseille), oral presentation: "Understanding of glycans of Mimivirus fibrils".
- ✓ DOC2AMU INTERDISCIPLINARY DOCTORAL DAY (30 November 2018, Salle de conférences, Campus Saint-Charles, Marseille), poster presentation: the fibrils of Mimivirus are heavly glycosylated.

- ✓ 9th European Carbohydrate Symposium EUROCARB (2-6 July 2017, Barcelona), oral and poster presentation:" The giagnt DNA virus Mimivirus exhibits heavily glycosylated fibrils surrounding its capsid".

- ✓ XV Convegno –Scuola sulla chimica dei carboidrati (19-22 June 2016, Siena), poster presentation: "The giagant DNA virus Mimivirus exhibits heavily glycosylated fibrils surrounding its capsid".