AIX-MARSEILLE UNIVERSITE
FACULTE DE MEDECINE - LA TIMONE
ECOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

THESE

GENOMES OF MIMIVIRUSES OF AMOEBA

Présentée par

Niyaz Yoosuf

En vue de l'obtention du grade de Doctorat d'Aix-Marseille Université
Spécialité: Maladies Infectieuses

Soutenue le 10 Décembre 2013

COMPOSITION DU JURY

Président du Jury	Professeur Jean-Marc Rolain
Rapporteur	Professeur Bruno Pozzetto
Rapporteur	Docteur Hervé Lecoq
Directeur de Thèse	Professeur Philippe Colson

Unité de Recherche sur les Maladies Infectieuses Tropicales et Emergentes
MR CNRS 7278, INSERM U1095, IRD 198

TABLE OF CONTENTS

Abstract

The members of families *Mimiviridae* and *Marseilleviridae*, which infect and replicate in *Acanthamoeba* spp. and other phagocytic protists, were discovered during the past decade and linked to a monophyletic group of viruses named the Nucleocytoplasmic Large DNA viruses (NCLDVs), which infect a broad variety of eukaryotes including diverse unicellular organisms. Recently, it has been proposed to reclassify the NCLDVs into a new viral order named the *Megavirales*. The discovery a decade ago of *Acanthamoeba polyphaga* mimivirus by co-culture with amoeba, with particle and genome sizes that are in the same order of magnitude than those of small bacteria, raised issues regarding evolutionary biology and fostered interest for these megaviruses and their huge and remarkable gene content that challenge the definition of viruses. Subsequently to the Mimivirus isolation, several dozens of giant viruses of amoeba have been isolated but the genome of very few has been extensively studied. We studied the genomes of these giant viruses of amoeba to gain a better understanding of their gene repertoire and evolutionary importance. The phylogenetic analysis of giant viruses of amoeba clearly distinguished three lineages, named lineages A, B and C. We studied in detail the genome of *Acanthamoeba polyphaga* moumouvirus, the leader member of lineage B to decipher its gene content and its evolutionary relationship with other organisms. We further studied the genomes of Terra1 virus and Terra2 virus, which belong to lineages C and A, respectively, and were isolated from soil samples whereas previously described mimiviruses of amoeba were isolated from fresh or marine water. Furthermore, we described the genome of Courdo11 virus, which belongs to lineage C, and is closely related to the first mimivirus isolated from a human, who exhibited unexplained pneumonia. Finally, we searched for sequences related to mimiviruses and marseilleviruses in human metagenomes to gain a better insight of their prevalence in humans and their potential pathogenicity, and obtained results indicating that these giant viruses can be present in the human body

Keywords: Mimivirus, *Mimiviridae*, *Marseilleviridae*, Nucleocytoplasmic large DNA viruses, *Megavirales*, Amoeba

Résumé

Les membres des familles *Mimiviridae* et *Marseilleviridae*, qui infectent et se répliquent dans *Acanthamoeba* spp. et d'autres protistes phagocytaires, ont été découverts au cours de la dernière décennie et rattachés à un groupe monophylétique de virus nommés les grands virus à ADN nucléocytoplasmiques (NCLDVs), qui infectent un large éventail d'eucaryotes y compris différents organismes unicellulaires. Récemment, il a été proposé de reclasser les NCLDVs dans un nouvel ordre viral nommé les *Megavirales*. La découverte, il y a une dizaine d'années de *Acanthamoeba polyphaga* mimivirus par co-culture sur amibes, avec des tailles de la particule et de son génome qui sont du même ordre de grandeur que celles de petites bactéries, a soulevé des questions dans le domaine de la biologie évolutionnaire et favorisé l'intérêt pour ces megavirus et leur énorme et remarquable répertoire de gènes qui remettent en question la définition de virus. Suite à l'isolement de Mimivirus, plusieurs dizaines de virus géants des amibes ont été isolés, mais le génome de peu d'entre eux a été étudié de façon approfondie. Nous avons étudié les génomes de ces virus géants d'amibe afin d'acquérir une meilleure compréhension de leur répertoire de gènes et leur importance évolutionnaire. L'analyse phylogénétique des virus géants d'amibe distingue clairement trois lignées, nommées A, B et C. Nous avons étudié en détail le génome de *Acanthamoeba polyphaga* moumouvirus, le membre fondateur de la lignée B et avons déchiffré son contenu en gènes et sa relation évolutive avec d'autres organismes. Nous avons également étudié les génomes de Terra1 virus et Terra2 virus, qui appartiennent respectivement aux lignées C et A, et ont été isolés à partir d'échantillons de sol alors que les mimivirus décrits aupravant ont été isolés à partir d'eau douce ou de mer. En outre, nous avons décrit le génome du virus Courdo11, qui appartient à la lignée C, et est étroitement lié au premier Mimivirus isolé d'un humain, qui présentait une pneumonie inexpliquée. Enfin, nous avons recherché des séquences de mimivirus et marseillevirus dans les métagénomes humains afin de mieux connaître la prévalence de ces virus géants chez l'homme et leur potentielle pathogénicité, et avons obtenu des résultats indiquant que ces virus géants peuvent être présents dans le corps humain.

Mot-clés: Mimivirus, *Mimiviridae*, *Marseilleviridae*, Nucleocytoplasmic large DNA viruses, *Megavirales*, Amoeba

Avant-Propos

Le format de présentation de cette thèse correspond à une recommandation de la spécialité Maladies Infectieuses et Microbiologie, à l'intérieur du Master de Sciences de la Vie et de la Santé qui dépend de l'Ecole Doctorale des Sciences de la Vie de Marseille.

Le candidat est amené à respecter des règles qui lui sont imposées et qui comportent un format de thèse utilisé dans le Nord de l'Europe permettant un meilleur rangement que les thèses traditionnelles. Par ailleurs, la partie introduction et bibliographie est remplacée par une revue envoyée dans un journal afin de permettre une évaluation extérieure de la qualité de la revue et de permettre à l'étudiant de le commencer le plus tôt possible une bibliographie exhaustive sur le domaine de cette thèse. Par ailleurs, la thèse est présentée sur article publié, accepté ou soumis associé d'un bref commentaire donnant le sens général du travail. Cette forme de présentation a paru plus en adéquation avec les exigences de la compétition internationale et permet de se concentrer sur des travaux qui bénéficieront d'une diffusion internationale.

Professor Dider Raoult

Chapter One

INTRODUCTION

Chapter One

Introduction

The Nucleocytoplasmic large DNA viruses (NCLDVs) correspond to a monophyletic group of viruses that infect animals and diverse eukaryotes including unicellular organisms. The NCLDVs include members of the families *Poxviridae*, *Asfarviridae*, *Asco-iridoviridae*, *Phycodnaviridae*, *Mimiviridae* and newly proposed *Marseilleviridae* (Iyer et al. 2001; Iyer et al. 2006; Yutin & Koonin 2012; Yutin & Koonin 2009). Recently it has been proposed to reclassify NCLDVs into a new viral order named the *Megavirales* (Colson et al. 2012). All members of the NCLDVs shares five core genes, namely the major capsid protein, helicase-primase (D5), DNA polymerase subunit family B, DNA-packaging ATPase (A32), and viral late transcription factor 3 (A2L). Moreover, 47 genes were assigned to the common ancestor of the group, although missing in some NCLDV (Yutin et al. 2009). Mimiviruses are giant viruses with particle and genome sizes that are in the same order of magnitude than those of small bacteria. In addition mimiviruses encode many genes that have not been reported earlier in viruses, in particular the multiple components of the translation system such as aminoacyl-tRNA synthetases (Arslan et al. 2011; Colson et al. 2011a; Raoult et al. 2004; Yutin & Koonin, 2012). The founding member of the family *Mimiviridae* is *Acanthamoeba polyphaga* mimivirus, discovered in 2003 from the water collected from a

cooling tower in Bradford, England (La Scola et al. 2003) In 2008, La Scola et al. reported a new strain of *Acanthamoeba polyphaga* mimivirus, isolated by co-culture with amoeba, named *Acanthamoeba castellanii* mamavirus. The further observation of Mamavirus revealed a novel virus-like agent called Sputnik that is icosahedral in shape and forms small viral particles (50 nm in size), which coexisted in the cytoplasm of the infected amoebae and inside the mamavirus factories. Sputnik only multiplied within *A. castellanii* if these cells are co-infected with mimivirus or mamavirus (La Scola et al. 2008). Since then, several mimiviruses with the capsid sizes ranging between 150-600 nm have been isolated from freshwater, saltwater and soil using the amoebal co-culture method (Boughalmi et al. 2013; La Scola et al. 2010). The genomes of the members of the order *Megavirales*, which are capable to infect a wide range of eukaryotic hosts, encompass huge and remarkable gene repertoires (Colson et al. 2012). We reviewed literature to summarize the knowledge on the composition and evolution of these gene repertoires for each of the families that compose the order *Megavirales*, and particularly depicted the core genome, genes acquired by lateral gene transfer, duplicated genes, and ORFans (**Chapter Two**).

Subsequently to the Mimivirus discovery, several new giant viruses recovered by co-culturing on amoebae have been described and phylogeny reconstructions based on highly conserved genes

delineated three lineages within mimiviruses of amoebae (referred to as A, B and C). The genomes of *Acanthamoeba polyphaga* mimivirus and *Megavirus chilensis*, the leading members of lineage A and lineage C, respectively, were described in detail (Arslan et al. 2011; Raoult et al. 2004). The *Acanthamoeba polyphaga* moumouvirus, which was isolated from cooling tower water in southeastern France, represents the leading member of lineage B.

We studied the gene repertoire of the moumouvirus to identify similarities and differences including new characteristics of this new giant virus. In addition, genomic comparisons of the members of the *Mimiviridae* showed substantial gene loss in the Moumouvirus lineage (**Chapter Three, 3.1**).

Since the discovery of Mimivirus in 2003, four new genomes of mimiviruses have been studied in detail. Among these viruses, Mimivirus, Mamavirus (another strain of mimivirus) and Moumouvirus have been isolated from water collected from cooling tower (Colson et al. 2011a; La Scola et al. 2003; La Scola et al. 2008; Raoult et al. 2004; Yoosuf et al. 2012). The two other viruses have been isolated from marine water, *Megavirus chilensis* being isolated from water collected from the coast of Chile by culturing on *Acanthamoeba* spp and *Cafeteria roenbergensis* virus (Crov) being isolated from water collected in Texas, USA from *Cafeteria roenbergensis*, a phagocytic protist (Arslan et al. 2011; Fischer et al.

2010). We described therefore the first instance of two giant viruses, Terra1 virus (lineage C) and Terra 2 virus (lineage A), recovered from soil samples by co-culturing on *Acanthamoeba* spp. (**Chapter Three, 3.2**).

We further described the gene content of a new mimivirus named Courdo11 virus. Courdo11 virus was revealed being closely related to two mimiviruses of amoeba of lineage C, LBA111 and Shan, which were isolated from the bronchoalveolar fluid and the stools, respectively, of Tunisian patients presenting pneumonia, which further emphasizes the earlier findings that mimiviruses may cause pneumonia (Colson et al. 2013; Saadi et al. 2013a; Saadi et al. 2013b; Vincent et al. 2010) (**Chapter Three, 3.3**).

Finally, the identification of mimiviruses from pneumonia patients fostered interest in the possible pathogenicity of mimiviruses. Taken together with earlier results, these recent findings indicate that these giant viruses may be causative agents of pneumonia (La Scola et al. 2005; Raoult et al. 2007). Experimentally, Mimivirus was found to be capable of inducing pneumonia in mice and infecting macrophages through a phagocytosis-like mechanism (Ghigo et al. 2008; Khan et al. 2007). We carried out analysis to search into metagenomic databases for sequences related to mimiviruses and other *Megavirales* members, we called the megavirome. Our results, added to the serendipitous detection of Mimivirus- and

Marseillevirus-like sequences in stools from an asymptomatic Senegalese man and to findings from earlier studies, indicate that mimiviruses can be present in humans (**Chapter Four**).

Chapter Two

Gene Repertoire of members of the *Megavirales* (Review)

Yoosuf N, Colson P

manuscript under preparation

Chapter Two

Gene Repertoire of members of the *Megavirales*

The Nucleocytoplasmic large DNA viruses comprises a monophyletic group of viruses (Iyer et al. 2001) the nucleo-cytoplasmic large DNA viruses (NCLDVs), a monophyletic group of viruses that encompasses members of the families *Poxviridae*, *Phycodnaviridae*, *Iridoviridae*, *Ascoviridae* and *Asfarviridae* primarily based on a limited set of core genes shared by all of these viruses (Iyer et al. 2006; Iyer et al. 2001). It has recently been proposed that the NCLDVs should be reclassified into a new viral order called "*Megavirales*" (Colson et al. 2012). These virus infects a wide range of eukaryotic hosts including green and brown algae (phycodnaviruses), various protists (mimiviruses and marseilleviruses) or *Metazoa* (poxviruses, iridoviruses, asfarviruses) (Yutin & Koonin, 2012), hence these viruses possess highly diverse gene content and the possibilities of gene exchange are numerous. The core genes remains under constant selective pressure, keeping its functional importance. The evolution of genomes mainly shaped up by two forces, gene duplications and horizontal gene transfers. Hence we summarise the gene content of the members of *Megavirlaes* based on its core genes, horizontal gene transfer genes, duplicated genes and ORFans

TITLE PAGE

Full-length title: Gene Repertoire of members of the *Megavirales*

Author list: Niyaz Yoosuf[1,2], Philippe Colson[1,2 *]

Affiliations: [1] Aix-Marseille Univ., Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE) UM63 CNRS 7278 IRD 198 INSERM U1095, facultés de Médecine et de Pharmacie, 27 boulevard Jean Moulin, 13385 Marseille cedex 05, France; [2] Fondation Institut Hospitalo-Universitaire (IHU) Méditerranée Infection, Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone, Assistance Publique – Hôpitaux de Marseille, 264 rue Saint-Pierre, 13385 Marseille cedex 05, France

*** Corresponding author:**Philippe Colson, Institut Hospitalo-Universitaire (IHU) Fondation Méditerranée Infection, Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone, Assistance Publique – Hôpitaux de Marseille, 264 rue Saint-Pierre, 13385 Marseille cedex 05, France. Tel. +33 491 385 522 ; email: philippe.colson@univ-amu.fr

Keywords: *Mimiviridae*; *Marseilleviridae*; *Phycodnaviridae*; *Poxviridae*; *Asfarviridae*; *Ascoviridae, Iridoviridae*; *Megavirales*; Nucleocytoplasmic large DNA viruses; Giant viruses, Amoeba, Eukaryotic viruses

TEXT (5,782 words)

INTRODUCTION

Nucleo-cytoplasmic large DNA viruses (NCLDV) constitute an apparently monophyletic group that was first coined in 2001 (Iyer et al. 2001) and consists of seven viral families, namely *Poxviridae*, *Ascoviridae*, *Iridoviridae*, *Asfarviridae*, *Phycodnaviridae*, *Mimiviridae* and *Marseilleviridae* infecting a broad variety of eukaryotes (Figure 1). Thus, these viruses infect a widespread range of eukaryotic hosts including green and brown algae (phycodnaviruses), various protists (mimiviruses and marseilleviruses) or *Metazoa* (poxviruses, iridoviruses, asfarviruses) (Koonin and Yutin 2010) and they either replicate exclusively in the cytoplasm of the host cells, or possess both cytoplasmic and nuclear stages in their life cycle (Moss, 2001). The NCLDVs encompass a considerably broad range of viruses that infect hosts composing a major part of the whole range of eukaryotic diversity. Besides, these viruses share a common ancestral origin as indicated by a set of ancestral genes and common virion architecture and virus reproduction within cytoplasmic factories, which support the classification of all the NCLDV families into a new viral order, named the "*Megavirales*" in reference to the large or giant size of the virions and their genomes. The gene repertoire of the *Megavirales* members encompasses several groups of genes among which core genes that are shared by all or a majority of viruses, genes transferred laterally, duplicated genes and ORFan genes. In the present review, we will summarize these gene contents.

GENE CONTENT OF THE NCLDVS

Core genes

Iyer et al. described in 2001 the monophyletic origin of members of four viral families, *Poxviridae*, *Asfarviridae*, *Iridoviridae* and *Phycodnaviridae*, and gathered them in a superfamily, the nucleocytoplasmic large DNA viruses, to encompass all these viruses based on their large size, their DNA genome and the nucleic or cytoplasmic stages observed during the viral replication cycle (Iyer et al. 2001). In 2006, this work was updated by analyzing Mimivirus, discovered in 2003, and additional genomes of iridoviruses, phycodnaviruses and poxviruses (Iyer et al. 2006). Core genes were identified for these viruses that were classified as class I when found in all families, class II when missing in some species despite being present in all families, class III when absent from one family, and class IV when absent from more than one (Iyer et al. 2006; Iyer et al. 2001). Nine genes were found to be shared by all members of all families of NCLDVs including a VV D5-type ATPase, a DNA polymerase (B-family), a VV A32 virion packaging ATPase, a VV A18 helicase, a capsid protein (D13), a thiol oxidoreductase, a VV D6/D11-like helicase, a S/T protein kinase, and a transcription factor (VLTF2). In addition, members of at least three of the four families shared 22 other core genes. Recent analyses delineated about 50 core genes in the NCLDVs (Yutin and Koonin 2012). In 2009, Yutin et al. described NCLDV clusters of orthologous groups of proteins (COGs), they named Nucleo-Cytoplasmic Virus Orthologous Groups

(NCVOGs) (Yutin et al. 2009). A total of 1,445 NCVOGs were identified among which 177 are represented in more than one NCLDV family and a set of 47 conserved genes was identified by a maximum-likelihood reconstruction, which were likely present in the genome of the common ancestor of the megaviruses. Also, five NCVOGs were identified that are shared by all the NCLDV genomes namely, the major capsid protein (orthologs of vaccinia virus D13 protein), primase-helicase (VV D5), Family B DNA polymerase (VV E9), packaging ATPase (VV A32), and transcription factor (VV A2). The majority of the core genes of the Megavirales members encode enzymes involved in DNA metabolism and replication, or structural proteins. Megavirales members therefore encode a nearly complete DNA replication apparatus in addition to key enzymes involved in the final steps of the DNA metabolism (Iyer et al. 2006; Koonin & Yutin, 2010; Yutin & Koonin, 2012; Yutin et al. 2009). The Megavirales core genes seem to have originated from different sources including homologous genes of bacteriophages, bacteria and eukaryotes, which suggests origin of these viruses at an early stage of the evolution of eukaryotes through extensive mixing of genes from widely different genomes (Koonin & Yutin, 2010; Yutin et al. 2009) and more recent analyses highlighted substantial complexity and diversity of these evolutionary scenarios (Yutin & Koonin, 2012).

Poxviruses

The poxviruses (family *Poxviridae*) are a family of double-stranded DNA (dsDNA) viruses with very large genomes (130–360 kilobase pairs (kbp) in length), usually encoding more than 150 genes per genome (Table 1) (Lefkowitz et al. 2006, Moss 2001). Poxviruses are well known for the two member viruses namely, *Variola virus* (VARV) and *Vaccinia virus* (VACV). VARV is the causative agent of smallpox, a disease that ravaged the human population until its eradication in 1977 by a worldwide vaccination campaign. Poxvirus replication occurs in the cytoplasm, thus preventing the virus from using nuclear enzymes of the host and requiring it to encode its own enzymes for DNA replication (Lefkowitz et al. 2006). The discovery of homologs of vertebrate immune system signaling molecules in the genomes of poxviruses and herpesviruses sparked the interest in studying horizontal transfer of host genes to poxviruses (Hughes & Friedman, 2005; McFadden, 1995). Many of the host-derived genes apparently hold the function of immunomodulatory genes and genes involved in nucleic acid metabolism. Viral proteins that are very identical to host genes are predicted to be functional proteins which interfere with a variety of host immune defense mechanisms including antigen display, cytokines and their receptors, cytoplasmic signaling resulting from immune activation, and genes involved in resistance of cells to oxidative stress and apoptosis. The two important studies to identify horizontal gene transfer events in poxviruses were carried out by Hughes and Friedman with a systematic search for horizontally

transferred genes by phylogenetic methods (Hughes & Friedman, 2005), while Bratke and McLysaght rather studied gene order around putative horizontally transferred genes to identify single and multiple gene events (Bratke and McLysaght 2008). They used two basic principles (a) horizontally transferred genes at conserved positions relatively to neighboring genes supports a single transfer event; (b) horizontally transferred gene at different genomic locations supports several transfer events (Bratke & McLysaght, 2008). Austin L. Hughes has done a detailed study on origin and evolution of viral interleukin-10 and other DNA virus genes with vertebrate homologues (Hughes 2002). There were cases in which the phylogenies provided strong evidence that poxvirus genes originated well prior to the origin of vertebrates, including casein kinase-related PK2 in poxviruses prior to deuterostome–protostome divergence, rpoA and rpoB prior to animal–fungus divergence. Interestingly, poxvirus proteins that originated early in the history of life include proteins playing fundamental roles in DNA replication, such as rpoA and rpoB. The presence of such proteins may have been necessary for the origin of DNA viruses themselves. Gene duplications in poxviruses were often lineage specific, and the most extensively duplicated viral gene families were found in only a few of the genomes. Twenty two gene families were present in at least one of the species of subfamily *Entomopoxvirinae* and at least one of the species of subfamily *Chordopoxvirinae*. A total of 1005 gene families were found in at

least one of the 17 poxvirus genomes, while 95 families included two or more members in at least one of the genomes (Hughes, 2002).

Ascoviruses

Ascoviridae is a family of double stranded large DNA viruses which infect insects, where they produce large enveloped virions that are 150 by 400 nm in size and cause chronic fatal disease, with cytopathology resembling that of apoptosis (Bigot et al. 1997; Federici et al. 2000). Ascoviruses have circular genomes, size ranging from 116 to 190 kbp (Table1). In ascoviruses, lateral gene transfers were identified by BLASTp that detected homologs in eukaryotic, bacterial genomes, and genomes from other megaviruses than ascoviruses and from viruses that do not belong to the proposed order Megavirales. Six open reading frames (ORFs) have been identified as of eukaryotic (Zinc-dependent metalloprotease, Unknown protein, Endonuclease, Serine/Threonine protein kinase, Hydroxysteroid (17-beta) dehydrogenase, Metallo-hydrolase) and bacterial origin (two Metallo-hydrolase, Acyl-Coenzyme A Binding Protein, BRO-like protein 12, CK1 family protein kinase, RedQ-like DEAD helicase and four ORFs from other Megavirales or non-Megavirales (IAP-like protein, Unknown protein, Ubiquitin, NTPase/helicase) (Bigot et al. 2008). The *bro* gene and *bro*-like genes were identified in viral families *Ascoviridae* and *Iridoviridae* but not in other invertebrate or vertebrate genomes, vertebrate viruses, transposons, nor in prokaryotic genomes except in prophages or bacterial transposons. The phylogenetic

analysis of *bro* genes suggested that they have resulted from the recombination of viral genomes that allowed the duplication and loss of genes and, on the same time, the acquisition of genes by horizontal transfer over evolutionary time (Bideshi et al. 2003). The common feature of many eukaryotic dsDNA viruses is the presence of multigene families. Major capsid protein is one of the genes studied in detail. In the genome of *Trichoplusia ni* (TnAV2), there are two ORFs coding the major capsid protein and the sequences shared 100 % identity, which is not common in multigene families. Likewise, thymidine kinase has two homologues and baculovirus repeated open reading frame (*bro*) had three homologues (Wang et al. 2006). The DpAV genome contains 6-8 interspersed repeated sequences of 494 bp with two imperfect palindromes and similar enhancer motifs of the ubiquitous and virus early transcription factors. These homologous regions are earlier noted in baculoviruses, which are implicated in viral DNA replication (Bigot et al. 1997). Five repeat regions were found in the entire genome of *Heliothis virescens* (HvAV3) with 94-100 % of identity among the repeats. These repeat regions code for a putative protein. The C terminus of this protein consists of a conserved transposase domain. This putative domain was conserved in most of the transposase proteins. The presence of putative transposable elements within the repeat regions indicates that DNA might have been transfered to the ascovirus genome from the host. This element may be the possible reason for the duplication of the gene in the genome (Asgari et al. 2007).

Iridoviruses

The *Iridoviridae* is a family of linear double stranded large DNA viruses (~120-200 nm) (He et al. 2002; Jakob et al. 2001; Shi et al. 2010). The genome size of iridoviruses ranges from 105 to 212 kbp (Table1). This family of viruses infects vertebrates (*Ranavirus*, *Megalocytivirus*, *Lymphocystivirus*) and invertebrate (*Iridovirus*, *Chloriridovirus*) hosts. The important characteristic of this family of viruses is its ability to infect diverse array of hosts, which likely at least partly explains the diversity of their gene content between different genera. The iridovirus genomes are circularly permuted and terminally redundant. During the co-evolution of iridoviruses and their hosts, gene gains and losses are likely to have host-specific effects. The gained genes could help evasion from host defenses while lost genes could be associated with loss of antigenic signal to the host cell immune system or the increase of virulence (Bubić et al. 2004; McLysaght et al. 2003). Horizontal gene transfers in iridoviruses that involve their hosts may have a high rate due to the nuclear stage of iridovirus DNA replication (Chinchar et al. 2009; Williams et al. 2005). Huang et al studied the gene gain and gene loss events based on the presence of clusters of orthologs genes for the 13 genomes that were sequenced (Huang et al. 2009). The phylogenetic tree based on eleven concatenated proteins indicated that gene loss could occur throughout the tree, reptile ranavirus and amphibian ranavirus (+2/-) have less gene gain-and-loss events than fish ranavirus (+50/-24), fish lymphocystivirus (+65/-26), fish megalocytivirus (+86/-19) and insect

iridovirus (+105/-). In iridoviruses, major replicative and transcription enzymes possibly originated from their eukaryotic hosts and the presence of these genes in all genera of this family indicates that the ancestral iridovirus must have acquired genes from its eukaryotic hosts and later differentiated into the five current genera. The enzyme ribonucleotide reductase that comprises small and large subunits RR-1 and RR-2, have homologs in all iridoviruses except members of the genus *Megalocytivirus,* and they are thought to be derived from *Rickettsia*–like eubacteria (Gammon et al. 2010). This enzyme plays an important role in eukaryotic DNA synthesis. In addition, the RR-1 gene from members of genus *Iridovirus* possesses an intein whereas viruses from genera *Ranavirus* and *Lymphocystivirus* do not. The presence of an intein in the RR-1 gene is very rare and noticed only in certain bacteria and phage. In contrast, megalocytiviruses encode only the RR-2 gene with low homology with those from other members of iridoviruses. Phylogenetic analysis suggests that the megalocytivirus RR-2 gene originated from a previous eukaryotic host.

Asfarviruses

African swine fever virus (ASFV) is a unique and complex pathogen that infects wild and domestic swine and members of the family *Argasidae* composed of soft-bodied ticks (Dixon et al. 1990; Lubisi, et al. 2007). The ASFV double-stranded DNA genome differs in length from about 170 to 193 kbp depending on the isolate (Table1). Due to the gain or loss of ORFs from the multigene families,

ASFV encodes between 151 and 167. Short tandem repeats are present in asfarviruses that vary in different isolates, being either located within genes or within intergenic regions, leading to small length variations in the genome The genes are distributed equally on both positive and negative strands. Multiple gene families are very common in asfarviruses, approximately 30% of paralogous genes being present in the genome, their number differing between different isolates (Agüero et al. 1990; Jones et al. 1987; Pires et al. 1997; Yozawa et al. 1994). The genes composing these familes of proteins are arranged adjacent to each other and have the same orientation, which indicates that they are evolved by gene duplication (De La Vega et al. 1994; Rodriguez et al. 1990). It also has been noted that these genes tend to be positioned at the terminal regions of the genome, their copies being distributed on the either end of the asfarvirus genome, and these genes was proposed to be transformed during genome replication and resolution. Some of these genes may have different function due to the presence of extra functional domains and large sequence divergence.

Phycodnaviruses

The phycodnaviruses are DNA viruses that infect algae (Dunigan et al. 2006). These viruses have a wide range of hosts (including algae from both marine and fresh water) and this is associated with considerable genetic diversity, though morphology is similar. The family name *Phycodnaviridae* has been quoted mainly because of two

of their characteristics: “Phyco” comes from their algal hosts and “dna” comes from their double stranded DNA genomes. Phycodnaviruses are grouped into six genera named on the basis of the viral host, namely *Chlorovirus*, *Coccolithovirus*, *Prasinovirus*, *Prymnesiovirus*, *Phaeovirus* and *Raphidovirus* (Dunigan et al. 2006; Wilson et al. 2009). These genomes have size ranging from 100 kbp to over 550 kbp (Table1) (Dunigan et al. 2006; Van Etten & Meints, 1999). Phycodnaviruses infect a wide range of hosts and hence possibilities of gene exchange are numerous. The chloroviruses encode enzymes required for the synthesis and glycosylation of structural proteins, namely two UDP-D-glucose 4,6-dehydratases (UGDs) and bifunctional UDP-4-keto-6-deoxy-D-glucose epimerase/reductase (UGER). Phylogeny showed that there was a possible recent horizontal gene transfer of UGD gene from a green algal host. At the same time, UGER was absent in *Acanthocystis turfacea* chlorella virus 1, but the host, chlorella, may encode this enzyme. Both of these genes are late genes that plays an important role in posttranslational modification of capsid proteins (Parakkottil Chothi et al. 2010). *Ostreococcus tauri* virus OtV-2 likely acquired cytochrome b5, RNA polymerase sigma factor and a high-affinity phosphate transporter encoding gene from its host, the three proteins showing a high homology with the osterococcus proteins. Moreover the genes encoding cytochrome b5, RNA polymerase sigma factor genes and four unknown functional proteins were arranged adjacent to each other in the viral genome. This may be a possible so-called “hot

spot" region in the *Ostreococcus tauri* virus2, which is more prone to harbor host genes (Weynberg et al. 2011). In addition, *Emiliania huxleyi* virus 86 (EhV-86) acquired seven genes involved in sphingolipid biosynthesis pathway from its host, microalga *Emiliania huxleyi* (Monier et al. 2009; Wilson et al. 2005). Insertion elements present in phycodnaviruses belong to bacterial and archaeal IS607 family (Frost et al. 2005). These insertion sequences do not occur between genes of bacterial origin and other genes, instead they co-localize with the stretches of bacterial-like genes, which support that they have been inherited from bacterial genomes along with bacterial-like genes (Filée et al. 2007). The number of bacterial-like genes seems to depend on the host, hosts that engulf bacteria being able to provide ecological niche for viral access to bacterial gene pools. The two phycodnaviruses ESV-1 and EHV86, which infect *Ectocarpus siliculosus* and *Emilinia huxleyi*, respectively, two free-living algae that are not learned to ingest bacteria, have very few mobile genetic elements (only two copies of IS4 family element), but other phycodnaviruses that infect *Chlorella* spp. with a symbiosis lifestyle show considerable gain of bacterial-like genes (Filée et al. 2008).

Mimiviruses

The discovery of *Acanthamoeba polyphage* mimivirus (APMV) by co-culturing with *Acanthamoeba* hosts changed dramatically the outlook of viruses because of its particle size and its gene content (La Scola et al. 2003; Raoult et al. 2004). Mimivirus genome size

ranges from 617 kb to 1,259 kbp (Table1). In Mimivirus, homologs were identified for 9/9 class I core genes (100 %), 6/8 class II core genes (75 %), 11/14 class III core genes (79 %) and 16/30 class IV core genes (53 %) (Raoult et al. 2004). Among class II core genes, Mimivirus lacks two genes which are important for the biosynthesis of 3'-deoxythymidine-5'-triphosphate: thymidylate kinase and 3'-deoxipyridine-5'-triphosphate pyrophosphatase (dUTPase), but class IV core genes thymidylate synthase and thymidine kinase have a homolog in Mimivirus. Likewise, Mimivirus misses class III core gene adenosine 5'-triphosphate (ATP)-dependent DNA ligase, which was replaced by class IV core gene nicotinamide adenine dinucleotide (NAD)-dependent ATP ligase. The Mimivirus genome is rich in nucleotide synthesis enzymes including a deoxynucleoside kinase, a cytidine deaminase and a nucleoside diphosphate kinase, reported to be the first found in a double stranded DNA virus. Raoult et al. have identified in Mimivirus several unique genes that were not previously reported in viruses, includes proteins coding for translation associated proteins, DNA repair enzymes, chaperones and new enzymatic pathways and genes that are believed being trademark genes of cellular organisms (La Scola et al. 2003; Legendre et al. 2011; Raoult et al. 2004). Since 2008, several new mimiviruses including close relatives to Mimivirus that form three lineage A, B and C (Mamavirus, Terra2 virus, Moumouvirus, Courdo11 virus, *Megavirus chilensis*) and others more distantly related (*Cafeteria roenbergensis* virus (CroV)) have been isolated from different phagocytic protists in

different niches, including fresh water, ocean and soil (Arslan et al. 2011; Fischer et al. 2010; La Scola et al. 2010; Yoosuf et al. 2012). Only 4.6% of the Mimivirus gene repertoire is composed of NCLDV core genes, which indicates that this gene content is lineage specific. Lateral gene transfer and gene duplications have also strongly influenced the composition of the Mimvirus genome (Filée et al. 2008; Iyer et al. 2006; Raoult et al. 2004). Moreira and Brochier-Armanet specifically studied a set of 198 Mimivirus proteins attributed to COG families (Moreira & Brochier-Armanet, 2008; Tatusov et al. 2003). A total of 126 ORFs with clear homologs were retrieved, the phylogenetic analyses inferring an eukaryotic origin for 60 of the 126 Mimiviral ORFs that have reliable homologs in cellular species, approximately 10% of which appeared to be acquired from amoebae. Filee et al. also identified 96 genes of bacterial origin. The bacterial-like genes show a strong bias in Mimivirus (and at least one phycodnavirus, NY2A) toward DNA replication and repair (20% of proteins) and cell envelope (12.5% of proteins) in COGs functional gene categories. Three consecutive open reading frames encoding a sugar transaminase, a glycosyltransferase, and a protein of unknown function were identified in the Mimivirus genome that are syntenic with three ORFs in the genome of *Clostridium acetobutylicum* indicating the inheritance of these bacterial-like genes as a short contiguous block; in addition, the bacterial-like genes tended to be clustered toward the extremities of the Mimivirus genomes. Furthermore, a 38-kb genomic region of

putative bacterial origin was identified in the CroV genome that encodes 34 ORFs, 14 being most similar to bacterial proteins, among which 7 are predicted to function in carbohydrate metabolism (Fischer et al. 2010). These findings further support the speculation that these genes may have been acquired from a bacterium by the frequent encounters of CroV and phagocytosed bacteria inside the host cytoplasm. These findings suggest that eukaryotic hosts using bacteria as food may work as a hotspot for the exchange of DNA between replicating viruses and bacteria, thus providing a biological niche with access to bacterial genes (Filée et al. 2007; Fischer et al. 2010; Raoult & Boyer, 2010). The Moumouvirus genome analysis revealed substantial gene loss compared to *Megavirus chiliensis*, indicating that genomes of mimivirus form this lineage experienced genome reduction. In comparison with the *Megavirus chiliensis* genome, A total of 85 genes located in the terminal regions of the *Megavirus chiliensis* genome have been apparently lost in the moumouvirus lineage; an alternative, less parsimonious evolutionary scenario would involve independent acquisition of these genes in the Mimivirus and the Megavirus lineages. Two genes encoding metabolic enzymes, cysteine dioxygenase and NAD-dependent epimerase/dehydratase, are shared by Moumouvirus and CroV to the exclusion of other *Megavirales* members (Yoosuf et al. 2012). Mobile genetic elements have been detected in the Mimivirus genome that were previously thought to be specific of prokaryotes (Filée et al. 2007). They include insertion sequences, two homing endonucleases,

and an intein, considered as major agents of lateral gene transfer in prokaryotes. The insertion sequences contain two ORFs, a transposase and a protein of unknown function (Frost et al. 2005; Ton-Hoang et al. 2005). In addition, the concurrent presence of gene typically detected in prophages of bacteria and a nearby HNH endonuclease supports the hypothesis of acquisition by lateral gene transfer from a bacteriophage (Filée et al. 2007). Duplicated genes were found to compose about one-third of the Mimivirus gene content (Suhre, 2005). Using PSI-BLAST with various e-values (1e-5 to 1e-25), 244 to 398 paralogous genes were identified that compose 58 and 86 families, respectively. Moreover, duplicated genes are inserted about twice as frequently in the parallel orientation as in the antiparallel orientation, with respect to the coding direction of the matching gene (20 vs. 12%). Large paralogous families in Mimivirus are related to virus-host interactions. Ankyrin double-helix repeat containing proteins are the most repetitive protein (66 homologs). These proteins are ubiquitously found in large paralogous families in both viral and bacterial genomes. WD repeats, L cluster, Pfam FNIP repeats and protein kinases are other paralogous proteins prevalent in Mimivirus. Also, glycosyltransferases, poxvirus transcription factors, transposase site-specific integrase-resolvases and collagen triple helix repeat containing proteins are widely present in Mimivirus genome. These proteins have wide range of functions including virus-host interactions, host signaling or other regulatory processes.

In 2008, La Scola et al. described a new strain of Mimivirus, named Mamavirus (La Scola et al. 2008). The further observation of Mamavirus revealed a novel virus-like agent called Sputnik which is icosahedral in shape and small (50 nm in size) and coexisted in the amoebal cytoplasm of the infected cells and inside the mamavirus factories. Sputnik was named a virophage, because of its functional analogy to bacteriophages, as it only multiplies within *A. castellanii* if these cells are co-infected with Mimivirus or Mamavirus. The Sputnik genome encodes a protein with homologs in a marine metagenome that belongs to the family of bacterial insertion sequence transposase DNA-binding subunits, and the Sputnik ORF 10 is closely related to integrases of the tyrosine recombinase family from archaeal viruses and proviruses. The virophage could be a vehicle mediating lateral gene transfer between giant viruses (La Scola et al. 2008). In 2011, Fischer et al. identified Mavirus, another virophage that parasitizes *Cafeteria roenbergensis* virus (Fischer & Suttle, 2011). Yau et al thereafter reported a new virophage that preys on phycodnaviruses of prasinophytes (Yau et al. 2011). Sputnik 2 was the fourth virophage described thus far and was isolated from a human-associated sample (Cohen et al. 2011). Recently Santini et al discovered another virophage infecting *Phaeocystis globosa* virus PgV-16T named as PgVV the genome of which has a length of 19,527 bp (Santini et al. 2013). It has been recently shown that the virophages of the mimiviruses have a broad host range and thus can serve as vectors for gene exchanges among the three different lineages of amoeba-

associated mimiviruses (Clarke et al. 2013; Desnues et al. 2012; Gaia et al. 2013; Yutin & Koonin, 2009). Construction in a recent study of Clusters of Mimivirus Orthologous Genes (mimiCOGs) led to reclassify Organic lake phycodnaviruses and *Phaeocystis globosa* viruses as members of the family *Mimiviridae*, though these viruses were initially classified within the family *Phycodnaviridae*, which further indicates that only viruses within the family *Mimiviridae* support so far the reproduction of virophages (Yutin et al. 2013)

Marseilleviruses

The family *Marseilleviridae* encompasses viruses with a double stranded DNA genome (Colson et al. 2013). Marseillevirus, the founding member of this family discovered in 2008 has a circular DNA (Boyer et al. 2009) while the genome of Lausannevirus, another marseillevirus described in 2011, was found to be either a linear molecule with terminal repeats or a circularized molecule (Thomas et al. 2011). So far, three genomes of *Marseilleviridae* has been sequenced and annotated, including the recently reported Cannes8 virus (Aherfi et al. 2013; Boyer et al. 2009; Thomas et al. 2011). The size of these genomes ranges from 346 kbp to 374 kbp (Table1). In the Marseillevirus genome, 28 of the 457 predicted ORFs are bona fide NCLDV core genes, out of the 41 previously defined classes I–III genes (Boyer et al. 2009). Six ORFs are universal NCLDV proteins, and 17 are shared with Mimivirus/Mamavirus but are absent in other Megavirales members. Based on phylogenetic

analysis, 51 Marseillevirus ORFs might be of NCLDV origin. As in other megaviruses, including Mimivirus, the proportion of Marseillevirus ORFs that belong to the NCLDV core gene set is very small (6.1%). All core genes reported in Marseillevirus have orthologs in Lausannevirus, including a thymidine kinase (Thomas et al. 2011). Comparative genomics and phylogenetic analysis of Marseillevirus genes have strongly highlighted the mosaicism of the Marseillevirus genome and identified gene exchange with bacteria, archaea, other viruses and eukaryotes including amoeba (Boyer et al. 2009). Interestingly, non-random connection between inferred origins and functions of marseillevirus genes was observed. Notably, genes encoding defense and repair functions, in particular nucleases, tended to be of bacterial and bacteriophage origin, genes encoding metabolic enzymes and proteins implicated in protein and lipid modification or degradation tended to be of bacterial and eukaryotic origins and genes related to signal transduction tended to be of eukaryotic origin. The Marseillevirus and Lausannevirus were found to encode three histone-like proteins (Boyer et al. 2009; Thomas et al. 2011). Histone-like proteins have been described in several viruses including H3-H4 protein in *Heliothis Zea* virus, H4 protein in bracoviruses and H2B protein in Ostreid herpesvirus integrated to amphioxus genome (Cheng et al. 2002; De Souza et al. 2010; Gad & Kim, 2008). Viral histones may interact with the host cell DNA or regulate the viral DNA. Eukaryotic organisms acquired 4 copies of histones, H2A, H2B, H3 and H4, which help to form the nucleosome and wrap the

DNA (Talbert & Henikoff, 2010). Histones are present in all archaeal phyla including the deepest branching phylum *Thaumarchaeota* (Cubonová et al. 2005; Sandman & Reeve, 2006). Ancestral marseilleviruses may haveacquired histone doublets from an unknown eukaryote (Thomas et al. 2011). MORN repeat-containing proteins, various endonucleases and serine/threonine kinases, F-box containing proteins and ubiquitins were abundantly present in members of the family *Marseilleviridae* (Aherfi et al. 2013; Boyer et al. 2009; Thomas et al. 2011). The membrane occupation and recognition nexus (MORN) repeat domains enhances membrane-membrane or membrane-cytoskeleton interactions (Gubbels et al. 2006).

ORFans in NCLDV

ORFans refers to genes without detectable homologs in sequence databases (Fischer & Eisenberg, 1999). ORFan genes have a limited phylogenetic distribution and homologous genes are either restricted to closely related organisms or not detectable at all in other organisms. Another observation is that the proportion of ORFans continues to remain same even though the number of sequenced genomes is increasing (Yin & Fischer, 2006). There are various hypothesis has been made about the origin of ORFans, some believe that ORFans are originated from genome duplication, lateral gene transfer or might correspond to de novo created genes (Daubin & Ochman, 2004; Davids et al. 2003). Several studies emphasize the fact that ORFans represents genes of viral origin. Viral genomes possess higher

proportion of ORFans compared to other microrganisms. Boyer et al. carried out a study to decipher the importance of ORFans in Megavirales families (*Ascoviridae, Iridoviridae, Poxviridae, Phycodnaviridae, Asfarviridae, Mimiviridae* and *Marseilleviridae*) (Boyer et al. 2010). At least one representative member was selected in each family, and its genome was submitted to new ORF prediction to bring normalization in viral genome prediction. A total of 38% of predicted ORFs in all viral genomes showed no match against RefSeq and were classified as ORFans. However ORFan percentage showed a large range of variation [between 2.8% (PBCV-NY2A) and 75.2% (EhV-86)] according to the type of virus (Table 2). The metaORFans are ORFans having homologs in environmental databases. MetaORFans proportions in megavirus genomes were 3.5%. The detailed number of ORFans and metaORFans in each genome of Megavirales members are represented in Table 2. Some members of families *Iridoviridae*, *Poxviridae*, *Phycodnaviridae* and *Ascoviridae*, found no significant match against the environmental databases. In contrast, more than 10% of asfarvirus ORFans were converted to metaORFans. In all megavirus genomes analyzed, mean ORFan length (587 bp) was significantly shorter than non-ORFan length (1,149 bp), indicating that ORFans are over-represented among the shorter ORFs in these genomes. Besides, ORFans and non-ORFans exhibit a similar nucleotide composition pattern.

Conclusion

Mimiviruses and marseillevirus have fostered studies on members of the proposed order *Megavirales*. As these viruses share several genes with cellular organisms, this catalyzed the debate about the definition of viruses and their classification in the living world (Raoult & Forterre, 2008; Raoult, 2009). Indeed, the tree of life was initially based on ribosomal analyses that delineated three branches of life, *Eukarya, Bacteria and Archaea, while* viruses were not included on this classification because they lack ribosomes (Moreira & López-García, 2009). From the outset, Mimivirus has been proposed to compose a fourth branch of life (Raoult et al. 2004). Then, phylogenetic and phyletic analyses of information genes, involved in nucleotide biosynthesis, transcription and translation (for Mimivirus), allowed to show a four branch topology where Megavirales members stand as a monophyletic group aside *Eukarya, Bacteria and Archaea* (Boyer et al. 2010) . This issue is still controversial but strengthened by an increasing body of evidence (Williams et al. 2011; Nasir et al. 2012). Studies of the pangenome of the order i and its viral families have shown a substantial amount of lateral gene transfers in the viral genomes, though the core gene set indicate a common ancestral origin. Finally, *Megavirales* members have considerably expanded the diversity of the viral world and the recent discovery of the largest viruses described so far, *Pandoravirus salinus and P. dulcis* (Philippe et al. 2013), shows that still amazing viruses will undoubtedly be discovered in the future.

LEGENDS

Figure 1: Phylogenetic treesconstructed based on the family B DNA polymerase from selected members of the family "*Megavirales*" and *Pandoravirus dulcis* and *P. salinus* using the maximum likelihood method

The numbers at tree nodes indicate bootstrap replicates of 100. The line indicates the group of viruses infecting diverse hosts.

Table 1: General viral characteristics of the members of the order *Megavirales*

Table 2: ORFan classification in the selected members of the order *Megavirales*

Percentages were calculated in comparison with total number of ORF for each species

Figure 1

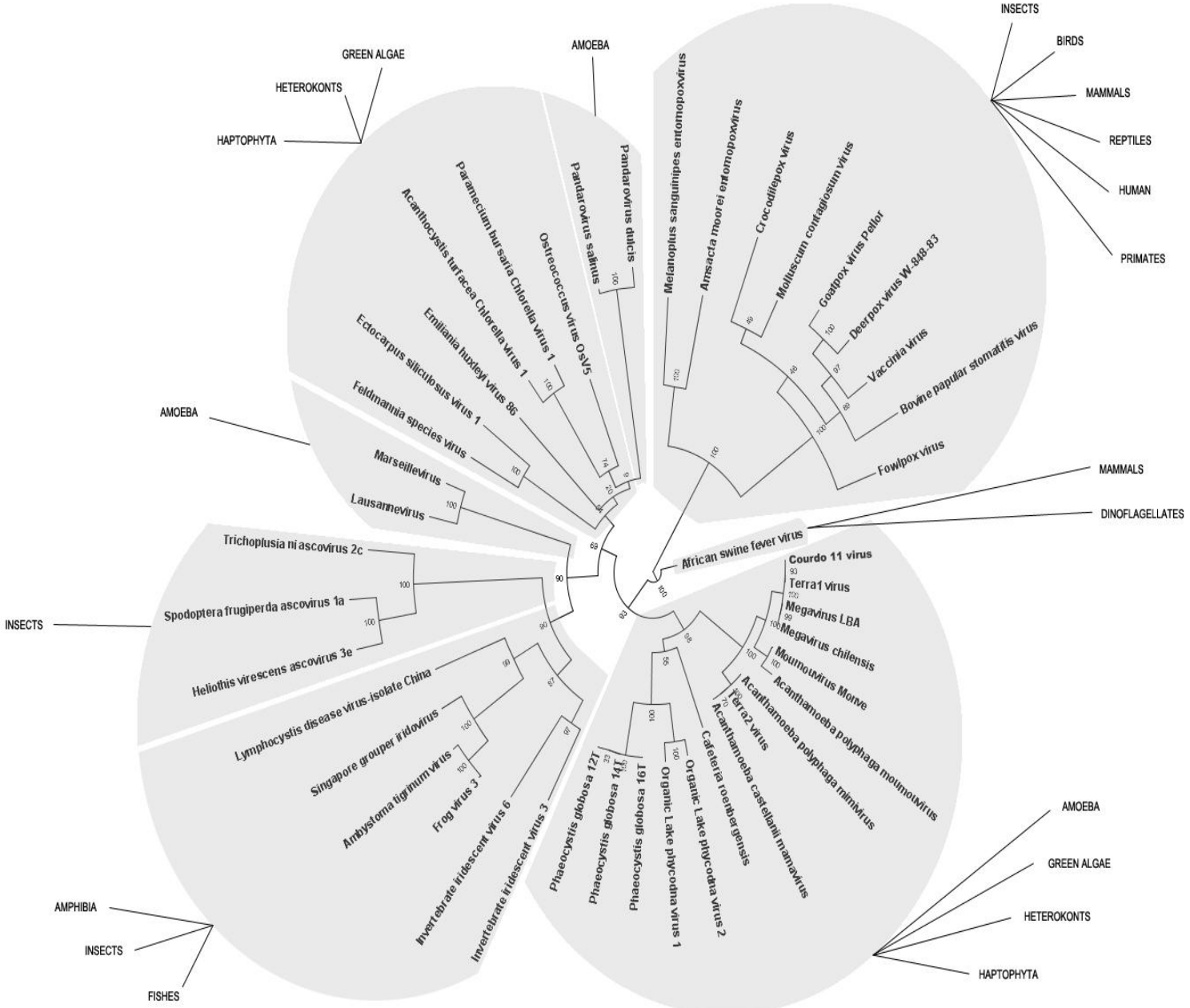

INSECTS
BIRDS
MAMMALS
REPTILES
HUMAN
PRIMATES
AMOEBA
GREEN ALGAE
HETEROKONTS
HAPTOPHYTA
Pandoravirus dulcis
Pandoravirus salinus
Ostreococcus virus OsV5
Paramecium bursaria Chlorella virus 1
Acanthocystis turfacea Chlorella virus 1
Emiliania huxleyi virus 86
Ectocarpus siliculosus virus 1
Feldmannia species virus
Marseillevirus
Lausannevirus
Melanoplus sanguinipes entomopoxvirus
Amsacta moorei entomopoxvirus
Crocodilepox virus
Molluscum contagiosum virus
Goatpox virus Pellor
Deerpox virus W-848-83
Vaccinia virus
Bovine papular stomatitis virus
Fowlpox virus
African swine fever virus
MAMMALS
DINOFLAGELLATES
Courdo 11 virus
Terra1 virus
Megavirus LBA
Megavirus chilensis
Mounouvirus Monve
Acanthamoeba polyphaga mounouvirus
Acanthamoeba polyphaga mimivirus
Terra2 virus
Acanthamoeba castellanii mamavirus
Cafeteria roenbergensis
Organic Lake phycodna virus 2
Organic Lake phycodna virus 1
Phaeocystis globosa 16T
Phaeocystis globosa 14T
Phaeocystis globosa 12T
Invertebrate iridescent virus 3
Invertebrate iridescent virus 6
Frog virus 3
Ambystoma tigrinum virus
Singapore grouper iridovirus
Lymphocystis disease virus-isolate China
Heliothis virescens ascovirus 3e
Spodoptera frugiperda ascovirus 1a
Trichoplusia ni ascovirus 2c
INSECTS
AMOEBA
GREEN ALGAE
HETEROKONTS
HAPTOPHYTA
AMPHIBIA
INSECTS
FISHES

Table 1

Virus Family	Replication site	Host range	Genome Size (bp)		No of Proteins	
			Min	Max	Min	Max
Poxviridae	Cytoplasm	Animals: insects, reptiles, birds, mammals	134,431	359,853	130	328
Asfarviridae	Cytoplasm	Mammals, dinoflagellates	170,101	182,284	151	163
Ascoviridae	Nucleus and Cytoplasm	Insects	119,343	186,262	123	180
Iridoviridae	Nucleus and Cytoplasm	Insects, Fishes, amphibia	102,653	212,482	95	468
Phycodnaviridae	Nucleus and Cytoplasm	Green algae, haptophyta, hetrokonts,	154,641	407,339	150	886
Mimiviridae	Cytoplasm	Amoeba, green algae hetrokonts, haptophyta	617,453	1,259,19	544	1120
Marseillviridae	Nucleus and Cytoplasm	Amoeba	346,754	368,454	428	444

Table 2

Megavirales	Species	No of ORFs	ORFan (%)	% MetaORFan	% Species ORFan
Poxviridae	CPV	312	31 (9.9)	-	9.9
	MSEV	224	65 (29.0)	0.9	28.1
Asfarviridae	ASFV	134	91 (67.9)	10.4	57.5
Ascoviridae	HvAV-3e	165	31 (18.8)	-	18.8
Iridoviridae	ATIV	123	45 (36.6)	4.9	31.7
	IIV-6	200	86 (43)	4.5	38.5
	SGIV	136	45 (33.1)	-	33.1
	ISKNV	111	64 (55.7)	0.9	56.8
	LDV-IC	135	8 (5.9)	-	5.9
Phycodnaviridae	EhV-86	459	345 (75.2)	6.9	70.4
	PBCV-NY2A	390	11 (2.8)	0.3	2.6
	EsV-1	238	30 (12.6)	-	12.6
	OtV-1	231	8 (3.5)	1.7	1.7
Mimiviridae	APMV	984	474 (48.1)	6.9	41.7
Marseilleviridae	MarV	449	305 (67.9)	4.8	62.1

Poxviridae: Canarypox virus (CPV), *Melanoplus sanguinipes* entomopoxvirus (MSEV) ; ***Asfarviridae***: African swine fever virus (ASFV) ; ***Ascoviridae***: *Heliothis virescens* ascovirus 3e (HvAV-3e) ; ***Iridoviridae***: *Aedes taeniorhynchus* iridescent virus (ATIV), Invertebrate iridescent virus 6 (IIV-6), Singapore grouper iridovirus (SGIV), Infectious spleen and kidney necrosis virus (ISKNV), Lymphocystis disease virus- isolate China (LDV-IC) ; ***Phycodnaviridae***: *Emiliania huxleyi* virus 86 (EhV-86), *Paramecium bursaria Chlorella* virus NY-2A (PBCV-NY2A), *Ectocarpus siliculosus* virus 1 (EsV-1), *Ostreococcus tauri* virus 1 (OtV-1) ; ***Mimiviridae***: *Acanthamoeba polyphaga* mimivirus (APMV) ; ***Marseilleviridae***: Marseillevirus (MarV)

REFERENCES

Agüero, M., Blasco, R., Wilkinson, P., Viñuela, E. (1990). Analysis of naturally occurring deletion variants of African swine fever virus: multigene family 110 is not essential for infectivity or virulence in pigs. *Virology*, *176*, 195–204.

Aherfi, S., Pagnier, I., Fournous, G., Raoult, D., La Scola, B., Colson, P. (2013). Complete genome sequence of Cannes 8 virus, a new member of the proposed family "Marseilleviridae." *Virus genes*, 47(3):550-5

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., Claverie, J. M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences*, *108*, 1–6.

Asgari, S., Davis, J., Wood, D., Wilson, P., McGrath, A. (2007). Sequence and organization of the Heliothis virescens ascovirus genome. *The Journal of general virology*, *88*, 1120–1132.

Barker, J., Brown, M. (1994). Trojan horses of the microbial world: protozoa and the survival of bacterial pathogens in the environment. *Mircobiology*, *140*(6), 1253–1259.

Bideshi, D. K., Renault, S., Stasiak, K., Federici, B. A., Bigot, Y. (2003). Phylogenetic analysis and possible function of bro-like genes, a multigene family widespread among large double-stranded DNA viruses of invertebrates and bacteria. *The Journal of general virology*, *84*, 2531–2544.

Bigot, Y, Rabouille, A., Sizaret, P. Y., Hamelin, M. H., Periquet, G. (1997). Particle and genomic characteristics of a new member of the Ascoviridae: Diadromus pulchellus ascovirus. *The Journal of general virology*, *78 (Pt 5)*, 1139–1147.

Bigot, Yves, Samain, S., Augé-Gouillou, C., Federici, B. A. (2008). Molecular evidence for the evolution of ichnoviruses from ascoviruses by symbiogenesis. *BMC Evolutionary Biology*, *8*, 253.

Boughalmi, M., Saadi, H., Pagnier, I., Colson, P., Fournous, G., Raoult, D., La Scola, B. (2013). High-throughput isolation of giant viruses of the Mimiviridae and Marseilleviridae families in the Tunisian environment. *Environmental microbiology*, *15*, 2000–7.

Boyer, M., Gimenez, G., Suzan-Monti, M., Raoult, D.(2010). Classification and determination of possible origins of ORFans through analysis of nucleocytoplasmic large DNA viruses. *Intervirology*, *53*, 310–320.

Boyer, M., Madoui, M.-A., Gimenez, G., La Scola, B., Raoult, D. (2010). Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. *PLoS ONE*, *5*, 8.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M. G, Suzan -Monti, M., La Scola, B., Koonin, E. V., Raoult, D. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 21848–21853.

Bratke, K.A., McLysaght, A. (2008). Identification of multiple independent horizontal gene transfers into poxviruses using a comparative genomics approach. *BMC Evolutionary Biology*, *8*, 67.

Bubić, I., Wagner, M., Krmpotić, A., Saulig, T., Kim, S., Yokoyama, W. M., Jonjic, S., Koszinowski, U. H. (2004). Gain of Virulence Caused by Loss of a Gene in Murine Cytomegalovirus. *Journal of Virology*, *78*, 7536–7544.

Chen, N., Li, G., Liszewski, M. K., Atkinson, J. P., Jahrling, P. B., Feng, Z., Schriewer, J., Buck, C., Wang, C., Lefkowitz, E. J., Esposito, J. J., Harms, T., Damon, I. K., Roper, R. L., Upton, C., Buller, R. M (2005). Virulence differences between monkeypox virus isolates from West Africa and the Congo basin. *Virology*, *340*, 46–63.

Cheng, C. H., Liu, S. M., Chow, T. Y., Hsiao, Y. Y., Wang, D. P., Huang, J. J., Chen, H. H. (2002). Analysis of the complete genome sequence of the Hz-1 virus suggests that it is related to members of the Baculoviridae. *Journal of Virology*, 76(18):9024-34.

Cheng, X. W., Carner, G. R., Brown, T. M. (1999). Circular configuration of the genome of ascoviruses. *The Journal of general virology*, *80 (Pt 6)*, 1537–1540.

Chinchar, V. G., Hyatt, A., Miyazaki, T., Williams, T. (2009). Family Iridoviridae: poor viral relations no longer. *Current Topics in Microbiology and Immunology*, *328*, 123–70.

Clarke, M., Lohan, A. J., Liu, B., Lagkouvardos, I., Roy, S., Zafar, N., Bertelli, C., Schilde, C., Kianianmomeni, A., Bürglin, T. R., Frech, C., Turcotte, B.,

Kopec, K. O., Synnott, J. M., Choo, C., Paponov, I., Finkler, A., Heng Tan, C. S., Hutchins, A. P., Weinmeier, T., Rattei, T., Chu, J. S., Gimenez, G., Irimia, M., Rigden, D. J., Fitzpatrick, D. A., Lorenzo-Morales, J., Bateman, A., Chiu, C. H., Tang, P., Hegemann, P., Fromm, H., Raoult, D., Greub, G., Miranda-Saavedra, D., Chen, N., Nash, P., Ginger, M. L., Horn, M., Schaap, P., Caler, L., Loftus, B. J. (2013). Genome of Acanthamoeba castellanii highlights extensive lateral gene transfer and early evolution of tyrosine kinase signaling. *Genome biology*, *14*, R11.

Cohen, G., Hoffart, L., La Scola, B., Raoult, D., Drancourt,M. (2011). Ameba-associated Keratitis, France. *Emerging infectious diseases*, *17*(7), 1306–1308.

Colson, P., De Lamballerie, X., Fournous, G., Raoult, D. (2012). Reclassification of Giant Viruses Composing a Fourth Domain of Life in the New Order Megavirales. *Intervirology*, *55*, 321–332.

Colson, P., Fancello, L., Gimenez, G., Armougom, F., Desnues, C., Fournous, G., Yoosuf, N., Million, M., La Scola, B., Raoult, D. (2013). Evidence of the megavirome in humans. *Journal of clinical virology*, *57*(3), 191–200.

Colson, P., Gimenez, G., Boyer, M., Fournous, G., Raoult, D. (2011). The Giant Cafeteria roenbergensis Virus That Infects a Widespread Marine Phagocytic Protist Is a New Member of the Fourth Domain of Life. *PLoS ONE*, *6*, 11.

Colson, P., Pagnier, I., Yoosuf, N., Fournous, G., La Scola, B., Raoult, D. (2013). "Marseilleviridae", a new family of giant viruses infecting amoebae. *Archives of virology*, *158*, 915–20.

Colson, P., Raoult, D. (2010). Gene repertoire of amoeba-associated giant viruses. *Intervirology*, *53*, 330–343.

Colson, P., Yutin, N., Shabalina, S. A., Robert, C., Fournous, G., La Scola, B., Raoult, D., Koonin, E. V. (2011). Viruses with More Than 1,000 Genes: Mamavirus, a New Acanthamoeba polyphaga mimivirus Strain, and Reannotation of Mimivirus Genes. *Genome biology and evolution*, *3*, 737–742.

Cubonová, L., Sandman, K., Hallam, S. J., Delong, E. F., Reeve, J. N. (2005). Histones in crenarchaea. *Journal Of Bacteriology*, *187*, 5482–5485.

Daubin, V., Ochman, H. (2004). Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Research*, *14*, 1036–1042.

Davids, W., Fuxelius, H. H, Andersson, S. G (2003). The journey to smORFland. *Comparative and Functional Genomics*, *4*(5), 537–541.

De La Vega, I., González, A., Blasco, R., Calvo, V., Viñuela, E. (1994). Nucleotide sequence and variability of the inverted terminal repetitions of African swine fever virus DNA. *Virology*, *201*, 152–156.

De Souza, R. F., Iyer, L. M., Aravind, L. (2010). Diversity and evolution of chromatin proteins encoded by DNA viruses. *Biochimica et Biophysica Acta*, *1799*, 302–318.

Desnues, C., Boyer, M., Raoult, D. (2012). Sputnik, a virophage infecting the viral domain of life. *Advances in virus research*, *82*, 63–89.

Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., Jardot, P., Monteil, S., Campocasso, A., Koonin, E. V., Raoult, D. (2012). Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 18078–18083.

Dixon, L. K., Bristow, C., Wilkinson, P. J.,Sumption, K. J. (1990). Identification of a variable region of the African swine fever virus genome that has undergone separate DNA rearrangements leading to expansion of minisatellite-like sequences. *Journal of Molecular Biology*, *216*, 677–688.

Dunigan, D. D., Fitzgerald, L. A.,Van Etten, J. L. (2006). Phycodnaviruses: a peek at genetic diversity. *Virus Research*, *117*, 119–132.

Federici, B. A., Bigot, Y., Hamm, J. J., Granados, R. R., Vlak, J. M. Miller, L. K. (2000). Family Ascoviridae. In Virus Taxonomy. Seventh Report of the International Committee on Taxonomy of Viruses, pp. 261–265. Edited by M. H. V. van Regenmortel, C. M. Fauquet, D. H. L. Bishop, E. B. Carstens, M. K. Estes, S. M. Lemon, J. Maniloff, M. A. Mayo, D. J. McGeoch, C. R. Pringle R. B. Wickner. San Diego: Academic Press.

Filée, J, Siguier, P., Chandler, M. (2007). I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. *Trends in genetics TIG*, *23*, 10–15.

Filée, J., Pouget, N., Chandler, M. (2008). Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evolutionary Biology*, *8*, 320.

Fischer, D., Eisenberg, D. (1999). Finding families for genomic ORFans. *Bioinformatics*, *15*(759-762).

Fischer, M. G., Allen, M. J., Wilson, W. H., Suttle, C. A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences*, *107*, 19508–13.

Fischer, M. G., Suttle, C. A. (2011). A virophage at the origin of large DNA transposons. *Science*, *332*, 231–234.

Frost, L. S., Leplae, R., Summers, A. O., Toussaint, A. (2005). Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, *3*, 722–732.

Gad, W., Kim, Y. (2008). A viral histone H4 encoded by Cotesia plutellae bracovirus inhibits haemocyte-spreading behaviour of the diamondback moth, Plutella xylostella. *The Journal of general virology*, *89*, 931–938.

Gaia, M., Pagnier, I., Campocasso, A., Fournous, G., Raoult, D., La Scola, B. (2013). Broad spectrum of Mimiviridae allows its isolation using a Mimivirus reporter. *PLoS ONE*, *8*(4).

Gammon, D. B., Gowrishankar, B., Duraffour, S., Andrei, G., Upton, C., Evans, D. H. (2010). Vaccinia Virus–Encoded Ribonucleotide Reductase Subunits Are Differentially Required for Replication and Pathogenesis. *PLoS Pathogens*, *6*, 20.

Ghigo, E., Kartenbeck, J., Lien, P., Pelkmans, L., Capo, C., Mege, J. L., Raoult, D. (2008). Ameobal pathogen mimivirus infects macrophages through phagocytosis. *PLoS pathogens*, *4*, e1000087.

Gubbels, M. J., Vaishnava, S., Boot, N., Dubremetz, J. F., Striepen, B. (2006). A MORN-repeat protein is a dynamic component of the Toxoplasma gondii cell division apparatus. *Journal of Cell Science*, *119*, 2236–2245.

Hammarlund, E., Lewis, M. W., Carter, S. V, Amanna, I., Hansen, S. G., Strelow, L. I., Wong, S. W., Yoshihara, P., Hanifin, J. M., Slifka, M. K. (2005). Multiple diagnostic techniques identify previously vaccinated individuals with protective immunity against monkeypox. *Nature Medicine*, *11*, 1005–1011.

He, J. G., Lü, L., Deng, M., He, H. H., Weng, S. P., Wang, X. H., Zhou, S.Y., Long, Q, X., Wang, X. Z., Chan, S. M. (2002). Sequence analysis of the

complete genome of an iridovirus isolated from the tiger frog. *Virology*, *292*, 185–197.

Horn, M., Wagner, M. (2004). Bacterial endosymbionts of free living Amoebae. *J.Eukaryot.Microbiol.*, *51*, 509–514.

Huang, Y., Huang, X., Liu, H., Gong, J., Ouyang, Z., Cui, H., Cao, J., Zhao, Y., Wang, X., Jiang, Y., Qin, Q. (2009). Complete sequence determination of a novel reptile iridovirus isolated from soft-shelled turtle and evolutionary analysis of Iridoviridae. *BMC Genomics*, *10*, 224.

Hughes, A. L. (2002). Origin and evolution of viral interleukin-10 and other DNA virus genes with vertebrate homologues. *Journal of Molecular Evolution*, *54*, 90–101.

Hughes, A. L., Friedman, R. (2005). Poxvirus genome evolution by gene gain and loss. *Molecular Phylogenetics and Evolution*, *35*, 186–195.

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

Iyer, L.M., Aravind, L., Koonin, E. V. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. *Journal of virology*, *75*, 11720–11734.

Jakob, N. J., Müller, K., Bahr, U., Darai, G. (2001). Analysis of the first complete DNA sequence of an invertebrate iridovirus: coding strategy of the genome of Chilo iridescent virus. *Virology*, *286*, 182–196.

Jones, E. V, Puckett, C., Moss, B. (1987). DNA-dependent RNA polymerase subunits encoded within the vaccinia virus genome. *Journal of Virology*, *61*, 1765–1771.

Khan, M., La Scola, B., Lepidi, H., Raoult, D. (2007). Pneumonia in mice inoculated experimentally with Acanthamoeba polyphaga mimivirus. *Microbial Pathogenesis*, *42*(2-3), 56–61.

Koonin, E. V., Yutin, N. (2010). Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology*, *53*, 284–292.

La Scola, B, Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E.V., Raoult, D. (2008).

The Virophage as a Unique Parasite of Giant Mimivirus. *Nature*, *455* (7209):100-4.

La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J. M., Raoult, D. (2003). A giant virus in amoebae. *Science*, 299(5615):2033.

La Scola, B., Campocasso, A., N'Dong, R., Fournous, G., Barrassi, L., Flaudrops, C., Raoult, D. (2010). Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. *Intervirology*, *53*, 344–353.

La Scola, B., Marrie, T. J., Auffray, J. P., Raoult, D. (2005). Mimivirus in pneumonia patients. *Emerging infectious diseases*, *11*, 449–452.

Lefkowitz EJ, Wang C, Upton C. (2006). Poxviruses: past, present, and future. *Virus Research*, *117*(1), 105–118.

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

Lubisi, B. A., Bastos, A. D. S., Dwarka, R. M., Vosloo, W. (2007). Intra-genotypic resolution of African swine fever viruses from an East African domestic pig cycle: a combined p72-CVR approach. *Virus Genes*, *35*, 729–735.

McFadden, G. (1995). Viroceptors, Virokines, and Related Immune Modulators Encoded by DNA Viruses. Austin: Landes.

McLysaght, A., Baldi, P. F., Gaut, B. S. (2003). Extensive gene gain associated with adaptive evolution of poxviruses. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 15655–60.

Moliner, C., Fournier, P. E., Raoult, D. (2010). Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS Microbiology Reviews*, *34*, 281–294.

Monier, A., Pagarete, A., De Vargas, C., Allen, M. J., Read, B., Claverie, J. M., Ogata, H. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Research*, *19*, 1441–1449.

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

Moreira, D., López-García, P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, *7*, 306–311.

Moss, B. (2001). Poxviridae: the viruses and their replication. In G. DE Fields BN, Knipe DM, Howley PM (Ed.), *Fields Virology* (pp. 2849–2884). Philadelphia: Williams & Wilkins.

Nasir, A., Kim, K. M., Caetano-anolles, G. (2012). Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evolutionary Biology*, *12*, 156.

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

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J.M., Abergel, C. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science 341*, 281–6.

Pires, S., Ribeiro, G., Costa, J. V. (1997). Sequence and organization of the left multigene family 110 region of the Vero-adapted L60V strain of African swine fever virus. *Virus Genes*, *15*, 271–274.
Raoult, D., Forterre, P. (2008). Redefining viruses: lessons from Mimivirus. *Nature reviews Microbiology*, *6*, 315–319.

Raoult, D. (2009). There is no such thing as a tree of life (and of course viruses are out!). *Nature Reviews Microbiology*.

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

Raoult, D., Boyer, M. (2010). Amoebae as genitors and reservoirs of giant viruses. *Intervirology*, *53*, 321–329.

Raoult, D., La Scola, B., Birtles, R. (2007). The discovery and characterization of Mimivirus, the largest known virus and putative pneumonia agent. *Clin infect Dis*, *45*, 95–102.

Reynolds, M. G., Cono, J., Curns, A., Holman, R. C., Likos, A., Regnery, R., Treadwell, T., Damon, I. (2004). Human monkeypox. *The Lancet Infectious Diseases*, *10*, 604–605,

Rodriguez, J. M., Yañez, R. J., Pan, R., Rodriguez, J. F., Salas, M. L., Viñuela, E. (1990). Multigene families in African swine fever virus: family 505. *Journal of Virology*, *68*, 2064–2072.

Rodríguez-Zaragoza, S. (1994). Ecology of free-living amoebae. *Critical reviews in microbiology*, *20*, 225–241.

Saadi, H., Pagnier, I., Colson, P., Cherif, J. K., Beji, M., Boughalmi, M., Azza, S., Armstrong, N., Robert, C., Fournous, G., La Scola, B., Raoult, D (2013). First isolation of Mimivirus in a patient with pneumonia. *Clin Infect Dis*, *4*, 127-34.

Saadi, H., Reteno Ikanga, D., Colson, P., Aherfi, S., Minodier, P., Pagnier, I., Raoult, D., La Scola, B. (2013). Shan virus, isolation of a new Mimivirus from the stool of a Tunisian patient with pneumonia. *Intervirology*, *56*, 424-9

Sandman, K., Reeve, J. N. (2006). Archaeal histones and the origin of the histone fold. *Current Opinion in Microbiology*, *9*, 520–525.

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack, K. E., Noordeloos, A. A., Brussaard, C. P., Claverie, J. M. (2013). Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proceedings of the National Academy of Sciences*, *110*, 10800-5.

Senkevich, T. G., Koonin, E. V, Bugert, J. J., Darai, G., Moss, B. (1997). The genome of molluscum contagiosum virus: analysis and comparison with other poxviruses. *Virology*, *233*, 19–42.

Shi, C. Y., Jia, K. T., Yang, B., Huang, J. (2010). Complete genome sequence of a Megalocytivirus (family Iridoviridae) associated with turbot mortality in China. *Virology Journal*, *7*, 159.

Suhre, K. (2005). Gene and genome duplication in Acanthamoeba polyphaga Mimivirus. *Journal of virology*, *79*(22), 14095–101.

Talbert, P. B., Henikoff, S. (2010). Histone variants--ancient wrap artists of the epigenome. *Nature Reviews Molecular Cell Biology*, *11*, 264–275.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, M., Mekhedov, S. L., Nikolskaya, A. N., Rao, B.S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, *4*, 41.

Thomas, V., Bertelli, C., Collyn, F., Casson, N., Telenti, A., Goesmann, A., Croxatto, A., Greub, G. (2011). Lausannevirus, a giant amoebal virus encoding histone doublets. *Environmental Microbiology*, *13*, 1454–1466.

Thomas, V., Greub, G. (2010). Amoeba/amoebal symbiont genetic transfers: lessons from giant virus neighbours. *Intervirology*, *53*, 254–267.

Ton-Hoang, B., Guynet, C., Ronning, D. R., Cointin-Marty, B., Dyda, F., Chandler, M. (2005). Transposition of ISHp608, member of an unusual family of bacterial insertion sequences. *the The European Molecular Biology Organization Journal*, *24*, 3325–3338.

Van Etten, J. L., Meints, R. H. (1999). Giant viruses infecting algae. *Annual Review of Microbiology*, *53*, 447–494.
Vincent, A., La Scola, B., Papazian, L. (2010). Advances in Mimivirus pathogenicity. *Intervirology*, *53*, 304–309.

Wang, L., Xue, J., Seaborn, C. P., Arif, B. M., Cheng, X.-W. (2006). Sequence and organization of the Trichoplusia ni ascovirus 2c (Ascoviridae) genome. *Virology*, *354*, 167–177.

Weynberg, K. D., Allen, M. J., Gilg, I. C., Scanlan, D. J., Wilson, W. H. (2011). Genome sequence of Ostreococcus tauri virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. *Journal of Virology*, *85*, 4520–4529.

Williams, T. A., Embley, T. M., Heinz, E. (2011). Informational Gene Phylogenies Do Not Support a Fourth Domain of Life for Nucleocytoplasmic Large DNA Viruses. *PLoS ONE*, *6*, 11.

Williams, T., Barbosa-Solomieu, V., Chinchar, V. G. (2005). A decade of advances in iridovirus research. *Advances in Virus Research*, *65*, 173–248.

Wilson, W. H., Schroeder, D. C., Allen, M. J., Holden, M. T. G., Parkhill, J., Barrell, B. G., Churcher, C., Hamlin, N., Mungall, K., Norbertczak, H., Quail, M.A., Price, C., Rabbinowitsch, E., Walker, D., Craigon, M., Roy, D., Ghazal, P. (2005). Complete genome sequence and lytic phase transcription profile of a Coccolithovirus. *Science*, *309*, 1090–1092.

Yau, S., Lauro, F. M., DeMaere, M. Z., Brown, M. V, Thomas, T., Raftery, M. J., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J. M., Gibson, J. A., Cavicchioli, R. (2011). Virophage control of antarctic algal host-virus dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 6163–6168.

Yin, Y., Fischer, D. (2006). On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. *BMC Evolutionary Biology*, *6*, 63.

Yoosuf, N., Yutin, N., Colson, P., Shabalina, S. A., Pagnier, I., Robert, C., Azza, S., Klose, T., Wong, J., Rossmann, M. G., La Scola, B., Raoult, D., Koonin, E. V. (2012). Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the Megavirus lineage. *Genome biology and evolution*, *4*(12), 1324–1330.

Yozawa, T., Kutish, G. F., Afonso, C. L., Lu, Z., Rock, D. L. (1994). Two novel multigene families, 530 and 300, in the terminal variable regions of African swine fever virus genome. *Virology*, *202*, 997–1002.

Yutin, N., Colson, P., Raoult, D., Koonin, E. V (2013). Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virology Journal*, *10*(106).

Yutin, N., Koonin, E. V. (2009). Evolution of DNA ligases of nucleo-cytoplasmic large DNA viruses of eukaryotes: a case of hidden complexity. *Biology direct*, *4*, 51.

Yutin, N., Koonin, E. V. (2012). Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virology journal*, *9*, 161.

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

Chapter Three (3.1)

Related Giant Viruses in Distant Locations and Different Habitats: Acanthamoeba polyphaga moumouvirus Represents a Third Lineage of the Mimiviridae That is Close to the Megavirus Lineage

Yoosuf N*, Yutin N*, Colson P*, Shabalina SA, Pagnier I, Robert C, Azza S, Klose T, Wong J, Rossmann MG, La Scola, Raoult D, Koonin EV

Genome Biol. Evol. 4(12):1324–1330

* equal contribution

Chapter Three

Related Giant Viruses in Distant Locations and Different Habitats: *Acanthamoeba polyphaga* moumouvirus Represents a Third Lineage of the *Mimiviridae* That is Close to the Megavirus Lineage

The founding member of the family *Mimiviridae*, *Acanthamoeba polyphaga* mimivirus, was discovered in 2003 from a cooling tower in Bradford, England subsequently to the investigation of a pneumonia outbreak (La Scola et al. 2003; Raoult et al. 2004). Since then, the genomes of three other members of the family Mimiviridae have been extensively described (Arslan et al. 2011; Colson et al. 2011a; Fischer et al. 2010). The phylogenetic analysis of core genes of the mimiviruses infected *Acanthamoeba* spp. indicated three lineages. The *Acanthamoeba polyphaga* mimivirus, the leading member of lineage A, and *Megavirus chilensis*, the leader member of lineage C, were studied in detail (Arslan et al. 2011; Raoult et al. 2004). In contrast, no genome had been described in detail for mimiviruses of lineage B, until we studied the DNA of *Acanthamoeba polyphaga* moumouvirus, which has a size of 1,021,348 base pairs. This giant virus was isolated using co-culture

with amoebae from water of a cooling tower in southeastern France. The Moumouvirus genome encodes 930 proteins and three transfer RNAs, being the fourth largest viral genome reported so far at time of our analyses. Among the predicted proteins, 75% had close homologs in *Megavirus chilensis*. The Megavirus and Moumouvirus genomes showed a perfect collinearity in their central part and variations at the extremities. Moreover, comparison of mimivirus genomes showed substantial gene loss in the Moumouvirus lineage. The majority of the proteins of Moumouvirus had the closest homologs in other members of the *Mimiviridae*, while 27 genes had their closest homolog in bacteria. Further analysis of these genes based on phylogeny supported gene acquisition from diverse bacteria after the separation of the Moumouvirus and Megavirus lineages.

Related Giant Viruses in Distant Locations and Different Habitats: *Acanthamoeba polyphaga moumouvirus* Represents a Third Lineage of the *Mimiviridae* That Is Close to the *Megavirus* Lineage

Niyaz Yoosuf[1,†], Natalya Yutin[2,†], Philippe Colson[1,3,†], Svetlana A. Shabalina[2], Isabelle Pagnier[1], Catherine Robert[1], Said Azza[1], Thomas Klose[4], Jimson Wong[4], Michael G. Rossmann[4], Bernard La Scola[1,3], Didier Raoult[1,3], and Eugene V. Koonin[2,*]

[1]Aix-Marseille University, URMITE, Faculté de Médecine et de Pharmacie, Marseille, France

[2]National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, Maryland

[3]Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, IHU Méditerranée Infection, Assistance Publique-Hôpitaux de Marseille, Centre Hospitalo-Universitaire Timone, Marseille, France

[4]Department of Biological Sciences, Purdue University

[†]The authors contributed equally to this work.

*Corresponding author: E-mail: koonin@ncbi.nlm.nih.gov.

Accepted: November 23, 2012

Data deposition: Moumouvirus genome sequence has been deposited in GenBank under the accession number JX962719.

Abstract

The 1,021,348 base pair genome sequence of the *Acanthamoeba polyphaga moumouvirus*, a new member of the *Mimiviridae* family infecting *Acanthamoeba polyphaga*, is reported. The moumouvirus represents a third lineage beside mimivirus and megavirus. Thereby, it is a new member of the recently proposed *Megavirales* order. This giant virus was isolated from a cooling tower water in southeastern France but is most closely related to *Megavirus chiliensis,* which was isolated from ocean water off the coast of Chile. The moumouvirus is predicted to encode 930 proteins, of which 879 have detectable homologs. Among these predicted proteins, for 702 the closest homolog was detected in *Megavirus chiliensis*, with the median amino acid sequence identity of 62%. The evolutionary affinity of moumouvirus and megavirus was further supported by phylogenetic tree analysis of conserved genes. The moumouvirus and megavirus genomes share near perfect orthologous gene collinearity in the central part of the genome, with the variations concentrated in the terminal regions. In addition, genomic comparisons of the *Mimiviridae* reveal substantial gene loss in the moumouvirus lineage. The majority of the remaining moumouvirus proteins are most similar to homologs from other *Mimiviridae* members, and for 27 genes the closest homolog was found in bacteria. Phylogenetic analysis of these genes supported gene acquisition from diverse bacteria after the separation of the moumouvirus and megavirus lineages. Comparative genome analysis of the three lineages of the *Mimiviridae* revealed significant mobility of Group I self-splicing introns, with the highest intron content observed in the moumouvirus genome.

Key words: moumouvirus, mimivirus, giant virus, megavirus, *Mimiviridae*, *Megavirales*, horizontal gene transfer, viral genome, nucleo-cytoplasmic large DNA viruses.

The family *Mimiviridae* consists of giant viruses that together with five previously recognized viral families and the candidate Marseilleviridae family comprise a monophyletic group of viruses known as nucleo-cytoplasmic large DNA viruses (NCLDV) (Iyer et al. 2001, 2006; Yutin and Koonin 2012). Recently, it has been proposed to combine all the NCLDV families into a new virus order tentatively named the *Megavirales* (Colson et al. 2012). The family *Mimiviridae* includes by far the largest viral genomes sequenced to date (La Scola et al. 2003; Claverie et al. 2009; Claverie and Abergel 2010). This is the only group of viruses with genomes larger than 1 megabase, which exceeds the genome size of numerous parasitic and symbiotic bacteria. The genomes of three *Mimiviridae* members have been completely sequenced

and characterized in detail: *Acanthamoeba polyphaga mimivirus* (Raoult et al. 2004), the prototype of the family; *A. castellanii mamavirus,* which is a close relative, effectively a strain of the mimivirus (Colson et al. 2011); and *Megavirus chiliensis* that has been recently isolated from a marine environment (Arslan et al. 2011). In addition, 16 virus isolates of the family *Mimiviridae* have been identified and characterized by proteomic methods and/or partial sequencing (La Scola et al. 2010). Furthermore, marine metagenome analysis has revealed numerous homologs of mimivirus genes indicating that *Mimiviridae* is an abundant and diverse family of giant viruses whose host range remains unknown but includes organisms from habitats as different as marine water, fresh water, and soil (Monier et al. 2008; Kristensen et al. 2010; Yamada 2011).

Phylogenetic analysis of genes that are conserved in the majority of the NCLDV (Iyer, Aravind, et al. 2001; Iyer, Balaji, et al. 2006; Koonin and Yutin 2010) has shown that another giant virus that has been isolated from the marine microflagellate *Cafeteria roenbergensis* (CroV) is a distant member of the *Mimiviridae* (Fischer et al. 2010; Colson et al. 2012). In addition to genes that are shared with the other NCLDV, the giant viruses of the *Mimiviridae* family possess many genes that have not been previously detected in any viruses, in particular genes encoding components of the translation system such as aminoacyl-tRNA synthetases as well as a variety of metabolic enzymes (Raoult et al. 2004; Colson and Raoult 2010). The comparison of the mimivirus/mamavirus and the megavirus genomes has shown that only 77% of the megavirus proteins have readily detectable homologs in the mimivirus, suggestive of a large pangenome of the *Mimiviridae* (Arslan et al. 2011). Clearly, additional complete genomes of diverse members of the *Mimiviridae* are required for the characterization of this pangenome. Here, we describe the genome of another member of the *Mimiviridae* that we denoted *A. polyphaga moumouvirus*. The moumouvirus was isolated from water collected in a cooling tower but perhaps unexpectedly is most closely related to the megavirus that was identified in a marine environment.

The moumouvirus was isolated in February 2008 by inoculating *A. polyphaga*, as previously described, with water from an industrial cooling tower located in the south-east of France (La Scola et al. 2008). Some features of this virus have been briefly described previously (La Scola et al. 2010). Morphologically, the moumouvirus particles resemble the particles of other *Mimiviridae* (Klose et al. 2010; Arslan et al. 2011). The icosahedral capsid is approximately 420 nm in size and is covered by a dense layer of fibers (fig. 1). In comparison, *A. polyphaga mimivirus* and *Megavirus chiliensis* exhibit larger capsids with a diameter of approximately 500 and 520 nm, respectively (Klose et al. 2010; Arslan et al. 2011). In addition, these two mimiviruses harbor fibers that are approximately 125- and 75-nm long, respectively, whereas the size of the moumouvirus fibers is approximately 100 nm. Some of the

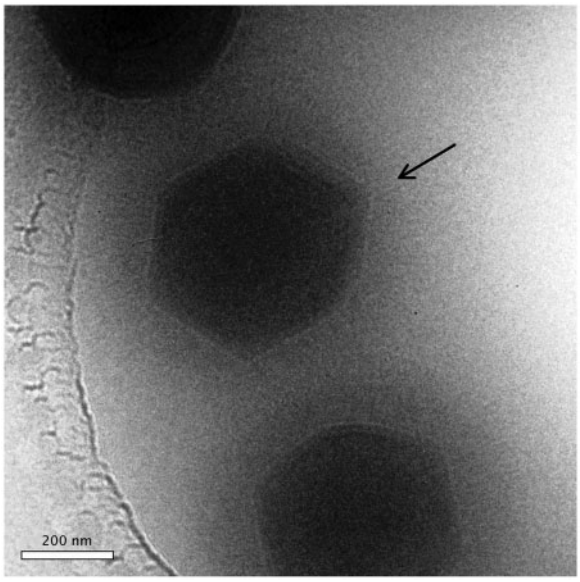


Fig. 1.—Cryo-electron micrograph of moumouvirus particles. The viral particles have a dense layer of fibers and their morphology resembles the shape of other *Mimiviridae* members, including a distinctive, starfish like vertex (arrow). Scale bar: 200 nm.

moumouvirus particles also exhibit, similar to other *Mimiviridae* members, a distinctive, starfish-like vertex (Klose et al. 2010; Arslan et al. 2011). Finally, viral factories were observed within the *A. polyphaga* cytoplasm during the replication cycle of the moumouvirus; the morphology of the moumouvirus factories is similar to that observed previously for *A. polyphaga mimivirus* and *Megavirus chiliensis* (Suzan-Monti et al. 2007; Arslan et al. 2011).

The moumouvirus genome DNA was sequenced using the 454-Roche GS20 device (Roche Diagnostics Corp., Branford, CT) (Raoult et al. 2004; Margulies et al. 2005) and then the AB SOLiD instrument (Life Technologies Corp., Carlsbad, CA). The genome assembly was performed using a combination of Roche 454 paired-end and AB SOLiD sequencing reads (supplementary methods, Supplementary Material online). The moumouvirus genome is 1,021,348 base pairs (bp) in length which is more than 200 kilobase (kb) shorter than the megavirus genome (the current record holder in viral genome size) and more than 100 kb shorter than the mimivirus and mamavirus genomes (the moumouvirus genome sequence was deposited in GenBank under the Accession Number JX962719). Using pulse-field gel electrophoresis, the moumouvirus genome was characterized as a linear DNA molecule of approximately 1 megabase (not shown). Using a combination of prediction tools (supplementary methods and file 1, Supplementary Material online), 930 open reading frames (ORFs) were identified as putative protein-coding genes, with the mean predicted protein size of 290 amino acids (aa). These ORFs are evenly distributed on both DNA strands,

with 470 predicted genes located on the "direct" strand and 460 on the "reverse" strand. The mean size of intergenic regions is 130 ± 166 nucleotides, with the predicted protein-coding density of 0.91 genes/kb (as compared with 0.89 genes/kb for the megavirus). In addition, three tRNA genes were predicted using the tRNAscan-SE method (Schattner et al. 2005). The ORFs were analyzed for evolutionary conservation, protein domain content and predicted functions by using PSI-BLAST search (Altschul et al. 1997) of the Refseq database at the NCBI (one iteration by default and up to 3 iterations when initial functional prediction was ambiguous), domain identification by RPS-BLAST search of the Conserved Domain Database (Marchler-Bauer and Bryant 2004), and assignment of proteins to clusters of orthologous NCLDV genes (NCVOGs) (Yutin et al. 2009).

Of the 930 predicted proteins of the moumouvirus, for 879 homologs were detected by protein sequence database search, and for the great majority, the most similar homolog was a megavirus protein (supplementary file 1, Supplementary Material online). For 656 predicted proteins of the moumouvirus, the megavirus homolog was a bidirectional best hit (BBH) (with the expect value cut-off of 10^{-3}); that is, a probable ortholog. The putative moumouvirus–megavirus orthologous protein pairs ranged in identity from 91% to 23%, with a median of 62%. An analogous comparison between moumouvirus and mimivirus yielded 548 putative orthologs, with a median 52% identity, indicating that the moumouvirus is more similar to the megavirus than it is to the mimivirus in terms of both the gene repertoire and sequence conservation. The evolutionary affinity of the moumouvirus and the megavirus was clearly supported by the results of phylogenetic analysis of concatenated conserved NCLDV proteins (fig. 2 and supplementary file 2, Supplementary Material online). Although the trees for individual conserved genes showed topological differences for other branches within *Phycodnaviridae* and *Mimiviridae*, the moumouvirus–megavirus clade and its monophyly with the mimivirus–mamavirus clade were invariably recovered (supplementary file 2, Supplementary Material online). In addition, a genomic dot-plot of the moumouvirus against the megavirus reveals near perfect collinearity of orthologous genes in the middle part of the genome (~650 kb), with rearrangements found only in the peripheral parts of the genomes (fig. 3*A*). This similarity of genome architectures contrasts the results of the comparison of the moumouvirus and mimivirus genomes that shows shorter, interrupted collinear regions and in addition a large inversion in the central part of the genomes (fig. 3*B*), similar to that described from the comparison of the megavirus and mimivirus genomes (Arslan et al. 2011). Conservation of the gene order in the middle of the genome with divergence at the genome ends seems to be a general feature of NCLDV evolution that was first noted in poxviruses (Senkevich et al. 1997) and has been more recently pointed out for *Chlorella phycodnaviruses* (Filee et al. 2007), marseillevirus and lausannevirus

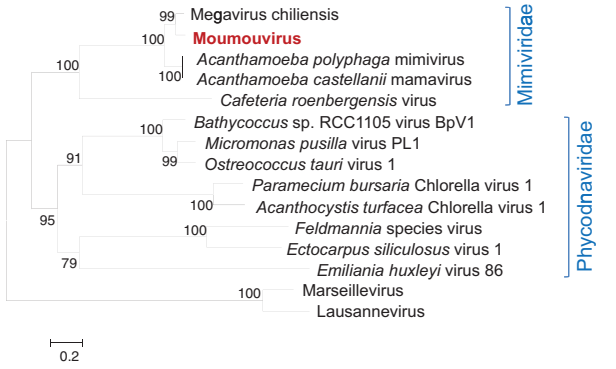


Fig. 2.—Phylogenetic tree of the *Mimiviridae* and selected *Phycodnaviridae* constructed from concatenated alignments of DNA polymerase, A32-like packaging ATPase, and A2-like Transcription Factor. Marseillevirus and lausannevirus were used as an outgroup. The alignment included 1,429 positions that were deemed reliably aligned. The bootstrap values (percentage points) are indicated for each internal branch (*Mimiviridae*).

(Thomas et al. 2011), mamavirus, mimivirus, and CroV (Boyer et al. 2011; Colson, Gimenez, et al. 2011; Colson et al. 2011).

Together, these observations indicate that megavirus and moumouvirus comprise a distinct branch of the *Mimiviridae*. Given the moderate sequence conservation between the orthologs and differences in the gene repertoire (discussed later), moumouvirus and megavirus clearly are distinct virus species unlike mimivirus and mamavirus that, at >98% mean identity between orthologous proteins and near perfect genomic collinearity, are most appropriately considered strains of the same species (Colson et al. 2011). These findings are in line with the recent demonstration of three evolutionary lineages within the *Mimiviridae* (Colson et al. 2012).

Although moumouvirus is the sister group of megavirus in the phylogenetic tree of the *Mimiviridae* (fig. 2), its genome is more than 200 kb smaller than the megavirus genome. Of the 1,120 predicted protein-coding genes of the megavirus (Arslan et al. 2011), for 464 no one-on-one ortholog has been detected in the moumouvirus. Analysis of these megavirus proteins showed that 219 are members of paralogous families common for *Mimiviridae* members; 139 are ORFans without detectable homologs; 21 apparently were acquired from sources outside the *Mimiviridae* (mainly from bacteria); and 85 are shared by megavirus and mimivirus/mamavirus but absent in the moumouvirus (supplementary file 3, Supplementary Material online). Thus, these 85 genes that are located in the terminal regions of the genome apparently have been lost in the moumouvirus lineage; an alternative, less parsimonious evolutionary scenario would involve independent acquisition of these genes in the mimivirus and megavirus lineages. Most of the genes that are inferred to have been lost by the moumouvirus are functionally uncharacterized but for some functions could be predicted, in particular in DNA repair (supplementary file 3, Supplementary Material online).

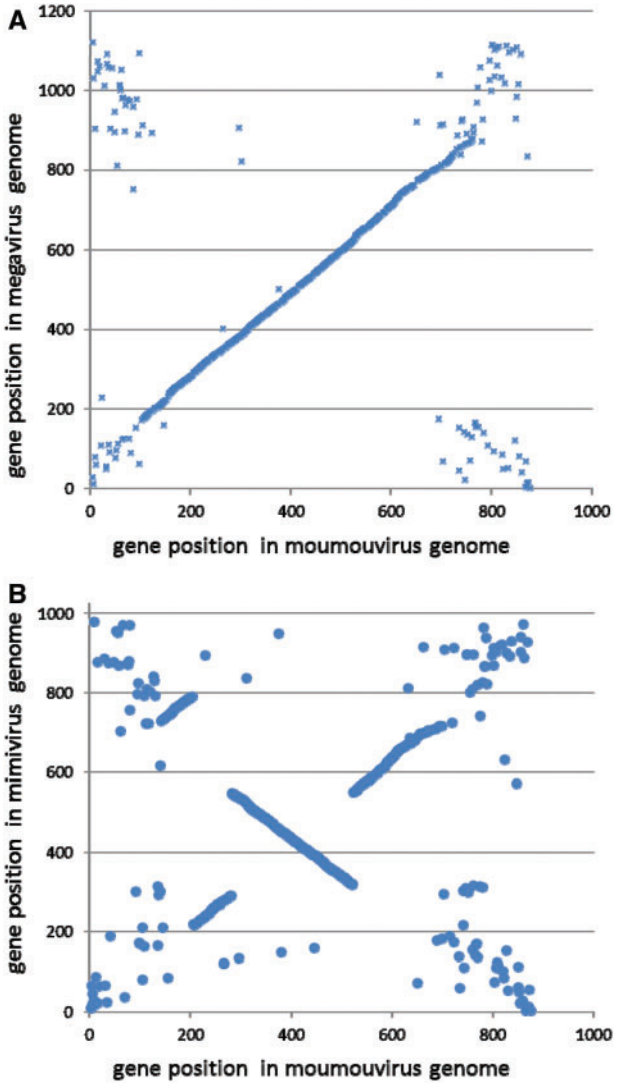


Fig. 3.—Genomic dot plots for the moumouvirus and other members of the *Mimiviridae*. (*A*) Moumouvirus versus *Megavirus chiliensis*. (*B*) Moumouvirus versus *Acanthamoeba polyphaga mimivirus*. Each point represents a pair of orthologous genes (BBHs in BLASTP searches).

Interestingly, one of the lost genes encodes the small polyA polymerase subunit/cap *O*-methyltransferase, a gene that is shared by megavirus, mamavirus, and poxviruses but is missing in the rest of the NCLDV, suggestive of multiple losses (Colson et al. 2011). The demonstration of extensive gene loss in the moumouvirus echoes the dramatic reduction in the mimivirus genome size after cultivation in germ-free amoeba; notably, the size of the terminal regions that have been eliminated from the mimivirus genome after multiple passages is approximately the same (~200 kb) as the difference in genome size between megavirus and moumouvirus (Boyer et al. 2011).

In addition to being the largest viral genome sequenced to date, the megavirus is notable for encoding the largest number of translation system components among all viruses including 7 aminoacyl-tRNA synthetases (aaRS) (Arslan et al. 2011). The moumouvirus encodes apparent orthologs of many but not all of these proteins, in particular 5 aaRS (supplementary files 1 and 3, Supplementary Material online).

The majority of the moumouvirus protein-coding genes have apparent conserved orthologs in the megavirus but the remaining genes showed some interesting evolutionary patterns. Two genes encoding metabolic enzymes, cysteine dioxygenase and NAD-dependent epimerase/dehydratase, are shared by moumouvirus and CroV to the exclusion of the other NCLDV. Phylogenetic analysis of both genes demonstrated monophyly of the two giant viruses, along with some uncharacterized environmental sequences (fig. 4*A* and *B*). This phylogeny implies the presence of these genes in the common ancestor of the *Mimiviridae* with at least two subsequent losses (in the mimivirus and megavirus branches) or the less likely evolutionary scenarios involving gene exchange between moumouvirus and CroV or independent acquisition of genes from related sources by the two viruses. Phylogenetic analysis of the moumouvirus genes with closest bacterial homologs supported the origin of these genes from diverse bacteria (two examples are shown in fig. 4*C* and *D*), in agreement with the previously noticed extensive gene exchange among symbionts and parasites of amoeba (Ogata et al. 2006; Moreira and Brochier-Armanet 2008; Boyer et al. 2009; Raoult and Boyer 2010).

Similarly to other *Mimiviridae* members, moumouvirus genes were found to contain 8 Group I self-splicing introns and three inteins (supplementary file 4, Supplementary Material online). All sequenced members of the *Mimiviridae* share an apparent ancestral intron in the gene for the largest subunit of the RNA polymerase (RNAP) and an ancestral intein in the DNA polymerase gene. The moumouvirus contains only a single intron in the major capsid protein gene, similar to the mamavirus, whereas the megavirus and the mimivirus contain two introns in this gene. In the HSP70 chaperone gene, megavirus and moumouvirus share an intron to the exclusion of the other *Mimiviridae* members but moumouvirus lacks the intron-encoded endonuclease ORF. In addition, the moumouvirus contains short inteins that consist of the cis-acting HINT protease domain alone in the genes encoding the repair ATPase MutS and an uncharacterized protein (supplementary file 4, Supplementary Material online). The positions of other introns in the genomes of the *Mimiviridae* members vary and several new introns were identified. In particular, unlike other *Mimiviridae* members, the moumouvirus contains the largest number of introns (5), along with an intein, in the second largest RNAP subunit gene (supplementary file 4, Supplementary Material online). These findings emphasize the dynamic evolution of the introns and inteins in the *Mimiviridae*.

Analysis of the mimivirus and megavirus genomes has revealed two distinct features of the transcripts, namely the

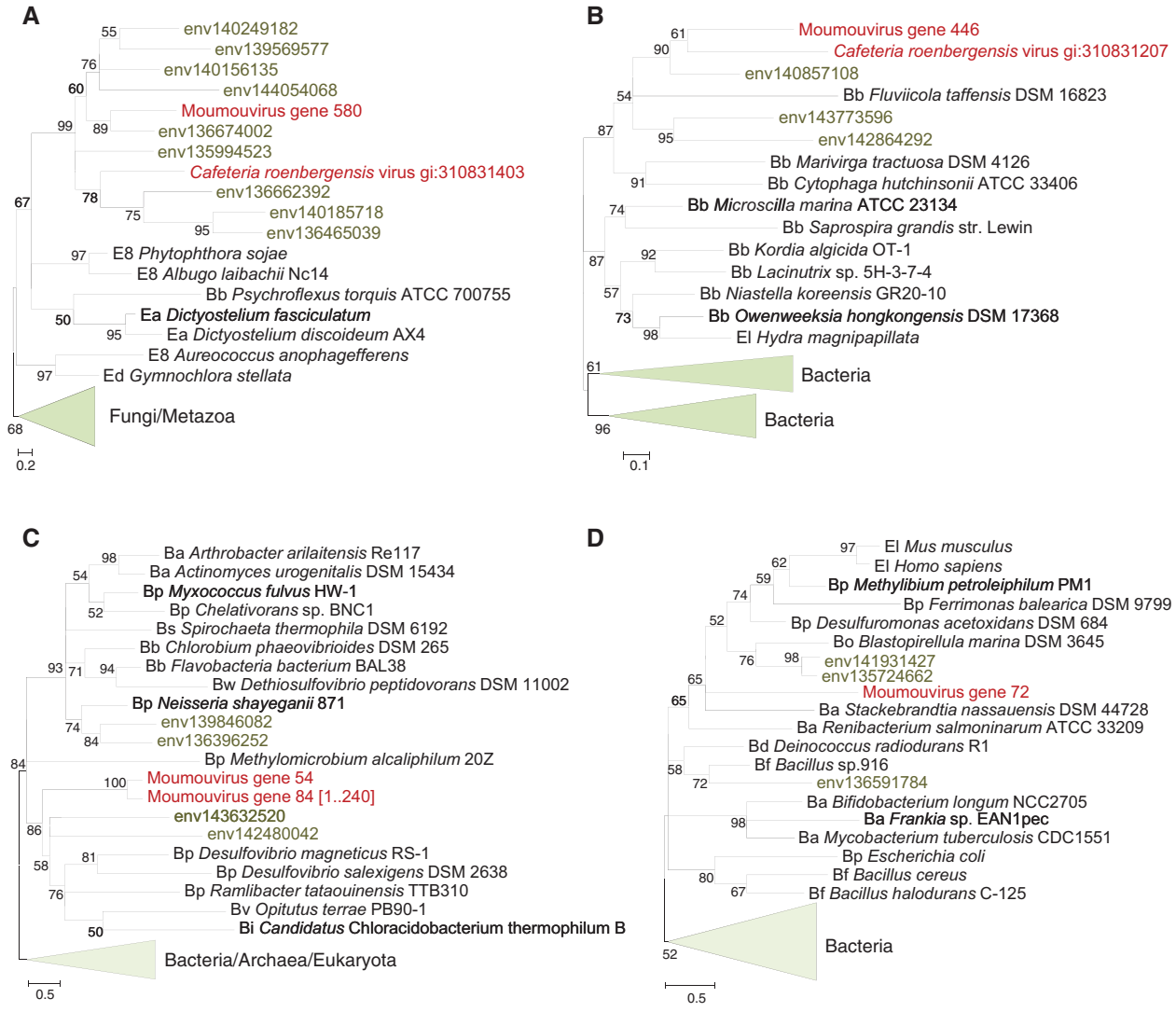


Fig. 4.—Phylogenetic trees of moumouvirus genes missing in other members of the *Mimiviridae*. (*A*) Cysteine dioxygenase, (*B*) NAD-dependent epimerase/dehydratase, (*C*) Methyltransferase, and (*D*) Nudix Hydrolase. The bootstrap values (percentage points) are indicated for each internal branch. Each sequence is denoted by taxa abbreviation and species name. GenBank Identification (GI) numbers for *Cafeteria roenbergensis* virus are shown on the trees. For other sequences, GI numbers are as follows: *Phytophthora sojae*, 348690046; *Albugo laibachii* Nc14, 325183169; *Psychroflexus torquis* ATCC 700755, 91214785; *Dictyostelium fasciculatum*, 328865331; *Dictyostelium discoideum* AX4, 66806929; *Aureococcus anophagefferens*, 323447704; *Gymnochlora stellate*, 193875832; *Fluviicola taffensis* DSM 16823, 327404160; *Marivirga tractuosa* DSM 4126, 313675758; *Cytophaga hutchinsonii* ATCC 33406, 110637252; *Microscilla marina* ATCC 23134, 124004204; *Saprospira grandis* str. Lewin, 379730028; *Kordia algicida* OT-1, 163754599; *Lacinutrix sp.* 5H-3-7-4, 336171271; *Niastella koreensis* GR20-10, 375148557; *wenweeksia hongkongensis* DSM 17368, 375013912; *Hydra magnipapillata*, 221105922; *Arthrobacter arilaitensis* Re117, 308177223; *Actinomyces urogenitalis* DSM 15434, 227495863; *Myxococcus fulvus* HW-1, 338534578; *Chelativorans sp.* BNC1, 110632790; *Spirochaeta thermophila* DSM 6192, 307718496; *Chlorobium phaeovibrioides* DSM 265, 145219574; *Flavobacteria bacterium* BAL38, 126664257; *Neisseria shayeganii* 871, 349575093; *Dethiosulfovibrio peptidovorans* DSM 11002, 288575048; *Methylomicrobium alcaliphilum* 20Z, 357407184; *Desulfovibrio magneticus* RS-1, 239906771; *Desulfovibrio salexigens* DSM 2638, 242280667; *Ramlibacter tataouinensis* TTB310, 337278313; *Opitutus terrae* PB90-1, 182415517; *Candidatus Chloracidobacterium thermophilum B*, 347756407; *Mus musculus*, 148701036; *Homo sapiens*, 344217763; *Methylibium petroleiphilum* PM1, 124266300; *Ferrimonas balearica* DSM 9799, 308049455; *Desulfuromonas acetoxidans* DSM 684, 95930587; *Blastopirellula marina* DSM 3645, 87312225; *Stackebrandtia nassauensis* DSM 44728, 291297650; *Renibacterium salmoninarum* ATCC 33209, 163840224; *Deinococcus radiodurans* R1, 6458004; *Bacillus* sp.916, 394994508; *Bifidobacterium longum* NCC2705, 23465838; *Frankia sp.* EAN1pec, 68197430; *Mycobacterium tuberculosis* CDC1551, 13879930; *Escherichia coli*, 50513417; *Bacillus cereus*, 51975946; *Bacillus halodurans* C-125, 10176350. Taxa abbreviations: Ba, *Actinobacteria*; Bb, *Bacteroidetes/Chlorobi* group; Bd, *Deinococcus-Thermus*; Bf, *Firmicutes*; Bi, *Acidobacteria*; Bo, *Planctomycetes*; Bp, *Proteobacteria*; Bs, *Spirochaetes*; Bv, *Chlamydiae/Verrucomicrobia* group; Bw, *Synergistetes*; E8, *stramenopiles*; Ea, *Amoebozoa*; Ed, *Rhizaria*; El, *Opisthokonta*.

conserved octameric motif AAAATTGA upstream of protein-coding sequences (Suhre et al. 2005) and stable hairpin structures or palindromic sequences at the 3′-ends of transcripts (Byrne et al. 2009). We searched for these two characteristic features in the moumouvirus genome. Overall, AAAATTGA sites were found within 150-nt regions upstream of the predicted start codons in 351 of the 930 predicted moumouvirus genes (37.7%). Among the orthologous genes between moumouvirus and megavirus, the fraction containing this motif was nearly the same. The AAAATTGA motif has been shown to function as an early promoter element in mimivirus and megavirus (Legendre et al. 2010). Together with the previously published comparison of megavirus and mimivirus genes containing this motif (Arslan et al. 2011), our observations on the presence of AAATTGA in upstream regions of moumouvirus genes imply that the expression pattern of orthologous genes is largely conserved among these three members of the *Mimiviridae*.

We also detected palindromic sites and predicted thermodynamically stable hairpins in the 3′ intergenic regions (150-nts downstream of the stop codon) of 725 of the 930 (78%) predicted protein-coding genes of the moumouvirus. Most of these structures are well-conserved between moumouvirus and megavirus. For example, we aligned and predicted the consensus structure of the hairpin element at the 3′-end of the major capsid gene for mimivirus, megavirus, and moumouvirus (supplementary file 5, Supplementary Material online) for which the presence of this element in the mature transcript has been experimentally validated by RNA sequencing of megavirus mg464 (Arslan et al. 2011). The polyadenylation of the megavirus gene occurs in the predicted conserved hairpin. The position of the experimentally verified polyadenylation sites is conserved between mimivirus and megavirus allowing us to predict the polyadenylation site of the moumouvirus gene (supplementary file 5, Supplementary Material online). These findings are in good agreement with the previous results demonstrating the conservation of these structural elements between mimivirus and megavirus (Arslan et al. 2011), suggesting that these hairpins function as transcription termination signals in moumouvirus.

Conclusions

Analysis of the moumouvirus genome confirms that it represents a third lineage amongst the *Mimiviridae*, in addition to those represented by mimivirus and megavirus. The moumouvirus genome further expands the pangenome of the *Mimiviridae* and emphasizes the dynamic evolution of the giant viruses, in particular extensive gene loss. The evolutionary relationship between the moumouvirus isolated from freshwater amoeba and *Megavirus chiliensis* that was isolated from a marine environment but shown to reproduce in the amoeba host of the other mimiviruses (Arslan et al. 2011) suggests that these giant viruses have a broad host range leading to ecological plasticity.

Supplementary Material

Supplementary methods and files S1–S5 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Gregory Gimenez and Ghislain Fournous for help with bioinformatic analyses and Lina Barrassi for technical assistance in the isolation of the moumouvirus. This work was supported by intramural funds of the US Department of Health and Human Services (National Library of Medicine) to N.Y., S.A.S., and E.V.K.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389–3402.

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM. 2011. Distant mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proc Natl Acad Sci U S A. 108:17486–17491.

Boyer M, et al. 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci U S A. 106:21848–21853.

Boyer M, et al. 2011. Mimivirus shows dramatic genome reduction after intraamoebal culture. Proc Natl Acad Sci U S A. 108:10296–10301.

Byrne D, et al. 2009. The polyadenylation site of mimivirus transcripts obeys a stringent "hairpin rule". Genome Res. 19:1233–1242.

Claverie JM, Abergel C. 2010. Mimivirus: the emerging paradox of quasi-autonomous viruses. Trends Genet. 26:431–437.

Claverie JM, Abergel C, Ogata H. 2009. Mimivirus. Curr Top Microbiol Immunol. 328:89–121.

Colson P, de Lamballerie X, Fournous G, Raoult D. 2012. Reclassification of giant viruses composing a fourth domain of life in the new order *Megavirales*. Intervirology 55:321–332.

Colson P, Gimenez G, Boyer M, Fournous G, Raoult D. 2011. The giant *Cafeteria roenbergensis* virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of Life. PLoS One 6: e18935.

Colson P, Raoult D. 2010. Gene repertoire of amoeba-associated giant viruses. Intervirology 53:330–343.

Colson P, et al. 2011. Viruses with more than 1,000 genes: mamavirus, a new *Acanthamoeba polyphaga* mimivirus strain, and reannotation of mimivirus genes. Genome Biol Evol. 3:737–742.

Filee J, Siguier P, Chandler M. 2007. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet. 23: 10–15.

Fischer MG, Allen MJ, Wilson WH, Suttle CA. 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci U S A. 107:19508–19513.

Iyer LM, Aravind L, Koonin EV. 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J Virol. 75:11720–11734.

Iyer LM, Balaji S, Koonin EV, Aravind L. 2006. Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 117:156–184.

Klose T, et al. 2010. The three-dimensional structure of mimivirus. Intervirology 53:268–273.

Koonin EV, Yutin N. 2010. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology 53:284–292.
Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. 2010. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. 18:11–19.
La Scola B, et al. 2003. A giant virus in amoebae. Science 299: 2033.
La Scola B, et al. 2008. The virophage as a unique parasite of the giant mimivirus. Nature 455:100–104.
La Scola B, et al. 2010. Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. Intervirology 53: 344–353.
Legendre M, et al. 2010. mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in mimivirus. Genome Res. 20:664–674.
Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 32:W327–W331.
Margulies M, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380.
Monier A, et al. 2008. Marine mimivirus relatives are probably large algal viruses. Virol J. 5:12.
Moreira D, Brochier-Armanet C. 2008. Giant viruses, giant chimeras: the multiple evolutionary histories of mimivirus genes. BMC Evol Biol. 8:12.
Ogata H, et al. 2006. Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene exchanges between intracellular pathogens. PLoS Genet. 2:e76.
Raoult D, Boyer M. 2010. Amoebae as genitors and reservoirs of giant viruses. Intervirology 53:321–329.
Raoult D, et al. 2004. The 1.2-megabase genome sequence of mimivirus. Science 306:1344–1350.
Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33:W686–W689.
Senkevich TG, Koonin EV, Bugert JJ, Darai G, Moss B. 1997. The genome of molluscum contagiosum virus: analysis and comparison with other poxviruses. Virology 233:19–42.
Suhre K, Audic S, Claverie JM. 2005. Mimivirus gene promoters exhibit an unprecedented conservation among all eukaryotes. Proc Natl Acad Sci U S A. 102:14689–14693.
Suzan-Monti M, La Scola B, Barrassi L, Espinosa L, Raoult D. 2007. Ultrastructural characterization of the giant volcano-like virus factory of Acanthamoeba polyphaga mimivirus. PLoS One 2:e328.
Thomas V, et al. 2011. Lausannevirus, a giant amoebal virus encoding histone doublets. Environ Microbiol. 13:1454–1466.
Yamada T. 2011. Giant viruses in the environment: their origins and evolution. Curr Opin Virol. 1:58–62.
Yutin N, Koonin EV. 2012. Hidden evolutionary complexity of nucleo-cytoplasmic large DNA viruses of eukaryotes. Virol J. 9:161.
Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J. 6:223.

Associate editor: Emmanuelle Lerat

SUPPLEMENTARY METHODS

Moumouvirus isolation, purification, and moumouvirus DNA extraction

Virus was isolated from water sampled in a cooling tower as described previously (La Scola et al., 2000). Briefly, large volumes of *A. castellanii* infected by the moumouvirus were cultured and filtered through 0.8-mm and 0.2-mm membranes. The moumouvirus preparation was obtained by washing the 0.2-mm membranes with K36 buffer.

Moumouvirus DNA extraction and genome sequencing

The moumouvirus DNA was extracted using the same procedure that was previously described for mimivirus (La Scola et al., 2008). Then, the moumouvirus genome was sequenced using the 454-Roche GS20 device (Roche Diagnostics Corp., Branford, CT, USA), and thereafter using the AB SOLiD instrument (Life Technologies Corp., Carlsbad, CA, USA) (Margulies et al., 2005; Raoult et al., 2004). The moumouvirus assembly was performed using a combination of Roche 454 paired-end and AB SOLiD sequencing reads. Roche 454 paired-end sequencing reads were first assembled de novo, then by mapping on the genomes of other mimiviruses using the Newbler assembly software (Margulies et al., 2005). These steps generated a moumouvirus genome sequence composed of two contigs. The single

gap was closed by PCR amplification and Sanger sequencing. Thereafter, sequencing reads obtained with the SOLiD technology were mapped on the previously assembled genome using the CLC Bio software (http://www.clcbio.com/index.php?id=28). This step notably led to correct, in the Moumouvirus genome, 454 sequencing errors consisting in A or T base insertions or deletions within A or T-homopolymeric sequences.

Genome annotation

Moumouvirus genome was translated by GeneMarkS software (http://exon.biology.gatech.edu/; Besemer and Borodovsky, 2005); ORFs shorter than 50 aa were discarded; long (>1000 nt) intergenic regions were checked for the presence of putative ORFs; ORFs ranging from 50 to 100 aa were kept if they showed a similarity to Refseq proteins or to a conserved domain (as of CDD).

Phylogenetic tree construction

Protein sequences were retrieved from the non-redundant database at the National Center for Biotechnology Information (NIH, Bethesda). The non-redundant protein sequence database was searched using the PSI-BLAST program (Altschul et al, 1997). Protein sequences were aligned using the MUSCLE program (Edgar, 2004); columns containing a large fraction of gaps (greater than 30%) and columns with low information content were removed from the alignment prior to the phylogenetic analysis. For the latter purpose,

each alignment column was assigned a homogeneity value between 0 and 1 by scaling the sum-of-pairs BLOSUM62 score within the column between those of a homogeneous column (the same residue in all aligned sequences) and a random column (Yutin et al, 2008), and only columns with homogeneity value greater than 0.2 were retained. The alignment was used to construct an initial maximum likelihood phylogenetic tree with the FastTree program with default parameters (Price et al, 2010). The initial tree and the alignment were fed to the ProtTest program (Darriba et al, 2011) to select the best substitution matrix. The LG substitution model (Le and Gascuel, 2008) with gamma-distributed site rates (LG+G) outperformed other models and was thereby selected for further analysis. Final maximum-likelihood trees were constructed using TreeFinder (LG matrix, G[Optimum]:4, 1,000 replicates, Search Depth 2).

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25(17):3389-3402.

Besemer J. and M. Borodovsky 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 33(Web Server issue):W451-4.

Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004, 32(5):1792-1797.

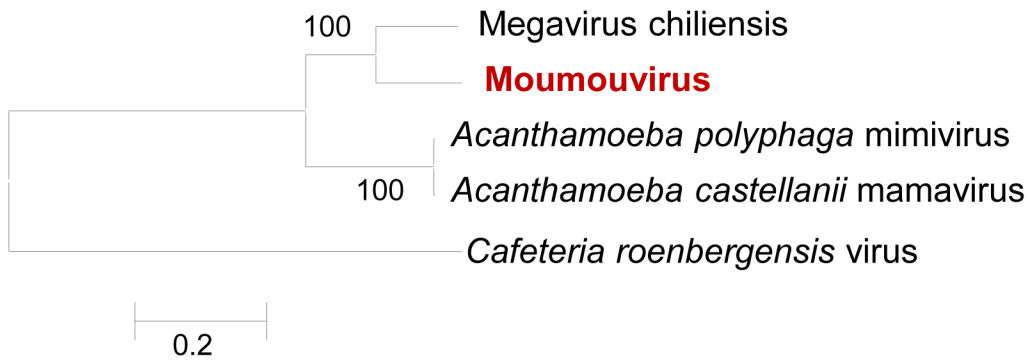
Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380. doi: nature03959 [pii] 10.1038/nature03959

Price MN, Dehal PS, Arkin AP: FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 2010, 5(3):e9490.

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

Yutin N, Makarova KS, Mekhedov SL, Wolf YI, Koonin EV: The deep archaeal roots of eukaryotes. Mol Biol Evol 2008, 25(8):1619-1630.

Supplemental File 2.

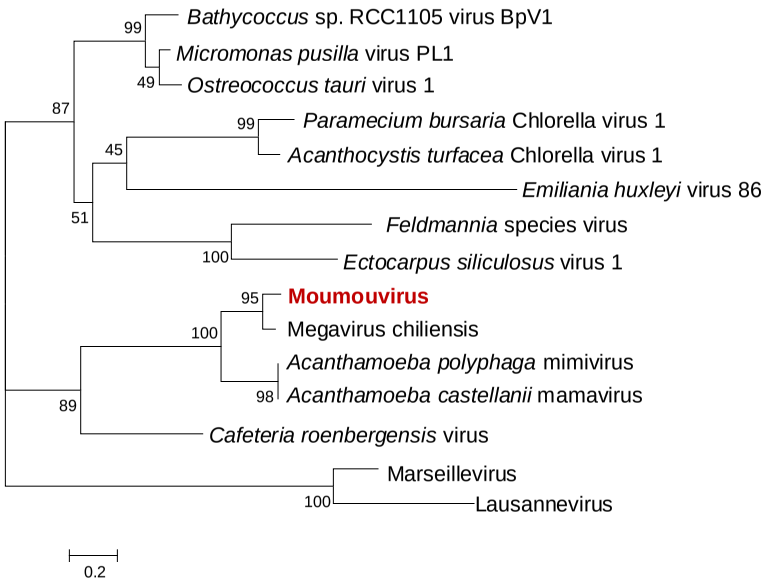


Protein list:

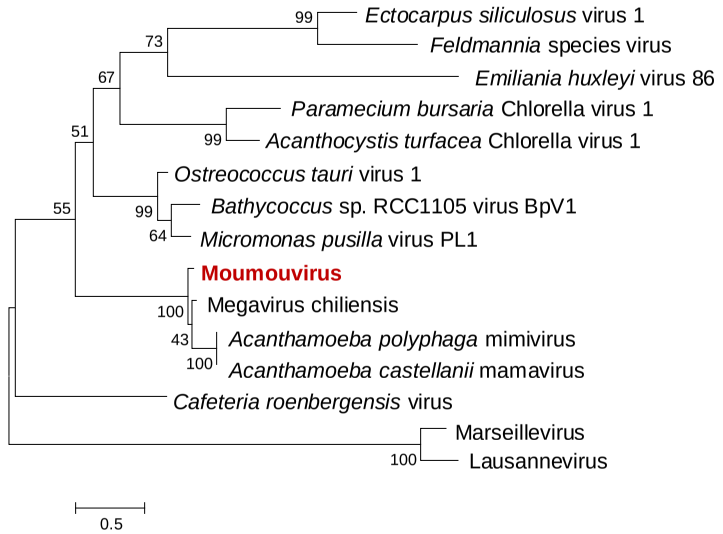
D5 family helicase-primase
RNA polymerase subunit 5
transcription initiation factor TFIIB
DNA directed RNA polymerase (II) subunit 1
DNA topoisomerase 2
A32-like packaging ATPase
mRNA-capping enzyme
Erv1 / Alr family protein
DNA polymerase type B
ribonucleotide reductase large subunit
ribonucleotide reductase small subunit
ribonuclease H
D6/D11-like helicase
Thiol oxidoreductase E10R
DNA directed RNA polymerase subunit 2

TreeFinder;
10,153 conserved positions

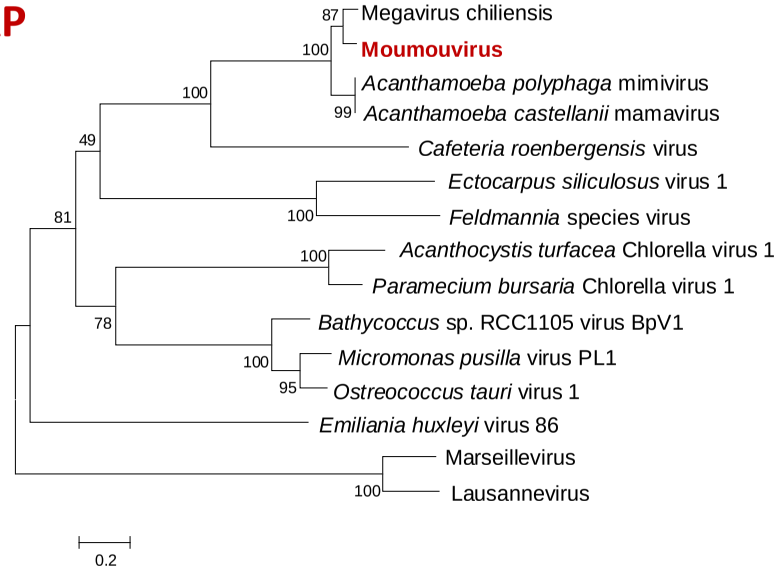
A32



A2



DNAP



TreeFinder

Supplementary file 4

gene with introns and/or inteins	mimivirus		mamavirus		megavirus		moumouvirus	
conserved	*protein name*		*protein name*		*protein name*		*ORF name*	
RNA polymerase largest subunit	R501	1 intron	MAMA_R595	1 intron	mg373	1 intron	Moumou_gene_290, 292	1 intron
DNA polymerase	R322	1 intein	MAMA_L400	1 intein	mg582	1 intein	Moumou_gene_486, 487	1 intein
different								
HSP70 chaperonin-like protein	L393	no introns	MAMA_L475	no introns	mg500	1 intron	Moumou_gene_411, 412	1 intron
major capsid protein	L425	2 introns	MAMA_L508	1 intron (different from Mimi)	mg464	2 introns (same as in Mimi)	Moumou_gene_377	1 intron
RNA polymerase second subunit	L244	3 introns	MAMA_L310	2 introns (both as in Mimi)	mg339	2 introns (one as in Mimi), 1 intein	Moumou_gene_253-258	5 introns, no intein
inteins unique to Moumouvirus								
DNA mismatch repair ATPase MutS	L359	no intein	MAMA_L440	no intein	mg543	no intein	Moumou_gene_451	1 intein
Hint (Hedgehog/Intein) domain-containing prote	-	-	-	-	-	-	Moumou_gene_88	1 intein

aligned region coordinates for mega/moumouvirus genes with introns

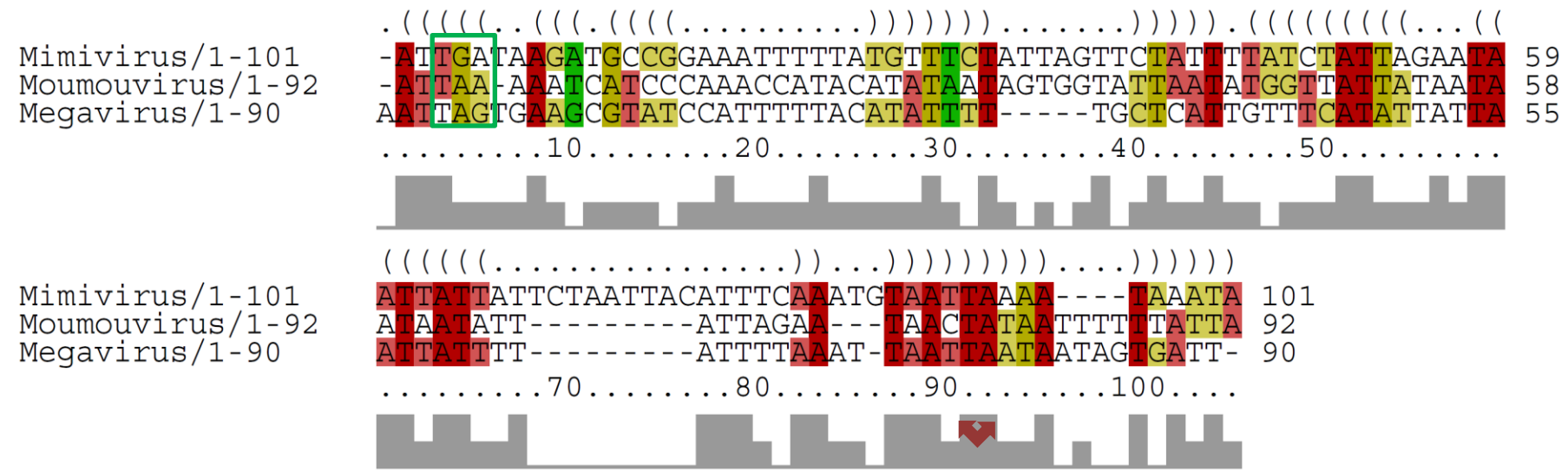
RNA polymerase largest subunit	exon 1 start	exon 1 end		exon 2 start	exon 2 end	comments
Megavirus NC_016072	388935	391898	intron, 1755 nt	393652	395163	*intron encodes an endonuclease*
Moumouvirus	316763	319730	intron, 1675 nt	321404	322915	*intron encodes an endonuclease*
HSP70 chaperonin-like protein	exon 1 start	exon 1 end		exon 2 start	exon 2 end	comments
Megavirus NC_016072	549977	550582	intron, 1146 nt	551727	553037	*intron encodes an endonuclease*
Moumouvirus	474317	474919	intron, 392 nt	475310	476620	*no intron-encoded endonuclease*
major capsid protein	exon 1 start	exon 1 end		exon 2 start	exon 2 end	
Megavirus NC_016072	508982	509037	intron, 1150 nt	510186	510213	intron, 437 nt
Moumouvirus	432263	432322	intron, 1246 nt	433567	433594	no intron
RNA polymerase second subunit	exon 1 start	exon 1 end		exon 2 start	exon 2 end	
Megavirus NC_016072	341204	341535	1602 nt	343136	345914	1397 nt
Moumouvirus	266755	267524	1588 nt	269111	272642	1394 nt

	exon 3 start	exon 3 end	
	510649	512334	
	433608	435280	
	exon 3 start	exon 3 end	
	347310	348857	
	274035	275582	

Supplemental File 5.

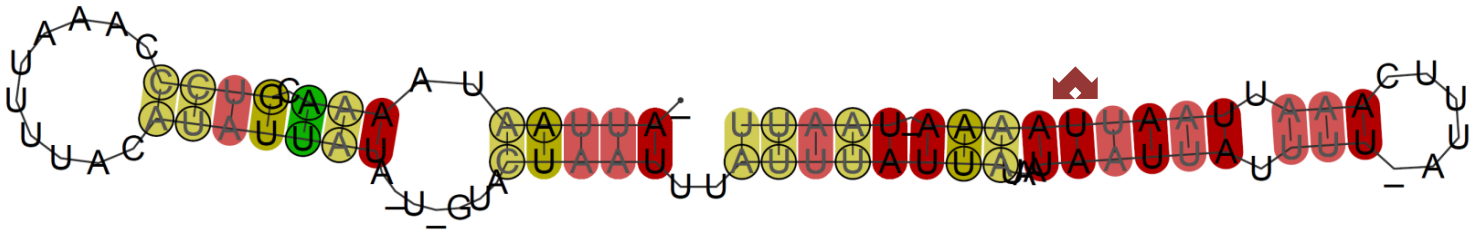
(A) Multiple alignment of the 3’UTR regions of the major capsid protein transcripts in mimivirus, megavirus and moumouvirus. Stop codon is shown in green box.

A



(B) The consensus secondary structure of the 3'UTRs is predicted by Alifold and Afold programs (Bernhart et al., 2008; Ogurtsov et al., 2006). Experimentally validated polyadenylation sites (red arrows) in mimivirus and megavirus are located in the conserved stable hairpin allowing prediction of the position of polyadenilation site in moumouvirus.

B



Chapter Three (3.2)

Draft genome sequences of Terra1 and Terra2 viruses, new members of the family *Mimiviridae* isolated from soil.

Yoosuf N, Pagnier I, Fournous G, Robert C, Raoult D, La Scola B, Colson P

Under revision in Virology

Chapter Three, 3.2

Draft genome sequences of Terra1 and Terra2 viruses, new members of the family *Mimiviridae* isolated from soil

Since the discovery of Mimivirus, the founding member of the family *Mimiviridae*, mimiviruses of amoeba have been delineated into three lineages named A, B and C (Raoult et al. 2004; Yoosuf et al. 2012). Apart from mimiviruses of amoeba, other closely related viruses classified within the family *Mimiviridae* have been described (Colson et al. 2011a). Before the present study, all the mimiviruses whose genome was described in detail had been isolated from water samples (Raoult et al. 2004; Fischer et al. 2010; Colson et al. 2011a; Arslan et al. 2011; Yoosuf et al. 2012). We analyzed and reported two draft genomes of mimiviruses of amoeba, Terra1 virus and Terra2 virus, which were recovered by co-culturing with *Acanthamoeba* spp isolated from soil samples. The Terra1 virus and Terra2 virus were predicted to encode 1,055 and 890 proteins, respectively. Based on phylogenetic analysis and comparative genomics, Terra1 virus was classified within lineage C and Terra 2 virus was classified within lineage A of amoeba-associated mimiviruses. Comparison of the genomic architecture of these two viruses with that of members from the same lineage and from the other lineages showed two different

patterns, with almost perfect collinearity within a given lineage and conservation in the central region and variance at the extremities between lineages. In addition, we identified a cluster of genes of bacterial origin in the genomes of lineage C mimiviruses, which is absent in the genomes of lineage A and B mimiviruses, indicating the possibility of HGT or gene loss in the genome of the ancestor of the respective lineages. Finally, we identified a small number of ORFans, but a large number of lineage ORFans, in both Terra1 virus and Terra2 virus genomes

TITLE PAGE

Full-length title: Draft genome sequences of Terra1 and Terra2 viruses, new members of the family *Mimiviridae* isolated from soil.

Authors list: Niyaz Yoosuf[1], Isabelle Pagnier[1], Ghislain Fournous[1], Catherine Robert[1], Didier Raoult[1, 2], Bernard La Scola[1,2], Philippe Colson[1, 2]

Affiliations: [1] Aix-Marseille Univ., Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE), UM63 CNRS 7278 IRD 198 INSERMU1095, IHU Méditerranée Infection, facultés de Médecine et de Pharmacie, Marseille, France; [2] Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie,Centre Hospitalo-Universitaire Timone, Assistance Publique – Hôpitaux de Marseille, Marseille, France

*** Corresponding author:** Philippe Colson, URMITE, Faculté de Médecine et de Pharmacie, Aix-Marseille Université, 27 Boulevard Jean Moulin, 13385 Marseille CEDEX 05, France.
Tel.: +33 491 324 375, Fax: +33 491 387 772,
E-mail: philippe.colson@univ-amu.fr

Key words: Terra1 virus; Terra2 virus; Mimivirus; Giant virus; "Megavirales"; Amoeba; Acanthamoeba; Soil; Environment

ABSTRACT

Since the discovery of Mimivirus, the founding member of the family *Mimiviridae*, three lineages, A, B and C, have been delineated among the mimiviruses of amoebae. To date, all giant viruses with detailed genomes have been isolated from water samples. Here, we describe the genome of two mimiviruses, Terra1 virus and Terra2virus, that were recovered by co-culturing on *Acanthamoeba* spp. from soil samples. These genomes are predicted to harbor 1,055 and 890 genes, respectively. Comparative genomics and phylogenomics show that Terra1 virus and Terra2 virus are classified within lineages C and A of the amoebae-associated mimiviruses, respectively. The genomic architecture of both viruses show conserved collinear central regions flanked by less conserved areas towards the extremities, when compared with other mimivirus genomes. A cluster of genes that are orthologous to bacterial genes and have no counterpart in other viral genomes except in lineage C mimiviruses was identified in Terra1 virus.

INTRODUCTION

Mimiviruses are giant viruses with particle and genome sizes that are on the same order of magnitude of small bacteria (Yutin & Koonin, 2012; Yutin, Colson et al., 2013; Colson, de Lamballerie et al., 2012). The founding member of the family *Mimiviridae*, is *Acanthamoeba polyphaga mimivirus* that was discovered in 2003 by coculturing on *Acanthamoeba polyphaga* from a water sample collected from a cooling tower in England in 1992 (La Scola, Audic et al., 2003). The Mimivirus genome is 1.18 kilobase pairs (kbp) and is predicted to encode for approximately 1,000 proteins, including proteins with functions that were believed to be the trademarks of cellular organisms, such as aminoacyl-tRNA synthetases (Raoult, Audic et al., 2004). Since the discovery of Mimivirus, approximately two dozen other giant viruses with capsid sizes ranging between 150-600 nm have been isolated from freshwater, saltwater and soil using the amoebal coculture method (La Scola, Campocasso et al., 2010; Boughalmi, Saadi et al., 2013; Colson, Raoult et al., 2013). To date, five mimivirus genomes have been described in detail. Three giant viruses, namely Mimivirus, Mamavirus (another Mimivirus strain) and Moumouvirus, have been recovered from water collected in cooling towers in the center of England; in Paris, France; and in southern France, respectively (La Scola, Audic et al., 2003; Raoult, Audic et al., 2004; La Scola, Desnues et al., 2008; Colson, Yutin et al., 2011;

Yoosuf, Yutin et al., 2012b). Two other viruses were isolated from marine coastal water. *Megavirus chilensis* was recovered from water collected in Chile by culturing with *Acanthamoeba* spp. (Arslan, Legendre et al., 2011), and the *Cafeteria roenbergensis* virus (Crov) was isolated from water collected in Texas, USA, from *Cafeteria roenbergensis*, a phagocytic protist belonging to the phylum *Chromalveolata* (Fischer, Allen et al., 2010). Three lineages, A, B and C, have been delineated among the mimiviruses of amoebae, and the leading members of these lineages are Mimivirus, Moumouvirus and *Megavirus chilensis*, respectively (Colson, de Lamballerie et al., 2012; Desnues, La Scola et al., 2012). More distantly related mimiviruses, including Crov, infect green algae, heterokonts and haptophyta (Yutin, Colson et al., 2013). In addition, smaller, though still giant, viruses that infect *Acanthamoeba* spp. have been isolated from environmental water samples since 2008, and these viruses compose a new proposed viral family called the family "Marseilleviridae", as the first of these viruses has been named Marseillevirus. Mimiviruses and marseilleviruses are common in environmental water, as demonstrated by the isolation of these viruses from approximately 20% of such samples and the frequent detection of sequences matching the DNA of these viruses in metagenomic studies (La Scola, Campocasso et al., 2010; Kristensen, Mushegian et al., 2010; Colson, Fancello et al., 2013). In addition, metagenomic reads matching the genomes of these giant viruses have also been detected (Colson, Fancello et al., 2013). Interestingly, two mimiviruses of amoebae classified within lineage C,

LBA111 virus and Shan virus, have recently been recovered from the bronchoalveolar fluid and the stool, respectively, of Tunisian patients presenting pneumonia (Saadi, Pagnier et al., 2013; Saadi, Ikanga Reteno et al., 2013), and marseilleviruses have been detected in the feces and the blood of two healthy people in rural Senegal and southeastern France, respectively (Lagier, Armougom et al., 2012; Colson, Fancello et al., 2013; Popgeorgiev, Boyer et al., 2013).

Mimiviruses and marseilleviruses have been assigned to the nucleo-cytoplasmic large DNA viruses (NCLDVs), a monophyletic group of viruses that encompasses members of the families *Poxviridae*, *Phycodnaviridae*, *Iridoviridae*, *Ascoviridae* and *Asfarviridae* primarily based on a limited set of core genes shared by all of these viruses (Iyer, Balaji et al., 2006; Raoult, Audic et al., 2004; Yutin & Koonin, 2012; Yutin, Wolf et al., 2009; Iyer, Aravind et al., 2001). It has recently been proposed that the NCLDVs should be reclassified into a new viral order called "Megavirales" (Colson, de Lamballerie et al., 2013). Two new giant viruses, *Pandoravirus salinus* and *Pandoravirus dulcis*, were isolated by coculturing on *Acanthamoeba* spp. from marine sediment collected in Chile and from the mud of a freshwater pond collected in Australia, respectively (Philippe, Legendre et al., 2013). These viruses have the largest particle and genome sizes among viruses. Notably, these viruses harbor 2.5 and 1.9 megabase-long genomes, respectively. In addition, ORFans compose approximately 93% of these viral genomes, and their morphology is unique among viruses.

To date, only mimviruses recovered from water samples have been described in detail. Here, we describe the genome of two mimiviruses recovered by coculturing with *Acanthamoeba* spp. from soil samples.

MATERIALS AND METHODS

The Terra1 and Terra2 viruses were isolated by inoculating *A. polyphaga*, as previously described (La Scola, Campocasso et al., 2010). The genomes of these viruses were sequenced by shotgun sequencing on the 454-Roche GS20 instrument (Boyer, Yutin et al., 2009). Sequence reads were assembled de novo using the Newbler tool (Margulies, Egholm et al., 2005) with 90% identity and 40 bp as overlap then using other mimivirus genomes as references with the CLC Bio software (http://www.clcbio.com/index.php?id=28).

Open reading frames were predicted using the GeneMarkS software with default parameters (Besemer & Borodovsky, 2005). The predicted protein sequences were searched against the GenBank database and the database of Clusters of Orthologous Groups of proteins (COGs) using BLASTp (Tatusov, Galperin et al., 2000; Altschul, Gish et al., 1990). The tRNAScanSE tool was used to search for transfer RNA genes (Schattner, Brooks et al., 2005). The strategy of reciprocal best BLASTp hits (Jordan, Rogozin et al., 2002) was performed to identify the orthologous set of genes using the

Proteinortho tool (Lechner, Findeiss et al., 2011). An e-value below 1e-03,an amino acid identity above 30% and a sequence coverage above 70% were used to consider hits as significant. ORFans were identified by BLASTp against the NCBI GenBank non-redundant protein sequence database (nr) with an e-value greater than 1e-3 and an alignment length greater than 80 amino acids (for alignment lengths <80 amino acids, we used an e-value of 1e-05). This e-value cut-off has been used in previous studies to define ORFans (Yin & Fischer, 2008; Yin & Fischer, 2006). Lineage ORFans, which are the ORFs that have homologs in a given taxonomic rank and no outside homolog (Yin & Fischer, 2006), were found with the same methodology as the ORFans. The genomic architectures were compared using MUMmer, Mauve and r2cat software (Kurtz, Phillippy et al., 2004; Darling, Mau et al., 2004; Husemann & Stoye, 2010). The sequence alignments were built using the muscle program (Edgar, 2004), and the alignments were trimmed using the trimAl tool (Capella-Gutierrez, Silla-Martinez et al., 2009). The phylogenetic trees were constructed using PhyML (Guindon, Lethiec et al., 2005) and were visualized using the FigTree software (http://www.umiacs.umd.edu/~morariu/figtree/).

RESULTS

Terra1 virus

The Terra1 virus was isolated from a soil sample collected in March 2009 in Marseille, France. Some of the features of this virus have been briefly described (La Scola, Campocasso et al., 2010).

The final assembly of the Terra1 virus genome yielded 12 contigs, including 8 large (> 1,500 bp) contigs and 4 small (<1,500 bp) contigs with mean coverages of 17 and 10 X, respectively. The Terra1 virus genome (GenBank Accession No.KF527229) is a double-stranded DNA molecule composed of approximately 1,233,835 bp. This genome is AT-rich (74.8%), which is similar to other mimiviruses. A total of 1,055 predicted proteins were tentatively identified in this genome, and 1,054 are larger than 100 amino acids. In addition, the Terra1 virus genome encodes two tRNAs (1 Leu-tRNA and 1 Trp-tRNA). The predicted genes were unevenly distributed on both DNA strands, with 640 located on the negative strand and 415 located on the positive strand. The length of the Terra1 predicted proteins ranges from 99 to 1,903 amino acids, with a mean length of 337 amino acids. A total of 1,044 (99.0%) of the 1,055 predicted proteins are homologous to a *Megavirus chilensis* protein with a mean amino acid sequence identity of 95.4%. In addition, 904 (85.7%) and 877 (82.2%) proteins from the Terra1 virus are homologous to Moumouvirus and Mimivirus proteins, respectively, with mean amino acid sequence

identities of 56.4% and 47.8%, respectively. *Megavirus chilensis* proteins were the best hit for 409 Terra1 virus proteins. In addition, 438 best hits were proteins of the Courdo7 virus that belongs to lineage C mimiviruses (La Scola, Campocasso et al., 2010; Desnues, La Scola et al., 2012), 4 best hits were proteins of Moumouvirus monve (La Scola, Campocasso et al., 2010; Desnues, La Scola et al., 2012) that belongs to the lineage B, and the 2 best hits were proteins of Mimivirus that belongs to the lineage A. Moreover, a BLASTp search against the predicted proteins from all previously published mimiviruses and against the COGs identified hits for 1,047 and 292 Terra1 virus proteins, respectively. The comparative analysis of the Terra1 virus using the Proteinortho tool with other annotated mimivirus genomes showed that the Terra1 virus shares a maximum number of orthologous genes (512, 48.5% of the gene repertoire) with *Megavirus chilensis* (lineage C). The Terra1 virus shares 433 orthologous genes (41.0%) with Moumouvirus (lineage B), 376 (35.6%) with Mimivirus and 367 (34.78%) with Mamavirus (lineage A). The *Cafeteria roenbergensis* virus, a distant member of the family *Mimiviridae*, shares 40 orthologs (3.8% of its proteome) with Terra1 virus proteins. Altogether, Mimivirus, Mamavirus, Moumouvirus and *Megavirus chilensis* share 287 orthologous genes with the Terra1 virus. Genomic dot plots of the Terra1 virus against the amoebae-associated mimiviruses of lineages A, B and C showed substantial levels of synteny, which decrease close to the genome extremities. The highest level of collinearity is with *Megavirus chilensis* (Figure 1).

The Terra1 virus genome shares a perfect collinearity with *Megavirus chiliensis*, with two inverted regions at the 5' extremity of the genome. The genomic dot-plot of the Terra1 virus against Mimivirus shows shorter, interrupted collinear regions with a larger inverted region in the central part of the genome. The phylogeny reconstruction based on family B DNA polymerase and a concatenation of core genes of the proposed order "Megavirales" indicates that the Terra1 virus belongs to lineage C of mimiviruses of amoebae, which is congruent with results from the comparative genomic analyses (Figures 2-3).

The BLASTp search against the GenBank nr database found hits for 1,050 of 1,055 Terra1 proteins and identified five ORFans (0.47% of the predicted proteome) (Figure 4). The Terra1 virus genome harbors 643 lineage ORFans, which represents 61% of the gene repertoire of this virus (Figure 4). The detailed analysis of the Terra1 virus proteome shows a set of genes that were either shared only by members of lineage C or that were shared by the members of lineages B and C but not lineage A. For example, Terra1_282 encodes a vacuolar sorting-associated protein that is essential for vacuolar biogenesis and maturation and is widely present in cellular organisms. This protein has been reported in *Megavirus chilensis*. Interestingly, this gene is only present in lineage C, which suggests that the lineage C ancestor might have acquired this gene. A less parsimonious scenario is that this gene was present in the ancestor of all of the lineages and was lost in lineages A and B. In addition, the Terra1_1006 gene that encodes a Cu/Zn superoxide dismutase, the

oxidoreductase enzyme that converts toxic superoxide radicals into molecular oxygen, has orthologs in all of the members of lineages B and C and in eukaryotic genomes. The distribution of this gene in mimiviruses suggests an evolutionary scenario in which this gene was present in the ancestor of all of the lineages and was lost in lineage A or that this gene was transferred in the ancestor of lineages B-C.

We have identified a set of proteins of putative bacterial origin in Terra1 virus that are contiguous and annotated as an UDP-N-acetylglucosamine2-epimerase; a dTDP-4-dehydrorhamnose reductase; a dTDP-d-glucose 4-6 dehydratase; a hypothetical protein; and an ExoV-like protein (Figure 5). The detailed comparative analysis of this set of genes with other mimivirus genomes identified orthologs for these genes only among the members of lineage C (except for the ExoV-like protein that is orthologous to mimivirus protein L143 and mamavirus protein L199), with no counterparts in the genomes of members of lineages A and B. A dTDP-4-dehydrorhamnose reductase and a dTDP-d-glucose 4-6 dehydratase are present in lineage A viral genomes, but these genes have different genomic locations because these genes correspond to Mimivirus genes R141 and L780, respectively, and are not orthologous to mimiviruses C proteins despite having the same functional annotation. In addition, we checked for the synteny of these genes in lineage C genomes and for corresponding regions in Mimivirus and Moumouvirus. The gene arrangement was conserved in the upstream and downstream regions of all the mimiviral genomes except in Moumouvirus (upstream). A

similar arrangement of contiguous genes of bacterial origin with predicted functions linked to carbohydrate metabolism was previously identified in Crov (Fischer, Allen et al., 2010). Fischer et al. hypothesized that the presence of these genes in the Crov genome could be the result of frequent encounters of this mimivirus with bacteria ingested by *Cafeteria roenbergensis*, the phagocytic protistan host, inside the host cytoplasm, and the presence of virally encoded transposases that might have enabled the integration of foreign DNA into the viral genome. The Terra1 virus genome also encodes a transposase (Terra1_894) and an integrase/resolvase (Terra1_895), which might have promoted the gain of the five aforementioned proteins that are widely distributed among bacterial species by mimiviruses of lineage C. A less parsimonious evolutionary scenario could be that the genes were gained by a common mimivirus A-C ancestor, and then the genes were lost in mimivirus lineages A and B.

Finally, we compared the gene repertoire of Terra1 virus with those of *P. dulcis* and *P. salinus* using BLASTp and 1e-3 as e-value cut-off. Bona fide orthologs were identified in the gene content of *P. dulcis* and *P. salinus* for 130 and 148 Terra1 virus proteins, respectively. Pandoraviruses share several core genes of the "Megavirales" with the Terra1 virus, including the class I core genes with the exception of the capsid protein.

Terra2 virus

The Terra2 virus was also isolated in March 2009 by inoculating *A. polyphaga* from a soil sample collected in Marseille, France. The icosahedral capsid is approximately 370 nm in size and is covered by a dense layer of fibers. The Terra2 virus capsid is smaller than that of the Terra1 virus, which is 420 nm in size.

The final assembly of the Terra2 virus genome yielded 18 contigs, including five large contigs (>1,500 bp) and 13 small contigs (<1,500 bp) with mean coverages of 20.3 and 9.8, respectively. The Terra2 virus genome (GenBank Accession No. KF527228) is a double-stranded, AT-rich (72%) DNA molecule composed of approximately 1,167,289 nucleotides. A total of 890 predicted proteins were tentatively identified in this genome, and 889 were larger than 100 amino acids. In addition, six tRNAs (1 His-tRNA, 1 Cys-tRNA, 1 Trp-tRNA and 3 Leu-tRNA) were detected. The protein-encoding genes were mostly distributed on the negative strand with 490 genes compared to 400 genes on the positive strand. The length of the Terra2 virus predicted proteins ranges from 99 to 2945 amino acids, with a mean length of 381 amino acids. The Terra2 virus encodes 82 proteins that have an amino acid length of more than 667 (ie., over 2 kilobase pair), of which three proteins have an amino acid length exceeding 2000, namely kinesin-like protein, Capsid protein and putative early transcription factor large subunit. A total of 883 (99%) of the 890 predicted Terra2 virus proteins have homologs in Mimivirus with a mean sequence identity of 95%, 884 (99%) have homologs in

Mamavirus with a mean sequence identity of 95%, 734 (82%) have homologs in Moumouvirus with a mean sequence identity of 48% and 769 (86%) have homologs in *Megavirus chilensis* with a mean sequence identity of 48%. A BLASTp search against all previously published mimivirus genomes and against COGs identified hits for 887 and 251 Terra2 virus genes, respectively. The comparative analysis of the Terra2 virus with other annotated mimivirus genomes using the Proteinortho tool yielded a maximum number of orthologs (728, 82%) with Mimivirus, followed by 721 (81%) orthologs with Mamavirus, 423 (48%) with *Megavirus chilensis* and 350 (39%) with Moumouvirus. In addition, 306 orthologs were shared by all these genomes, which represents 34% of the gene content of the Terra2 virus. The *Cafeteria roenbergensis* virus has 44 (5%) proteins orthologous to Terra2 virus proteins. These bidirectional best hit analyses showed that the Terra2 virus is most closely related to Mimivirus, the leading member of lineage A. Moreover, Mimivirus and Mamavirus proteins were the best hits for 662 and 182 Terra2 virus proteins, respectively. In addition, 23 of the best hits were from lentillevirus, another mimivirus of amoebae of lineage A (La Scola, Campocasso et al., 2010; Cohen, Hoffart et al., 2011). The best hits were from three mimiviruses of lineage C, *Megavirus chilensis*, Courdo11 virus and Courdo7 virus, for six, four and two proteins, respectively, and from Moumouvirus monve (La Scola, Campocasso et al., 2010; Desnues, La Scola et al., 2012), which belongs to lineage B, for two proteins. Genomic dot plots for the Terra2 virus against the

amoebae-associated mimiviruses of lineages A, B and C showed a high level of collinearity with Mimivirus, and a far lower collinearity was observed with Moumouvirus and *Megavirus chilensis* (Figure 1). The dot-plots of the Terra2 virus against Moumouvirus and *Megavirus chilensis* revealed shorter and interrupted collinear regions and a large inverted region located in the central part of the genome. In addition, the phylogeny reconstruction based on family B DNA polymerase and a concatenation of core genes of the proposed order "Megavirales" indicates that the Terra2 virus belongs to lineage A, which is in agreement with the comparative genomic results (Figures 2-3).

BLASTp searches against the nr database in GenBank found hits for 888 Terra2 proteins and identified two ORFans (0.22%) (Figure 4). In addition, the Terra2 virus genome harbors 524 lineage ORFans, which account for 59% of the viral gene repertoire (Figure 4). Similar to the Terra1 virus, the Terra2 virus possesses a set of genes shared only by members of a particular mimivirus lineage of amoebae. Terra2_291, a ricin-type lectin protein, only shares homologs with other members of lineage A, with an identity and a coverage higher than 90%. In addition, Terra2_822 protein encodes for a probable 7-dehydrocholesterol reductase, an enzyme that catalyzes the production of cholesterol from 7-dehydrocholesterol using NADPH. This gene is present in all the members of lineages A and B but has no counterpart in any member of lineage C. These findings suggest that gene gains and losses are common amongst the family *Mimiviridae*. Finally, we identified significant BLASTp hits in the gene content of

Pandoravirus dulcis and *Pandoravirus salinus* for 125 and 129 Terra2 virus proteins, respectively, whereas pandoraviruses were found to share several "Megavirales" core genes with the Terra2 virus, including the class I genes shared by all of the viruses of this proposed order, with the exception of the capsid protein.

Overall, we determined that the pan-genome of the family *Mimiviridae* has substantially expanded since the discovery of Mimivirus a decade ago. Indeed, the size of this pan-genome increased from 979 to 2,454 genes (Figure 6).

DISCUSSION

The comparative analyses of the Terra1 virus and the Terra2 virus genomes indicate that these giant viruses are bona fide members of the family *Mimiviridae*, are related to other mimiviruses that infect *Acanthamoeba* spp. and belong to lineages C and A, respectively. In addition, these analyses confirm previous findings regarding the architecture of the mimivirus genomes, with conserved collinear central regions flanked by less conserved areas towards the extremities. These features appear to be general trends of genome shape among the members of the proposed order "Megavirales", as initially highlighted in poxviruses and phycodnaviruses (Senkevich, Koonin et al., 1997; Filee, Siguier et al., 2007; McLysaght, Baldi et

al., 2003). The large inverted regions in the central part of the genomes and shorter collinear regions towards the tips identified by genome comparison were similar to those described during the comparisons of the Mimivirus genome with the *Megavirus chilensis* genome or the Moumouvirus genome (Arslan, Legendre et al., 2011; Yoosuf, Yutin et al., 2012a). These findings suggest that reshaping of the genomes can occur through the rearrangement of large fragments.

The evolutionary relationship between the Terra1 virus and the Terra2 virus, which were isolated from soil, and the other mimiviruses of amoebae, which have been isolated from fresh water and include Mimivirus, Moumouvirus and *Megavirus chilensis*, is an indication that mimiviruses have a wide habitat and may be found in many environmental samples inhabited by *Acanthamoeba* spp. (Yoosuf, Yutin et al., 2012a; Colson, Raoult et al., 2013). *Acanthamoeba* spp., the known hosts of mimiviruses, are phagocytic protists classified in the phylum *Amoebozoa* and are predominant among the organisms in soil and water (Barker & Brown, 1994; Moliner, Fournier et al., 2010; Thomas & Greub, 2010). These free living amoebae can ingest any particle with a size greater than 0.5 μm at the trophozoite stage and are known to graze on multiple organisms and microorganisms including bacteria, yeasts, fungi, viruses and algae.Therefore, these amoeba engulf large amounts of foreign DNA (Rodriguez-Zaragoza, 1994; Barker & Brown, 1994; Horn & Wagner, 2004). Moreover, *Acanthamoeba* spp.can resist various unfavorable conditions by differentiating into cysts (Raoult & Boyer, 2010; Bertelli & Greub,

2012). Interestingly, Mimivirus-like particles were observed by light microscopy within *Acanthamoeba* spp. in treated sewage sludge from a wastewater treatment plant in the UK (Gaze, Morgan et al., 2011). This finding suggested that amoebae could promote the dissemination of mimiviruses to land.

As in previous studies that analyzed the genomes of giant viruses, some evidence of lateral gene transfers between these genomes and genomes of other organisms from other branches of life have been found (Raoult, Audic et al., 2004; Boyer, Yutin et al., 2009; Filee, 2009). These transfers can be related to the particular lifestyle of mimiviruses within the amoebae where the viruses live sympatrically with other viruses and bacteria and can exchange genes with the viruses, bacteria and the eukaryotic host (Raoult & Boyer, 2010; Moliner, Fournier et al., 2010; Bertelli & Greub, 2012). The identification in the genome of the Terra1 virus of a cluster of genes that are orthologous to bacterial genes and have no counterpart in other viral genomes other than those of mimiviruses of lineage C exemplifies this capability to acquire genes. Such clusters of bacterial genes have been observed in the Mimivirus genome (Filee, Siguier et al., 2007) and the Crov genome (Fischer, Allen et al., 2010). In the case of Crov, which infects phagocytic protists other than *Acanthamoeba* spp., a 38 kbp genomic fragment was identified that encompasses 34 predicted genes, among which 14 were most similar to bacterial genes, and 7 were predicted to be involved in carbohydrate metabolism. Strikingly, the cluster of genes identified in the Terra1

virus genome also encodes genes involved in carbohydrate metabolism.

Since the discovery of Mimivirus, the founding member of the family *Mimiviridae*, the pan-genome of this viral family has shown a 2.5-fold expansion. Five and two ORFans were identified among the predicted genes from the Terra1 and Terra2 viruses, respectively. These small numbers of ORFans in the genome of these two new members of the family *Mimiviridae* suggest that the size of the pan-genome for mimiviruses of amoebae reached a plateau and could be considered closed based on currently available genomes. This finding may rely on the fact that these giant viruses were isolated through the same strategy of co-culturing the viruses with *Acanthamoeba polyphaga or castellanii*, and the majority of the amoebas were from water or soil samples, although these samples were collected from various geographical areas on three continents (La Scola, Campocasso et al., 2010; Pagnier, Raoult et al., 2013; Colson, Raoult et al., 2013). Notwithstanding, it has been recently shown that the family *Mimiviridae* is expanding through the reclassification of viruses formerly classified as phycodnaviruses (Yutin, Colson et al., 2013). In this view, it is worthy to note that the Crov genome, a mimivirus that is distantly related to Mimivirus, Moumouvirus and *Megavirus chilensis*, has been isolated in *Cafeteria roenbergensis*, a different host than *Acanthamoeba* spp. (Fischer, Allen et al., 2010; Massana, del et al., 2007). Finally, the present study underlines that the lineage ORFans compose a considerable part of the gene content of

mimiviruses that infect amoebae, and this observation, together with the large proportion of hypothetical proteins among this set of genes, stresses that much remains to be known about these viruses.

LEGENDS TO FIGURES

Figure 1. Genomic dot plots of the Terra1 virus against amoebae-associated mimiviruses of lineages A, B and C.

The Terra1 virus and Terra2 virus genomes were compared to the Mimivirus (a), Moumouvirus (b) and *Megavirus chiliensis* (c) genomes. Dot plots were constructed using the MUMmer3.22 software (Delcher, Salzberg et al., 2003): nucleotide-based alignments were performed with MUMmer. Dot plots were generated by the MUMmerplot script and the program gnuplot (www.gnuplot.info/docs_4.0/gnuplot.html). Aligned segments are represented by dots or lines. Colors indicate forward matches in red and reverse complement matches in blue.

Figure 2. Phylogenetic treesconstructed based on the family B DNA polymerase from selected members of the family "Megavirales" and *Pandoravirus dulcis* and *P. salinus* using the maximum likelihood method.

The numbers at tree nodes indicate bootstrap replicates of 100. The sequence alignment was generated using the muscle program (Edgar, 2004), and the trimAl tool (Capella-Gutierrez, Silla-Martinez et al., 2009) was used for automated alignment trimming. The phylogenic tree was constructed for 44 sequences (1,441 conserved positions)

using PhyML (Guindon, Lethiec et al., 2005) and visualized with FigTree software (http://www.umiacs.umd.edu/~morariu/figtree/).

Figure 3. The phylogenetic tree of the mimiviruses constructed from concatenated alignments of core genes of the proposed order "Megavirales".

Core genes used for the phylogeny reconstruction included primase-helicase, family-B DNA polymerase, packaging ATPase and A2L-like transcription factor. Marseillevirus was used as an outgroup. The alignment included 3,230 positions that were deemed reliably aligned. The bootstrap values are indicated for each internal branch.

Figure 4. The distribution of ORFans and lineage ORFans in the family *Mimiviridae* according to the timescale of the virus description.

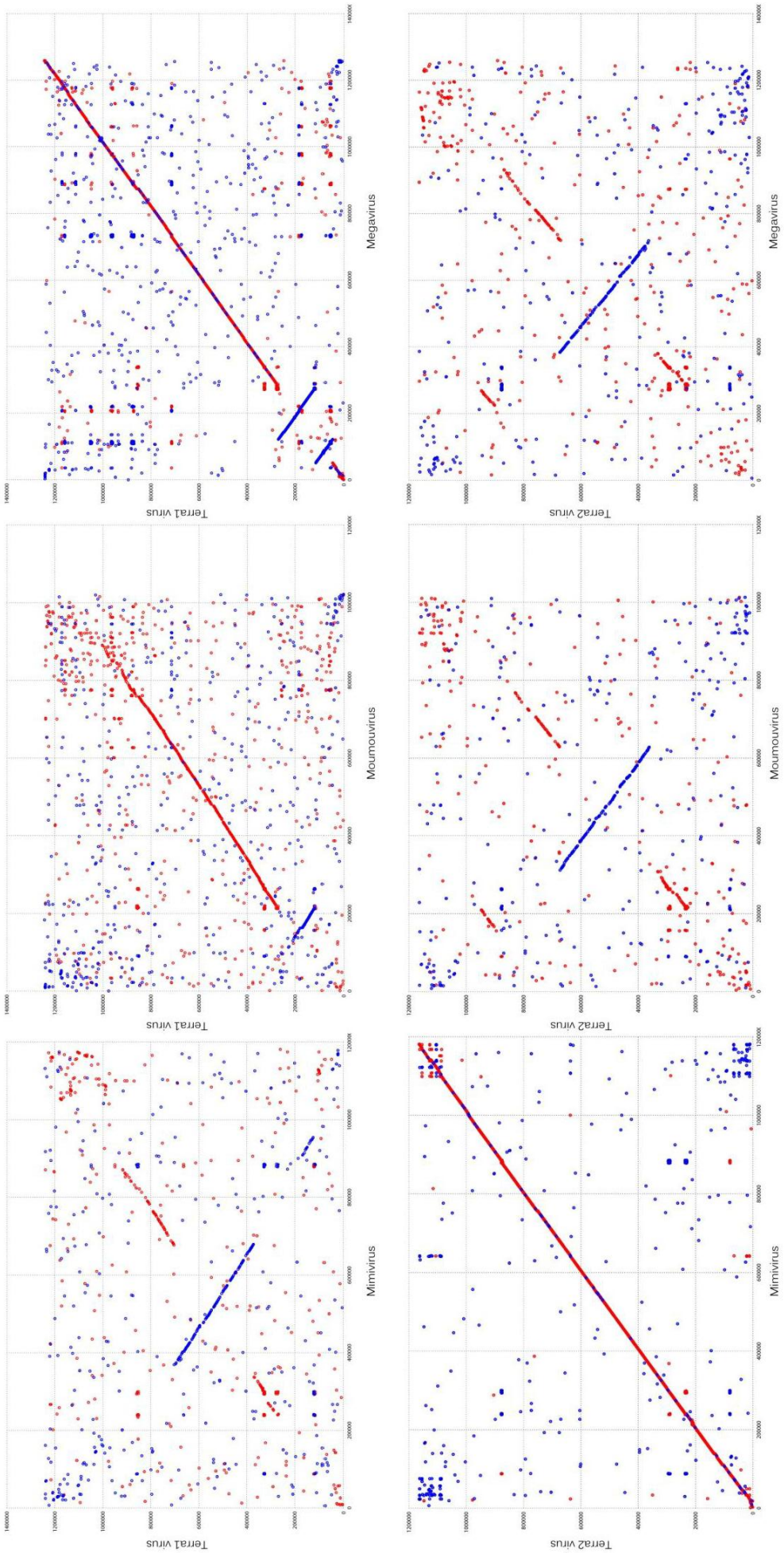
Figure 5. Comparative gene organization and gene synteny for a region of interest in the genomes of mimiviruses.

The genes and their orientation are depicted by polygons. Genes of interest across the genomes are depicted with the colors. The orientation of Terra1 virus, Mimivirus and Moumouvirus genomes are reversed for the feasibility of syntenic regions.

Figure 6. The number of genes composing the pan-genome of the mimiviruses, according to the time of description of the genomes of these giant viruses.

Crov, Cafeteria roenbergensis virus; OLPV, Organic Lake Phycodnavirus; PGV, *Phaeocystis globosa* virus

Fig. 1


Megavirus
Terra1 virus
Terra2 virus
Moumouvirus
Mimivirus

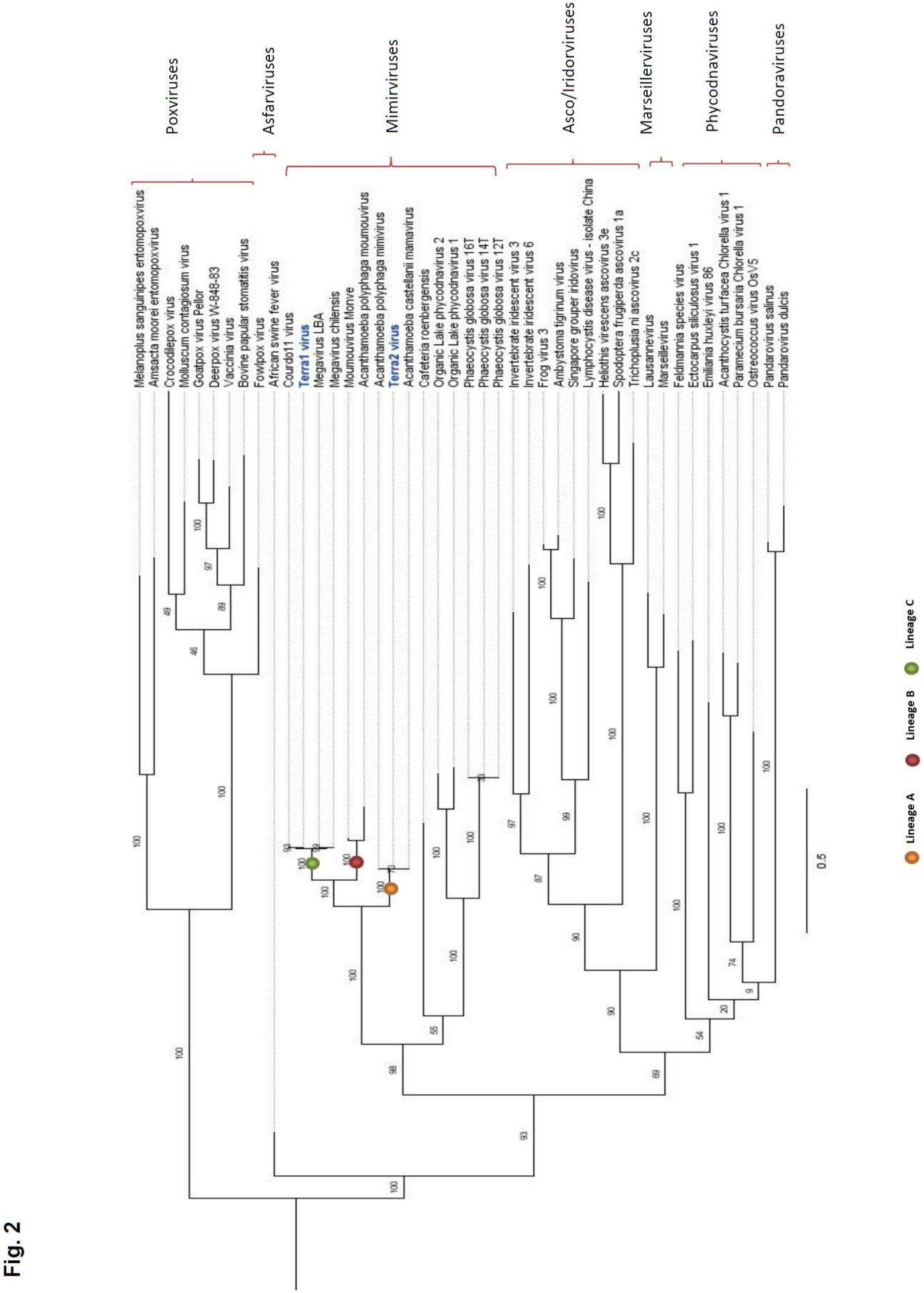

Poxviruses
Asfarviruses
Mimiviruses
Asco/Iridorviruses
Marseillerviruses
Phycodnaviruses
Pandoraviruses
Melanoplus sanguinipes entomopoxvirus
Amsacta moorei entomopoxvirus
Crocodilepox virus
Molluscum contagiosum virus
Goatpox virus Pellor
Deerpox virus W-848-83
Vaccinia virus
Bovine papular stomatitis virus
Fowlpox virus
African swine fever virus
Courdo11 virus
Terra1 virus
Megavirus LBA
Megavirus chilensis
Moumouvirus Monve
Acanthamoeba polyphaga moumouvirus
Acanthamoeba polyphaga mimivirus
Terra2 virus
Acanthamoeba castellanii mamavirus
Cafeteria roenbergensis
Organic Lake phycodnavirus 2
Organic Lake phycodnavirus 1
Phaeocystis globosa virus 16T
Phaeocystis globosa virus 14T
Phaeocystis globosa virus 12T
Invertebrate iridescent virus 3
Invertebrate iridescent virus 6
Frog virus 3
Ambystoma tigrinum virus
Singapore grouper iridovirus
Lymphocystis disease virus - isolate China
Heliothis virescens ascovirus 3e
Spodoptera frugiperda ascovirus 1a
Trichoplusia ni ascovirus 2c
Lausannevirus
Marseillevirus
Feldmannia species virus
Ectocarpus siliculosus virus 1
Emiliania huxleyi virus 86
Acanthocystis turfacea Chlorella virus 1
Paramecium bursaria Chlorella virus 1
Ostreococcus virus OsV5
Pandoravirus salinus
Pandoravirus dulcis
0.5
Lineage A
Lineage B
Lineage C

Fig. 2

Fig. 3

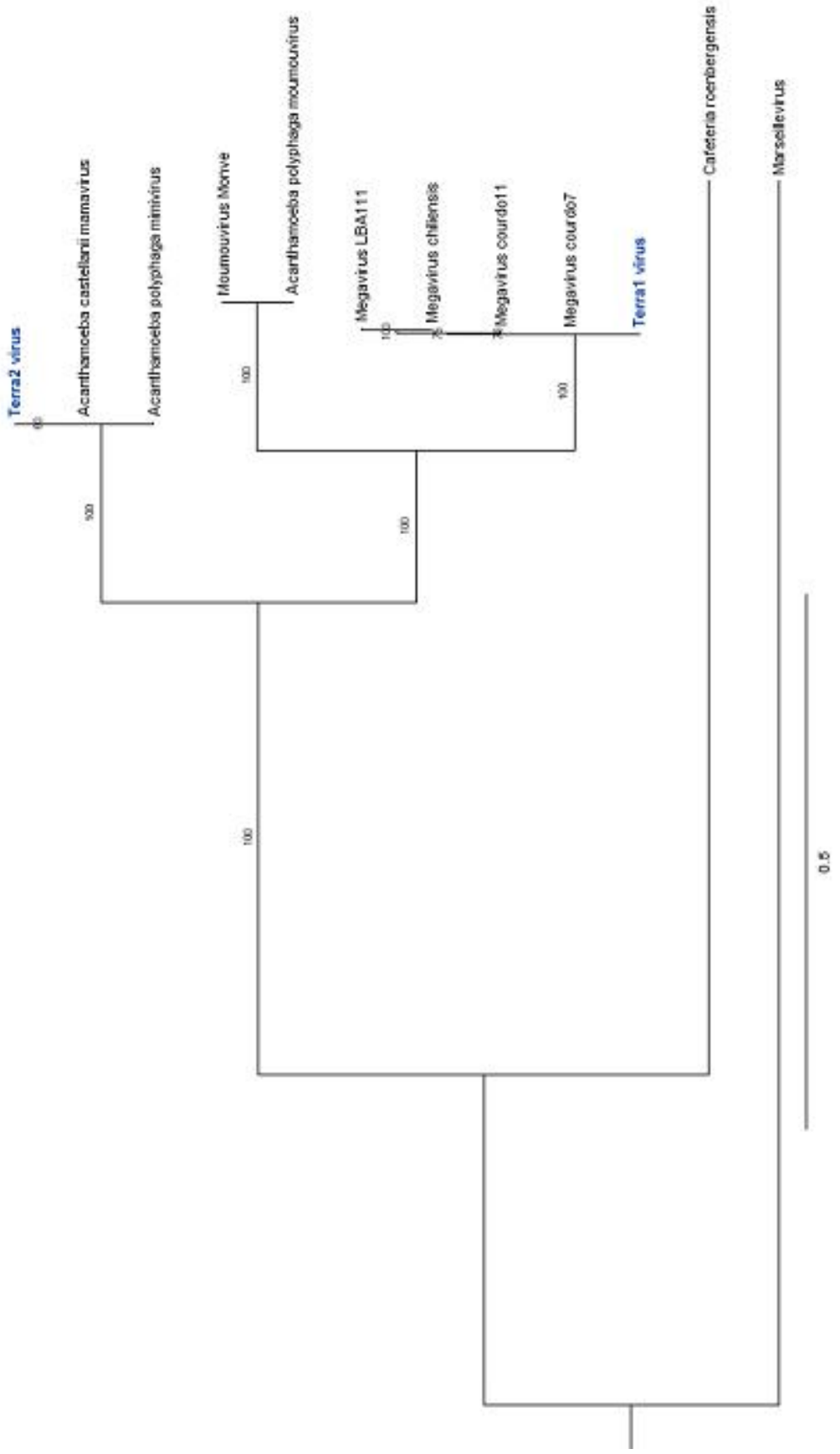

Terra2 virus
Acanthamoeba castellanii mamavirus
Acanthamoeba polyphaga mimivirus
Moumouvirus Monve
Acanthamoeba polyphaga moumouvirus
Megavirus LBA111
Megavirus chilensis
Megavirus courdo11
Megavirus courdo07
Terra1 virus
Cafeteria roenbergensis
Marseillevirus
100
100
100
100
100
100
100
0.5

Fig. 4

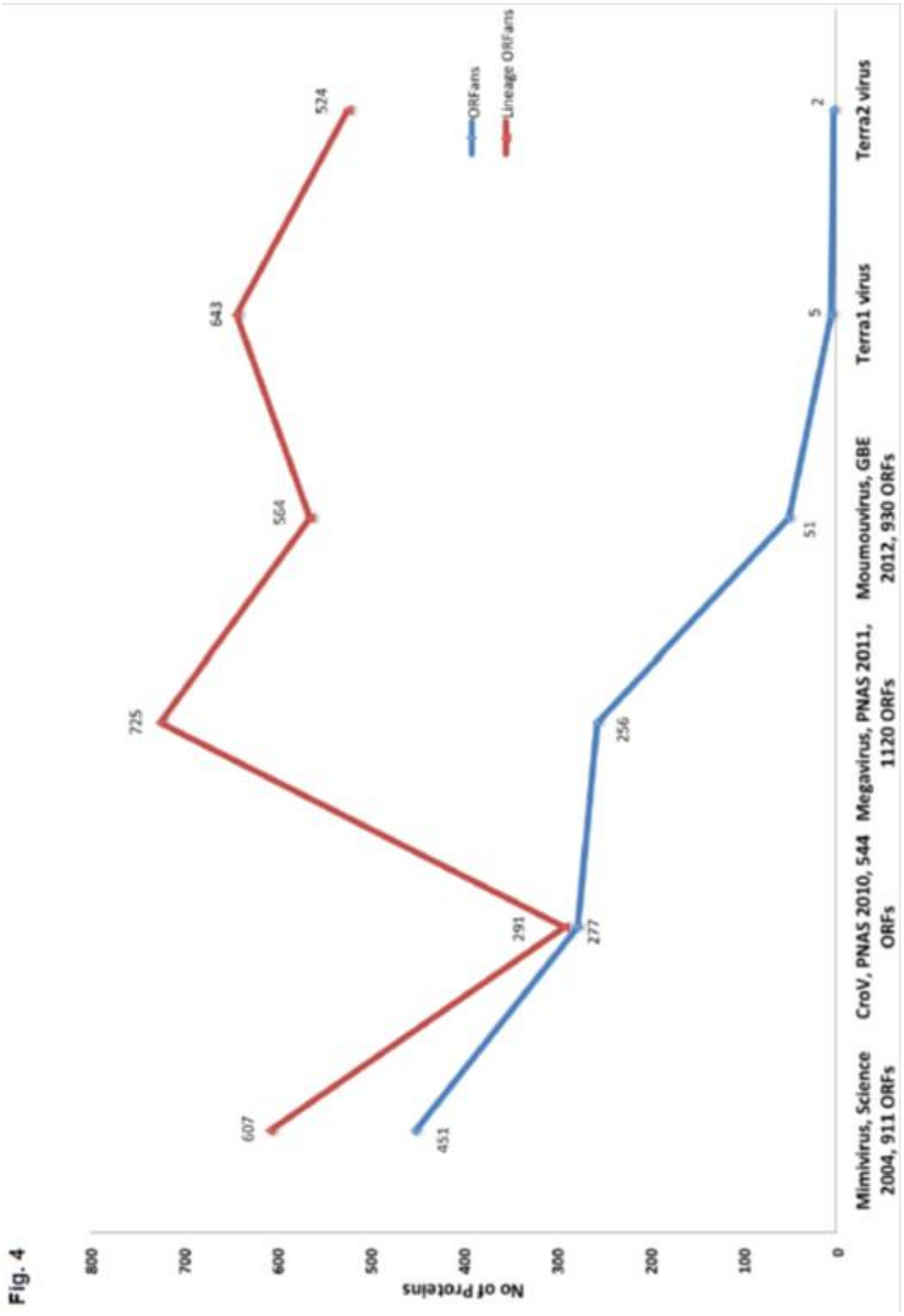

ORFans
Lineage ORFans
No of Proteins
800
700
600
500
400
300
200
100
0
451
607
277
291
256
725
51
564
5
643
2
524
Mimivirus, Science 2004, 911 ORFs
CroV, PNAS 2010, 544 ORFs
Megavirus, PNAS 2011, 1120 ORFs
Moumouvirus, GBE 2012, 930 ORFs
Terra1 virus
Terra2 virus

Fig. 4

Fig. 5

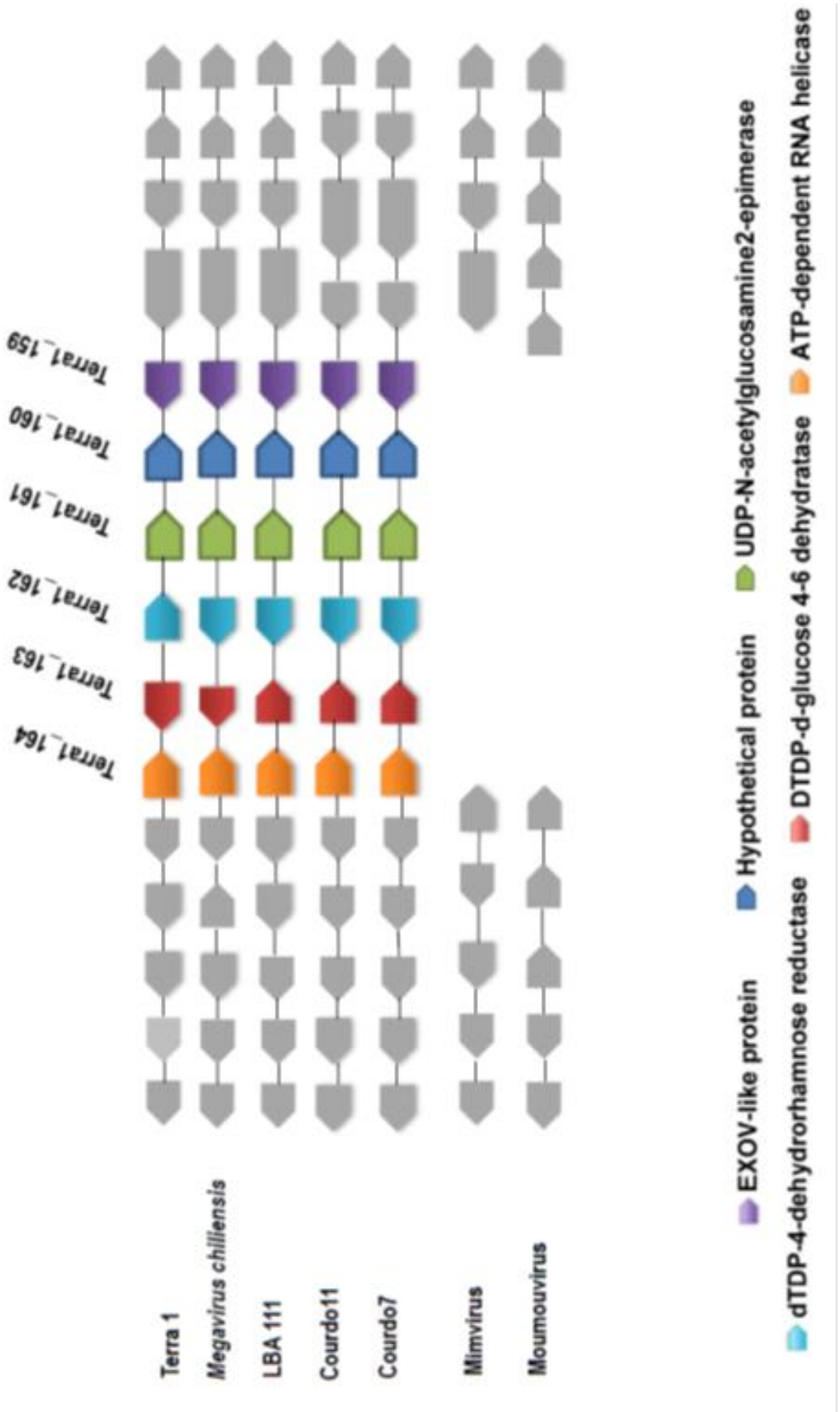
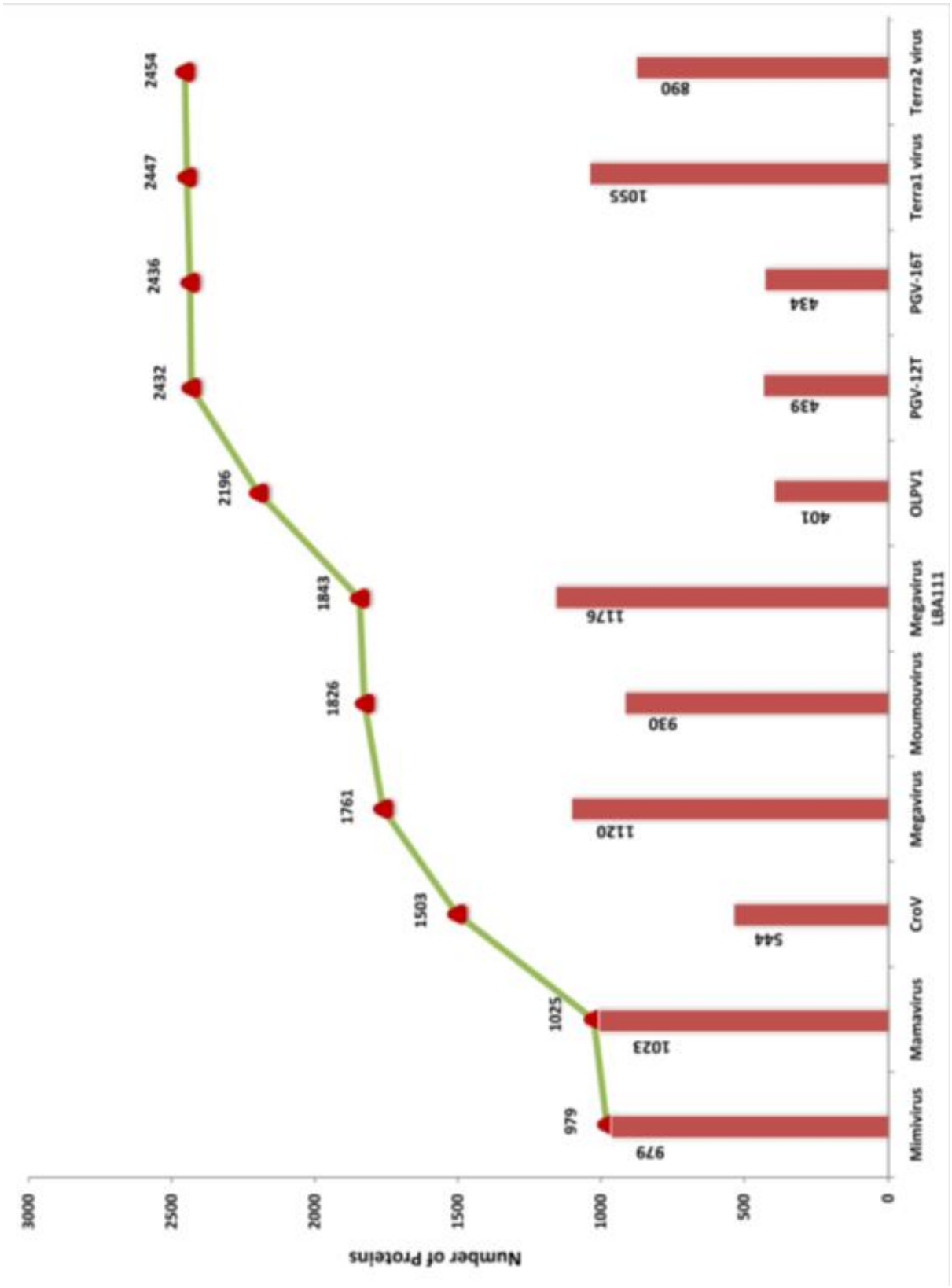

Terra 1
Megavirus chiliensis
LBA 111
Courdo11
Courdo7
Mimvirus
Moumouvirus
Terra1_159
Terra1_160
Terra1_161
Terra1_162
Terra1_163
Terra1_164
EXOV-like protein
Hypothetical protein
UDP-N-acetylglucosamine2-epimerase
dTDP-4-dehydrorhamnose reductase
DTDP-d-glucose 4-6 dehydratase
ATP-dependent RNA helicase

Fig. 6


Mimivirus
Mamavirus
CroV
Megavirus
Moumouvirus
Megavirus LBA111
OLPV1
PGV-12T
PGV-16T
Terra1 virus
Terra2 virus
979
1023
544
1120
930
1176
401
439
434
1055
890
2454
2447
2436
2432
2196
1843
1826
1761
1503
1025
979
0
500
1000
1500
2000
2500
3000
Number of Proteins

REFERENCE LIST

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool. J.Mol.Biol. 215, 403-410.

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., 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.

Barker, J., Brown, M. R., 1994. Trojan horses of the microbial world: protozoa and the survival of bacterial pathogens in the environment. Microbiology 140 (Pt 6), 1253-1259.

Bertelli, C., Greub, G., 2012. Lateral gene exchanges shape the genomes of amoeba-resisting microorganisms. Front Cell Infect.Microbiol. 2, 110.

Besemer, J., Borodovsky, M., 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 33, W451-W454.

Boughalmi, M., Saadi, H., Pagnier, I., Colson, P., Fournous, G., Raoult, D., La, S. 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., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M. G., Suzan-Monti, M., La, S. B., Koonin, E. V., Raoult, D., 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc.Natl.Acad.Sci.U.S.A 106, 21848-21853.

Capella-Gutierrez, S., Silla-Martinez, J. M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25, 1972-1973.

Cohen, G., Hoffart, L., La, S. B., Raoult, D., Drancourt, M., 2011. Ameba-associated Keratitis, France. Emerg.Infect.Dis. 17, 1306-1308.

Colson, P., de Lamballerie, X., Fournous, G., Raoult, D., 2012. Reclassification of Giant Viruses Composing a Fourth Domain of Life in the New Order Megavirales. Intervirology.

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., La Scola, B., Raoult, D., 2013. "Megavirales", a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch.Virol.

Colson, P., Fancello, L., Gimenez, G., Armougom, F., Desnues, C., Fournous, G., Yoosuf, N., Million, M., La Scola, B., Raoult, D., 2013. Evidence of the megavirome in humans. J.Clin.Virol. 57, 191-200.

Colson, P., Raoult, D., Pagnier, I., La Scola, B., 2013. Evidence of the presence of mimiviruses, their virophages, and marseilleviruses in environmental and clinical samples worldwide. Google maps content, URMITE laboratory. URL: http://maps.google.fr/maps/ms?vps=2&hl=fr&ie=UTF8&oe=UTF8&msa=0&msid=200914559094835369589.0004beba4af112f60dcf2.

Colson, P., Yutin, N., Shabalina, S. A., Robert, C., Fournous, G., La Scola, B., Raoult, D., Koonin, E. V., 2011. Viruses with more than 1000 genes: Mamavirus, a new Acanthamoeba castellanii mimivirus strain, and reannotation of mimivirus genes. Genome Biol.Evol.

Darling, A. C., Mau, B., Blattner, F. R., Perna, N. T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14, 1394-1403.

Delcher, A. L., Salzberg, S. L., Phillippy, A. M., 2003. Using MUMmer to identify similar regions in large sequence sets. Curr.Protoc.Bioinformatics. Chapter 10:Unit 10.3. doi: 10.1002/0471250953.bi1003s00., Unit.

Desnues, C., La Scola, B., Yutin, N., Fournous, G., Robert, C., Azza, S., Jardot, P., Monteil, S., Campocasso, A., Koonin, E. V., Raoult, D., 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. Proc.Natl.Acad.Sci.U.S.A. 109, 18078-18083.

Edgar, R. C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res.32, 1792-1797.

Filee, J., 2009. Lateral gene transfer, lineage-specific gene expansion and the evolution of Nucleo Cytoplasmic Large DNA viruses. J. Invertebr.Pathol. 101, 169-171.

Filee, J., Siguier, P., Chandler, M., 2007. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet. 23, 10-15.

Fischer, M. G., Allen, M. J., Wilson, W. H., 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.

Gaze, W. H., Morgan, G., Zhang, L., Wellington, E. M., 2011. Mimivirus-like Particles in Acanthamoebae from Sewage Sludge. Emerg.Infect.Dis. 17, 1127-1129.

Guindon, S., Lethiec, F., Duroux, P., Gascuel, O., 2005. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 33, W557-W559.

Horn, M., Wagner, M., 2004. Bacterial endosymbionts of free-living amoebae. J.Eukaryot.Microbiol. 51, 509-514.

Husemann, P., Stoye, J., 2010. r2cat: synteny plots and comparative assembly. Bioinformatics. 26, 570-571.

Iyer, L. M., Aravind, L., Koonin, E. V., 2001. Common origin of four diverse families of large eukaryotic DNA viruses. J.Virol. 75, 11720-11734.

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

Jordan, I. K., Rogozin, I. B., Wolf, Y. I., Koonin, E. V., 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res. 12, 962-968.

Kristensen, D. M., Mushegian, A. R., Dolja, V. V., Koonin, E. V., 2010. New dimensions of the virus world discovered through metagenomics. Trends Microbiol. 18, 11-19.

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S. L., 2004. Versatile and open software for comparing large genomes. Genome Biol 5, R12.

La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J. M., Raoult, D., 2003. A giant virus in amoebae. Science 299, 2033.

La Scola, B., Campocasso, A., N'Dong, R., Fournous, G., Barrassi, L., Flaudrops, C., Raoult, D., 2010. Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. Intervirology 53, 344-353.

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

Lagier, J. C., Armougom, F., Million, M., Hugon, P., Pagnier, I., Robert, C., Bittar, F., Fournous, G., Gimenez, G., Maraninchi, M., Trape, J. F., Koonin, E., Koonin, E. V., La Scola, B., Raoult, D., 2012. Microbial culturomics: paradigm shift in the human gut microbiome study. Clin.Microbiol.Infect.

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., Prohaska, S. J., 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC.Bioinformatics. 12:124. doi: 10.1186/1471-2105-12-124., 124-12.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., Rothberg, J. M., 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 437, 376-380.

Massana, R., del, C. J., Dinter, C., Sommaruga, R., 2007. Crash of a population of the marine heterotrophic flagellate Cafeteria roenbergensis by viral infection. Environ.Microbiol. 9, 2660-2669.

McLysaght, A., Baldi, P. F., Gaut, B. S., 2003. Extensive gene gain associated with adaptive evolution of poxviruses. Proc.Natl.Acad.Sci.U.S.A 100, 15655-15660.

Moliner, C., Fournier, P. E., Raoult, D., 2010. Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. FEMS Microbiol.Rev. 34, 281-294.

Pagnier, I., Raoult, D., La Scola, B., 2013. A collection of giant viruses from amoebae. Intervirology.

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J. M., Abergel, C., 2013. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. Science. 341, 281-286.

Popgeorgiev, N., Boyer, M., Fancello, L., Monteil, S., Robert, C., Rivet, R., Nappez, C., Azza, S., Chiaroni, J., Raoult, D., Desnues, C., 2013. Giant Blood Marseillevirus recovered from asymptomatic blood donors. J.Infect.Dis.

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

Raoult, D., Boyer, M., 2010. Amoebae as genitors and reservoirs of giant viruses. Intervirology 53, 321-329.

Rodriguez-Zaragoza, S., 1994. Ecology of free-living amoebae. Crit Rev.Microbiol. 20, 225-241.

Saadi, H., Ikanga Reteno, D. G., Colson, P., Aherfi, S., Minodier, P., Pagnier, I., Fenollar, F., Robert, C., Raoult, D., La Scola, B., 2013. Shan virus, isolation of a new Mimivirus from the stool of a Tunisian patient with pneumonia. Intervirology.

Saadi, H., Pagnier, I., Colson, P., Kanoun Cherif, J., Beji, M., Boughalmi, M., Azza, S., Armstrong, N., Robert, C., Fournous, G., La Scola, B., Raoult, D., 2013. First isolation of Mimivirus in a patient with pneumonia. Clin.Infect Dis. 57, e127-34.

Schattner, P., Brooks, A. N., Lowe, T. M., 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33, W686-W689.

Senkevich, T. G., Koonin, E. V., Bugert, J. J., Darai, G., Moss, B., 1997. The genome of molluscum contagiosum virus: analysis and comparison with other poxviruses. Virology. 233, 19-42.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., Koonin, E. V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33-36.

Thomas, V., Greub, G., 2010. Amoeba/amoebal symbiont genetic transfers: lessons from giant virus neighbours. Intervirology 53, 254-267.

Yin, Y., Fischer, D., 2006. On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. BMC.Evol.Biol 6, 63.

Yin, Y., Fischer, D., 2008. Identification and investigation of ORFans in the viral world. BMC.Genomics 9, 24.

Yoosuf, N., Yutin, N., Colson, P., Shabalina, S. A., Pagnier, I., Robert, C., Azza, S., Klose, T., Wong, J., Rossmann, M. G., La Scola, B., Raoult, D., Koonin, E. V., 2012. Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the Megavirus lineage. Genome Biol.Evol. 4(12), 1324-1330.

Yoosuf, N., Yutin, N., Colson, P., Shabalina, S. A., Pagnier, I., Robert, C., Azza, S., Klose, T., Wong, J., Rossmann, M. G., La, S. B., Raoult, D., Koonin, E. V., 2012b. Related giant viruses in distant locations and different habitats: Acanthamoeba polyphaga moumouvirus represents a third lineage of the Mimiviridae that is close to the megavirus lineage. Genome Biol Evol. 4, 1324-1330.

Yutin, N., Colson, P., Raoult, D., Koonin, E. V., 2013. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. Virol.J 10, 106.

Yutin, N., Koonin, E. V., 2012. Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. Virol.J. 9, 161.

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

Chapter Three (3.3)

Complete genome sequence of courdo11 virus, a member of the family *Mimiviridae*

Yoosuf N, Pagnier I, Fournous G, Robert C, La Scola B, Raoult D, Colson P

Accepted in Virus Genes

Chapter Three (3.3)

Complete genome sequence of courdo11 virus, a member of the family *Mimiviridae*

A decade ago, the founder member of the family *Mimiviridae*, *Acanthamoeba polyphaga* mimivirus was isolated from the water collected from a cooling tower in England, by co-culturing on *Acanthamoeba* spp (La Scola et al. 2003). Subsequently, several dozens of giant viruses that infect amoeba were isolated over the past decade, both from water and soil samples. *Megavirus chilensis* is the largest virus reported to date among the amoeba-associated mimiviruses (Arslan et al. 2011). Here, we describe the genome of another member of the family *Mimiviridae*, Courdo11 virus, which was isolated in 2010 by inoculating *Acanthamoeba* spp. with freshwater collected from a river in southeastern France. The Courdo11 virus genome is a double stranded DNA of 1,245,674 base pairs that is predicted to encode 1,166 proteins. Comparative genomics and phylogenetic analysis of Courdo11 virus with other members of the family *Mimiviridae* showed

that the Courdo11 virus is a member of lineage C of mimiviruses of amoebae, being most closely related to *Megavirus chilensis* and LBA 111, the first mimivirus isolated from a human (Arslan et al. 2011; Saadi et al. 2013a). We studied the gene content of Courdo11 virus and identified all the major characteristics of *Megavirus chilensis*, including the presence of three additional tRNAs in comparison with lineage

A mimiviruses. In addition, the genome architecture showed the same pattern than that pointed out for mimiviruses in earlier studies. We finally identified fourteen ORFans, which suggests that the pan-genome of mimiviruses might reach a plateau

Virus Genes
DOI 10.1007/s11262-013-1016-x

Complete genome sequence of Courdo11 virus, a member of the family *Mimiviridae*

Niyaz Yoosuf · Isabelle Pagnier · Ghislain Fournous · Catherine Robert · Bernard La Scola · Didier Raoult · Philippe Colson

Received: 30 September 2013 / Accepted: 13 November 2013

Abstract Giant viruses of amoebae were discovered 10 years ago and led to the description of two new viral families: *Mimiviridae* and *Marseilleviridae*. These viruses exhibit remarkable features, including large capsids and genomes that are similar in size to those of small bacteria and their large genetic repertoires include genes that are unique among viruses. The family *Mimiviridae* has grown during the past decade since the discovery of its initial member, Mimivirus, and continues to expand. Here, we describe the genome of a new giant virus that infects *Acanthamoeba* spp., Courdo11 virus, isolated in 2010 by inoculating *Acanthamoeba* spp. with freshwater collected from a river in southeastern France. The Courdo11 virus genome is a double stranded DNA molecule composed of 1,245,674 nucleotides. The comparative analyses of Courdo11 virus with the genomes of other giant viruses showed that it belongs to lineage C of mimiviruses of amoebae, being most closely related to *Megavirus chilensis* and LBA 111, the first mimivirus isolated from a human. Major characteristics of the *M. chilensis* genome were identified in the Courdo11 virus genome, found to encode three more tRNAs. Genomic architecture comparisons mirrored previous findings that showed conservation of collinear regions in the middle part of the genome and diversity towards the extremities. Finally, fourteen ORFans were identified in the Courdo11 virus genome, suggesting that the pan-genome of mimiviruses of amoeba might reach a plateau.

Keywords Courdo11 virus · Mimivirus · *Mimiviridae* · Giant virus · Megavirales · Nucleocytoplasmic large DNA viruses · Amoeba

N. Yoosuf · I. Pagnier · G. Fournous · C. Robert · B. La Scola · D. Raoult · P. Colson (✉)
Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE), UM63, CNRS 7278, IRD 198, INSERM U1095, Facultés de Médecine et de Pharmacie, Aix-Marseille Univ., 27 Boulevard Jean Moulin, 13385 Marseille Cedex 05, France
e-mail: philippe.colson@univ-amu.fr

I. Pagnier · B. La Scola · D. Raoult · P. Colson
Fondation Institut Hospitalo-Universitaire (IHU) Méditerranée Infection, Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone, Assistance Publique – Hôpitaux de Marseille, 264 rue Saint-Pierre, 13385 Marseille Cedex 05, France

Introduction

Giant viruses of amoebae were discovered 10 years ago and led to the description of two new viral families: *Mimiviridae* and *Marseilleviridae* [1–7]. These viruses exhibit remarkable features, including large capsids and genomes that are similar in size to those of small bacteria [7–10]. Furthermore, their large genetic repertoires include genes that encode components of the translation apparatus that are unique among viruses. Even larger amoebae viruses were recently described; the genomes of these viruses are 1.9 and 2.5 kbp in size [11]. Altogether, these findings have altered the definition of a virus [1, 11–13]. The family *Mimiviridae* has grown during the past decade since the discovery of its initial member, Mimivirus, and continues to expand [6, 9, 14]. The first group of this family is composed of mimiviruses that infect *Acanthamoeba* spp., among which three lineages, A, B and C, have been delineated with Mimivirus [1], Moumouvirus [5] and *Megavirus chilensis* [4] as the prototype members, respectively [7]. *Cafeteria roenbergensis* virus is a smaller,

Published online: 01 December 2013

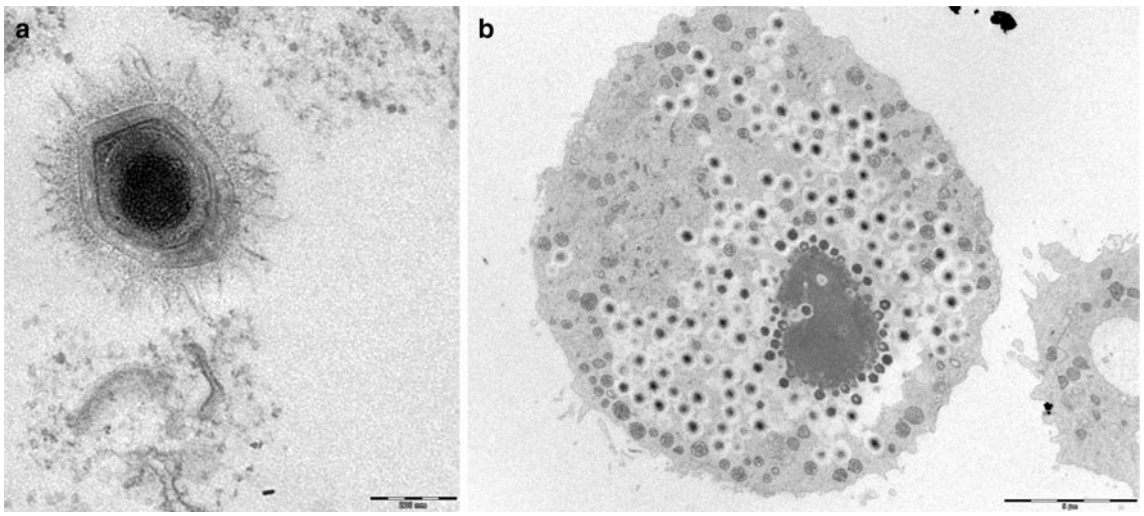


Fig. 1 Electron microscopy of Courdo11 virus particles (**a**, *left*) and *A. polyphaga* infected with Courdo11 virus 16 h post-infection (**b**, *right*). *Scale bars* represent 200 nm (**a**) and 5 μm (**b**)

distantly related mimivirus that infects a marine dinoflagellate [15]. Mimiviruses of amoebae can exchange DNA with other microorganisms that live sympatrically with the mimiviruses inside *Acanthamoeba* spp. [15], and they harbour a specific mobilome [16, 17]. These giant viruses are common in environmental samples, as shown by high rates of isolation from water and by metagenomic analyses [14, 18, 19]. In addition, LBA111 and Shan, two members of the lineage C of amoeba-associated mimiviruses were recently isolated from the bronchoalveolar fluid and stools, respectively, of Tunisian patients presenting pneumonia [20, 21]. These findings indicate that mimiviruses might cause pneumonia [22, 23]. Here, we describe the genome of a new mimivirus that infects *Acanthamoeba* spp.

Materials and methods

Courdo11 virus was isolated in 2010 by inoculating *A. polyphaga*, as previously described [24], with freshwater drawn by one of us (IP) slightly below the surface of the Le Peyron creek in Saint-Raphael city, southeastern France. Its capsid size is approximately 450 nm (Fig. 1). The Courdo11 virus genome was sequenced on a 454-Roche GS20 device as described previously [1, 5], and the obtained reads were assembled de novo with Newbler Assembly software [25]. A second set of reads was obtained using an AB SOLiD instrument and was mapped on the previously assembled genome using CLC Bio software (http://www.clcbio.com/index.php?id=28). Open reading frames were predicted using an in-house pipeline and GeneMarkS software [26]. The genomic architectures were compared using Owen [27], Mauve [28] and MUMmer [29] softwares. The strategy of reciprocal best BLASTp hits [30] was used to identify orthologous sets of genes using the Proteinortho tool [31]. An e-value below 1e-3, an amino acid identity above 30 % and sequence coverage above 70 % were used to consider hits as significant. The tRNAScanSE tool was used to search for transfer RNA genes [32]. BLASTp was performed against the NCBI GenBank non-redundant protein sequence database (nr) with an e-value lower than 1e-3 and alignment length greaterthan 80 amino acids (for alignment lengths <80 amino acids, we used an e-value of 1e-5) to identify ORFans. The same e-value cutoff has been used in previous studies to define ORFans [33]. The sequence alignments were built using the muscle program [34], and the alignments were trimmed using the trimAl tool [35]. The phylogenetic trees were constructed using PhyML [36] and were visualized using FigTree software (http://www.umiacs.umd.edu/~morariu/figtree/).

Results

The Courdo11 virus genome (GenBank Accession No. JX975216) is a double stranded DNA molecule composed of 1,245,674 nucleotides. It is 13,523 nt shorter than the *M. chilensis* genome [4] and is currently the second largest mimivirus genome that has been described. A total of 1,166 predicted proteins were tentatively identified in the Courdo11 virus genome, and 1,085 of these were greater than 100 amino acids in size. The sizes of the Courdo11 virus predicted proteins range from 48 to 2,605 amino acids in length with a mean size of 312 amino acids. The genome has a high coding density of 84 % with a mean distance of 176 nucleotides separating the coding sequences. The predicted genes are evenly distributed on both strands; 597

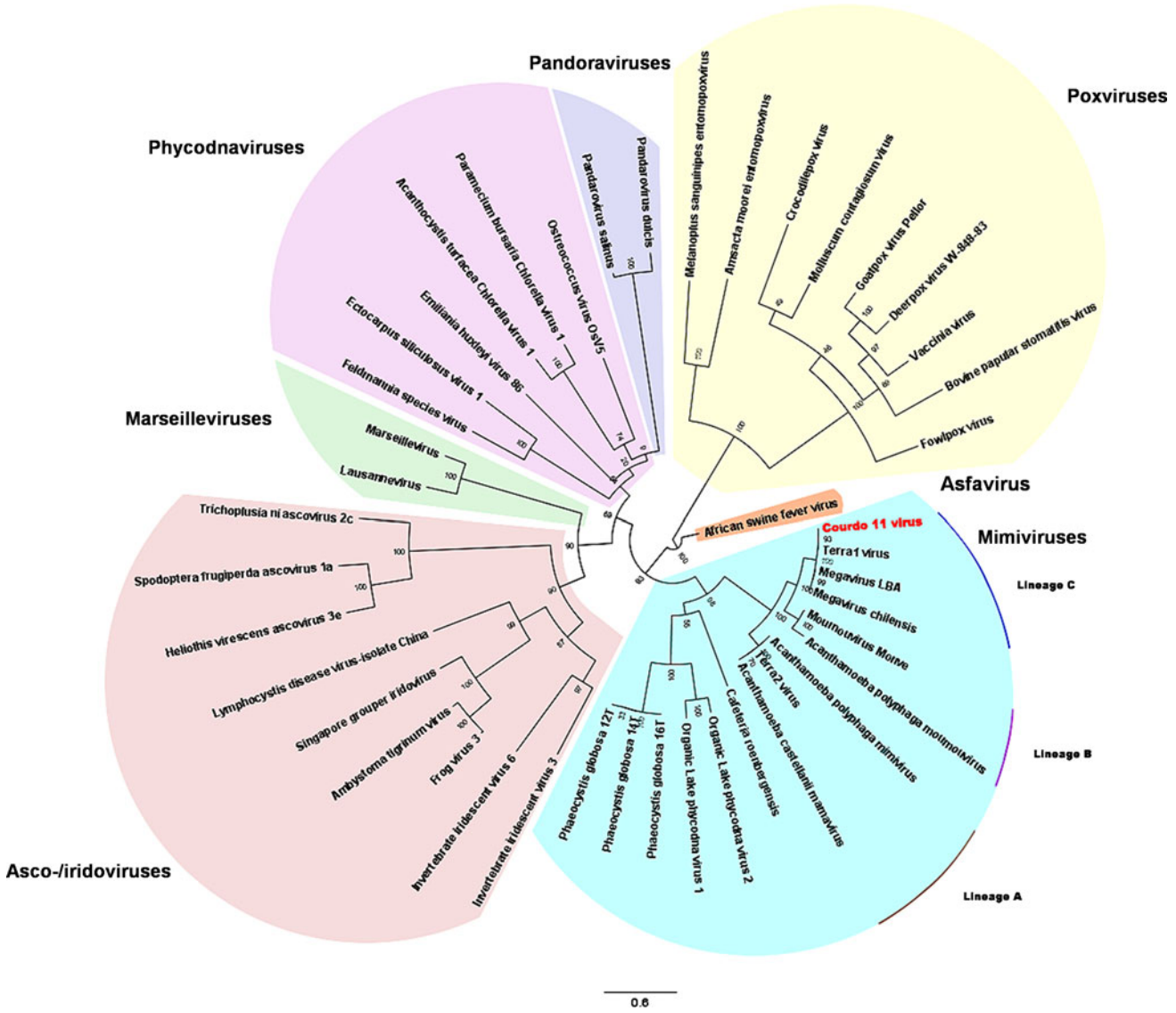


Fig. 2 The phylogenetic tree was constructed based on the family B DNA polymerases from selected members of the order "Megavirales" [7–9] and from *Pandoravirus dulcis* and *P. salinus* [11] using the maximum likelihood method. The sequence alignment was generated using the muscle program [34], and the trimAl tool [35] was used for automated alignment trimming. The phylogenetic tree was constructed using PhyML [36] and visualized with the FigTree software (http://www.umiacs.umd.edu/~morariu/figtree/). The numbers of tree nodes indicate bootstrap replicates of 100

and 569 predicted genes are located on the positive and negative strands, respectively. Phylogenetic reconstructions of the family B DNA polymerase of the proposed order "Megavirales" (that encompasses nucleocytoplasmic large DNA viruses) [7–9] indicates that Courdo11 virus belongs to the lineage C of amoeba-associated mimiviruses [7] and is closely related to *M. chilensis* recovered from marine coastal water in Chile (Fig. 2) [4].

Comparative analyses of Courdo11 virus and other mimiviruses showed that the genomes of Courdo11 virus and *M. chilensis* are highly similar and collinear, and the highest levels of divergence were located near the ends of the Courdo11 virus genome. Overall, the alignment of these two viral genomes using Owen revealed 70 regions larger than 5,000 nt with a mean nucleotide identity of 98.1 % (range, 93.8–99.6) (Fig. 3). A total of 1,137 of the 1,166 predicted proteins of Courdo11 virus (97.5 % of its gene repertoire) are homologous to *M. chilensis* with a mean amino acid identity of 95.5 %. Moreover, Courdo11 virus shares 1,130 (96.9 %) homologous proteins with LBA111. The reciprocal best hits strategy using Proteinortho to compare the Courdo11 virus predicted proteins with those of the members of family *Mimiviridae* showed that the Courdo11 virus shares a high number of orthologous genes, 860 (73.7 %) and 857 (73.4 %) of its gene repertoire, with *M. chilensis* and LBA 111, respectively,

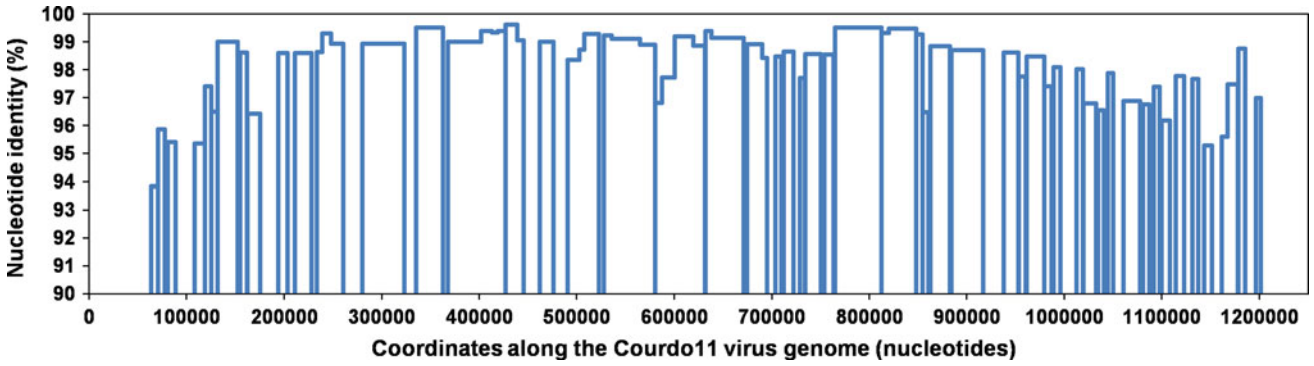


Fig. 3 Nucleotide identity along the Courdo11 virus and *Megavirus chilensis* [4] genomes for the largest (>5,000 nt) matching regions

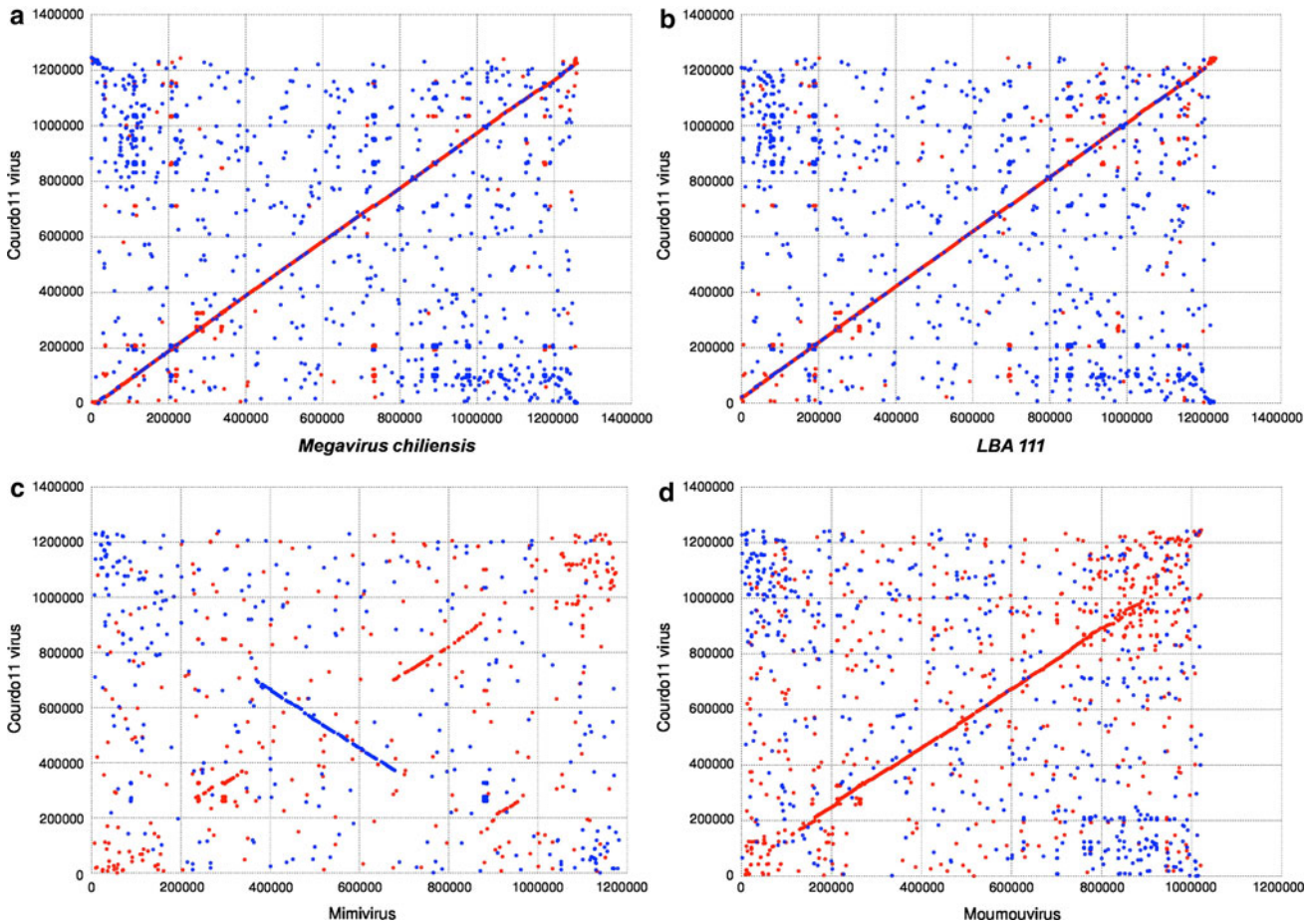


Fig. 4 Genomic dot plots for the Courdo11 virus against amoebae-associated mimiviruses belonging to lineages A, B and C. The Courdo11 virus genome was compared to the **a** *Megavirus chilensis* [4], **b** LBA 111, **c** Mimivirus [1] and **d** Moumouvirus [5] genomes. *Dot plots* were constructed using MUMmer 3.22 software [29], and nucleotide-based alignments were performed with MUMmer. *Dot plots* were generated by the MUMmerplot script and the program gnuplot (www.gnuplot.info/docs_4.0/gnuplot.html). The aligned segments are represented by *dots* or *lines*. Forward matches are shown in *red*, and reverse complement matches are shown in *blue* (Color figure online)

which are members of lineage C of mimiviruses of amoebae. In addition, Courdo11 virus shares 393 (33.7 %) and 450 (38.5 %) orthologous genes with Mimivirus, the founding member of lineage A and with Moumouvirus, the founding member of lineage B, respectively. The dot plots of gene repertoires for the Courdo11 virus genome with the genomes of members of lineage C, *M. chilensis* and LBA111, shows high levels of synteny and a near-perfect collinearity (Fig. 4a, b). The genomic dot plot of Courdo11 virus against Mimivirus [1] shows shorter interrupted

collinear regions with a large inversion in the central part of the genome (Fig. 4c), whereas the dot plot against Moumouvirus [5] shows a near-perfect collinearity in the middle part of the genome and rearrangements towards the extremities (Fig. 4d).

Overall, BLASTp searches against the NCBI GenBank non-redundant protein sequence database identified *bona fide* homologues for 1,152 Courdo11 virus predicted proteins (99 % of the gene repertoire), and 14 (1.2 %) ORFans were identified. The main components of the *M. chilensis* gene content are also found in the Courdo11 virus genome, including three amino acyl-tRNA synthetases (mg743, mg844, mg358) that are absent in Mimivirus, a putative DNA photolyase (mg779) and a uridine monophosphate kinase (mg431). In contrast, 6 tRNAs were detected in the Courdo11 virus genome, including three tRNA-Leu, one tRNA-Cys, one tRNA-His and one tRNA-Trp, whereas only 3 tRNAs were identified in *M. chilensis* (1 Trp, and 2 Leu). We extended our analysis of the Courdo11 virus by comparing it with the newly identified viruses *Pandoravirus dulcis* and *Pandarovirus salinus* [11]. A BLASTp search of the Courdo11 virus gene content against those of pandoraviruses using 1e-3 as e-value cut off yielded 150 (12.8 %) and 132 (11.3 %) significant hits with *P. salinus* and *P. dulcis,* respectively.

Discussion

The comparative analyses of Courdo11 virus with the genomes of other members of family *Mimiviridae* showed that this giant virus is a *bona fide* new member of the family *Mimiviridae* and belongs to lineage C of mimiviruses of amoebae. Genomic architecture comparisons mirrored previous findings that showed conservation of collinear regions in the middle part of the genome and diversity towards the extremities; this feature was indeed described in other mimiviruses and in poxviruses [4, 5, 37–39]. Further analyses showed that the Courdo11 virus genome is most closely related to *M. chilensis* [4] and LBA 111 [20], the first mimivirus isolated from a human. The evolutionary relationship between Courdo11 virus and other mimiviruses isolated from freshwater (Mimivirus and Moumouvirus), marine coastal water (*M. chilensis*) and soil (Terra1 virus and Terra2 virus) indicates that these giant viruses have a broad host range and can survive in different habitats. Major characteristics of the *M. chilensis* genome were identified in the Courdo11 virus genome, which was found to encode three more tRNAs. Fourteen ORFans were identified in the Courdo11 virus genome, suggesting that the pan-genome of mimiviruses of amoeba might reach a plateau.

References

1. D. Raoult, S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. La Scola, M. Suzan, J.M. Claverie, Science **306**, 1344–1350 (2004)
2. M. Boyer, N. Yutin, I. Pagnier, L. Barrassi, G. Fournous, L. Espinosa, C. Robert, S. Azza, S. Sun, M.G. Rossmann, M. Suzan-Monti, B. La Scola, E.V. Koonin, D. Raoult, Proc. Natl. Acad. Sci. USA **106**, 21848–21853 (2009)
3. V. Thomas, C. Bertelli, F. Collyn, N. Casson, A. Telenti, A. Goesmann, A. Croxatto, G. Greub, Environ. Microbiol. **13**, 1454–1466 (2011)
4. D. Arslan, M. Legendre, V. Seltzer, C. Abergel, J.M. Claverie, Proc. Natl. Acad. Sci. USA **108**, 17486–17491 (2011)
5. N. Yoosuf, N. Yutin, P. Colson, S.A. Shabalina, I. Pagnier, C. Robert, S. Azza, T. Klose, J. Wong, M.G. Rossmann, B. La Scola, D. Raoult, E.V. Koonin, Genome Biol. Evol. **4**, 1324–1330 (2012)
6. B. La Scola, A. Campocasso, R. N'Dong, G. Fournous, L. Barrassi, C. Flaudrops, D. Raoult, Intervirology **53**, 344–353 (2010)
7. P. Colson, X. de Lamballerie, G. Fournous, D. Raoult, Intervirology **55**, 321–332 (2012)
8. P. Colson, X. de Lamballerie, N. Yutin, S. Asgari, Y. Bigot, D.K. Bideshi, X.W. Cheng, B.A. Federici, J.L. Van Etten, E.V. Koonin, B. La Scola, D. Raoult, Arch. Virol. (2013). doi:10.1186/1743-422X-10-158
9. N. Yutin, P. Colson, D. Raoult, E.V. Koonin, Virol. J. **10**, 106 (2013)
10. N. Yutin, E.V. Koonin, Virol. J. **9**, 161 (2012)
11. N. Philippe, M. Legendre, G. Doutre, Y. Coute, O. Poirot, M. Lescot, D. Arslan, V. Seltzer, L. Bertaux, C. Bruley, J. Garin, J.M. Claverie, C. Abergel, Science **341**, 281–286 (2013)
12. D. Raoult, P. Forterre, Nat. Rev. Microbiol. **6**, 315–319 (2008)
13. E.V. Koonin, N. Yutin, Intervirology **53**, 284–292 (2010)
14. I. Pagnier, D.G. Reteno, H. Saadi, M. Boughalmi, M. Gaia, M. Slimani, T. Ngounga, M. Bekliz, P. Colson, D. Raoult, B. La Scola, Intervirology **56**, 354–363 (2013)
15. M.G. Fischer, M.J. Allen, W.H. Wilson, C.A. Suttle, Proc. Natl. Acad. Sci. USA **107**, 19508–19513 (2010)
16. C. Desnues, B. La Scola, N. Yutin, G. Fournous, C. Robert, S. Azza, P. Jardot, S. Monteil, A. Campocasso, E.V. Koonin, D. Raoult, Proc. Natl. Acad. Sci. USA **109**, 18078–18083 (2012)
17. N. Yutin, D. Raoult, E.V. Koonin, Virol. J. **10**, 158 (2013)
18. P. Colson, B. La Scola, D. Raoult, Intervirology **56**, 376–385 (2013)
19. P. Colson, D. Raoult, Megavirales composing a fourth domain of life: Mimiviridae and Marseilleviridae, *Viruses: essential agents of life* (Springer, Dordrecht, 2012)
20. H. Saadi, I. Pagnier, P. Colson, J. Kanoun Cherif, M. Beji, M. Boughalmi, S. Azza, N. Armstrong, C. Robert, G. Fournous, B. La Scola, D. Raoult, Clin. Infect. Dis. **57**, e127–e134 (2013)
21. H. Saadi, D.G. Ikanga Reteno, P. Colson, S. Aherfi, P. Minodier, I. Pagnier, F. Fenollar, C. Robert, D. Raoult, B. La Scola, Intervirology **56**, 424–429 (2013)
22. A. Vincent, B. La Scola, L. Papazian, Intervirology **53**, 304–309 (2010)
23. P. Colson, L. Fancello, G. Gimenez, F. Armougom, C. Desnues, G. Fournous, N. Yoosuf, M. Million, B. La Scola, D. Raoult, J. Clin. Virol. **57**(3), 191–200 (2013)
24. B. La Scola, C. Desnues, I. Pagnier, C. Robert, L. Barrassi, G. Fournous, M. Merchat, M. Suzan-Monti, P. Forterre, E. Koonin, D. Raoult, Nature **455**, 100–104 (2008)
25. M. Margulies, M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen,

S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, J.M. Rothberg, Nature **437**, 376–380 (2005)

26. J. Besemer, M. Borodovsky, Nucleic Acids Res. **33**, W451–W454 (2005)
27. A.Y. Ogurtsov, M.A. Roytberg, S.A. Shabalina, A.S. Kondrashov, Bioinformatics **18**, 1703–1704 (2002)
28. A.C. Darling, B. Mau, F.R. Blattner, N.T. Perna, Genome Res. **14**, 1394–1403 (2004)
29. A.L. Delcher, S.L. Salzberg, A.M. Phillippy, Curr. Protoc. Bioinformatics (2003). doi:10.1002/0471250953.bi1003s00. chapter 10:Unit 10.3
30. I.K. Jordan, I.B. Rogozin, Y.I. Wolf, E.V. Koonin, Genome Res. **12**, 962–968 (2002)
31. M. Lechner, S. Findeiss, L. Steiner, M. Marz, P.F. Stadler, S.J. Prohaska, BMC Bioinformatics **12**, 124 (2011)
32. P. Schattner, A.N. Brooks, T.M. Lowe, Nucleic Acids Res. **33**, W686–W689 (2005)
33. Y. Yin, D. Fischer, BMC Genomics **9**, 24 (2008)
34. R.C. Edgar, Nucleic Acids Res. **32**, 1792–1797 (2004)
35. S. Capella-Gutierrez, J.M. Silla-Martinez, T. Gabaldon, Bioinformatics **25**, 1972–1973 (2009)
36. S. Guindon, F. Lethiec, P. Duroux, O. Gascuel, Nucleic Acids Res. **33**, W557–W559 (2005)
37. T.G. Senkevich, E.V. Koonin, J.J. Bugert, G. Darai, B. Moss, Virology **233**, 19–42 (1997)
38. A. McLysaght, P.F. Baldi, B.S. Gaut, Proc. Natl. Acad. Sci. USA **100**, 15655–15660 (2003)
39. P. Colson, G. Gimenez, M. Boyer, G. Fournous, D. Raoult, PLoS One **6**, e18935 (2011)

Chapter Four

Evidence of megavirome in humans

Colson P, Fancello L, Gimenez G, Armougom A,

Desnues C, Fournous G, Yoosuf N, Million M,

Scola B, Raoult D

Chapter Four

Evidence of megavirome in humans

The *Megavirales* is a proposed new viral order that encompasses members of families *Poxviridae*, *Asfarviridae*, *Ascoviridae*, *Iridoviridae*, *Phycodnaviridae*, *Mimiviridae* and *Marseilleviridae* (Colson et al. 2012). Previously, these viral families were assigned to Nucleocytoplasmic large DNA viruses (NCLDVs) (Iyer et al. 2001; Iyer et al. 2006). There is a growing body of evidence supporting the role of Mimivirus, discovered while investigating a pneumonia outbreak using amoebal coculture, as a causative agent of pneumonia (Saadi et al. 2013a; Saadi et al. 2013b). The *Megavirales* virome, we referred to as the "megavirome", has been likely neglected because of technical steps prior sequencing, especially the use of filters with a pore size in the range 0.2–0.45-µm, which can prevent finding giant viruses. A study conducted in our laboratory based on ultra-deep sequencing of bacterial 16S ribosomal DNA (rDNA) from the stools of a healthy 20-year-old man living in rural Senegal included an alternative approach to the classical method that aimed at avoiding the PCR amplification bias. This strategy consisted of the complete enzymatic digestion of the fecal sample DNA with EcoO190I and BrsGI enzymes that are able to cleave sites inside primers and

generated fragments corresponding to 16S rDNA. Besides, most sequencing reads were unrelated to bacterial 16S rDNA and some were related to Mimivirus and Marseillevirus sequences. Subsequently, a new marseillevirus was isolated from the fecal sample. In addition, we searched for and detected sequences related to *Megavirales* members, including mimiviruses, in human metagenomes available from public databases. Taken together, these findings provide converging evidence of the presence of mimiviruses and marseilleviruses in humans.

Journal of Clinical Virology 57 (2013) 191–200

Contents lists available at SciVerse ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Evidence of the megavirome in humans

Philippe Colson[a,b], Laura Fancello[a], Gregory Gimenez[a], Fabrice Armougom[a], Christelle Desnues[a], Ghislain Fournous[a], Niyaz Yoosuf[a], Matthieu Million[a], Bernard La Scola[a,b], Didier Raoult[a,b,*]

[a] Aix-Marseille Univ., Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE) UM 63 CNRS 7278 IRD 198 INSERM U1095, Facultés de Médecine et de Pharmacie, Marseille, France
[b] IHU Méditerranée Infection, Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, Centre Hospitalo-Universitaire Timone, Assistance Publique – Hôpitaux de Marseille, Marseille, France

ARTICLE INFO

Article history:
Received 7 January 2013
Received in revised form 14 March 2013
Accepted 29 March 2013

Keywords:
Mimivirus
Giant virus
Marseillevirus
Megavirales
Metagenomics
Humans
Next-generation sequencing
Amoeba
Infectious diseases

ABSTRACT

Background: Megavirales is a proposed new virus order composed of Mimivirus, Marseillevirus and closely related viruses, as well as members of the families *Poxviridae*, *Iridoviridae*, *Ascoviridae*, *Phycodnaviridae* and *Asfarviridae*. The *Megavirales* virome, which we refer to as the megavirome, has been largely neglected until now because of the use of technical procedures that have jeopardized the discovery of giant viruses, particularly the use of filters with pore sizes in the 0.2–0.45-μm range. Concurrently, there has been accumulating evidence supporting the role of Mimivirus, discovered while investigating a pneumonia outbreak using amoebal coculture, as a causative agent in pneumonia.
Objectives: In this paper, we describe the detection of sequences related to Mimivirus and Marseillevirus in the gut microbiota from a young Senegalese man. We also searched for sequences related to *Megavirales* in human metagenomes publicly available in sequence databases.
Results: We serendipitously detected Mimivirus- and Marseillevirus-like sequences while using a new metagenomic approach targeting bacterial DNA that subsequently led to the isolation of a new member of the family *Marseilleviridae*, named Senegalvirus, from human stools. This discovery demonstrates the possibility of the presence of giant viruses of amoebae in humans. In addition, we detected sequences related to *Megavirales* members in several human metagenomes, which adds to previous findings by several groups.
Conclusions: Overall, we present convergent evidence of the presence of mimiviruses and marseilleviruses in humans. Our findings suggest that we should re-evaluate the human megavirome and investigate the prevalence, diversity and potential pathogenicity of giant viruses in humans.

1. Background

The story of giant viruses that infect phagocytic protists began in 1992 in England during the investigation of a pneumonia outbreak that led to isolate obligate intra-amoebal microorganisms in water of a cooling tower.[1–3] Then, in 2003, Mimivirus was discovered as part of this collection of intra-amoebal parasites.[1–4] Subsequently, Mamavirus, Marseillevirus and other giant viruses infecting phagocytic protists have been discovered; all of these viruses have been linked to nucleocytoplasmic large DNA viruses (NCLDVs), a monophyletic class of viruses composed of *Poxviridae*, *Iridoviridae*, *Ascoviridae*, *Phycodnaviridae* and *Asfarviridae*[5–11] for which we recently proposed the reclassification into a new viral order, the *Megavirales*.[12]

The question of pathogenicity of mimiviruses initially focused on the capability of Mimivirus to cause pneumonia due to the setting of its initial discovery as well as the involvement of some water-associated amoebae-resistant bacteria, including *Legionella pneumophila*, in such infections.[1,13,14] Experimentally, Mimivirus was found to be capable of inducing pneumonia in mice[15] and infecting macrophages through phagocytosis.[16] In humans, serological studies showed seroconversion to Mimivirus in several patients presenting with pneumonia.[13,18] Antibodies to Mimivirus were associated with pneumonia, re-hospitalization after discharge[13] and a poorer outcome in mechanically ventilated pneumonia patients (Table 1).[19] In contrast, several studies have failed to isolate Mimivirus from patients with pneumonia and Mimivirus DNA testing was positive in only a single patient.[13,20–23,25] However, DNA detection may have been hampered in these studies

* Corresponding author at: Aix-Marseille Université, Unité des Rickettsies, URMITE UM3 CNRS 7278 IRD 198 INSERM U1095, Faculté de Médecine, 27 Boulevard Jean Moulin, 13385 Marseille Cedex 05, France. Tel.: +33 491 324 375; fax: +33 491 387 772.
E-mail address: didier.raoult@gmail.com (D. Raoult).

1386-6532/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jcv.2013.03.018

Table 1
Summary of the clinical, microbiological and metagenomic data on mimivirus and pneumonia.

Reference	Year(s) of sample collection	Country	Population size	Subgroup size	Main characteristics of patients	Respiratory sample type	PCR	Serology
13	1985–1997	Canada	887	121	Adults; ambulatory/com-acquired pneumonia patients, Nova Scotia	N.t.	N.t.	36 positive (9.7%)
				255	Adults; patients hospitalized with com-acquired pneumonia, multiple centers across Canada			
				511	Adults; healthy control subjects, Nova Scotia	N.t.	N.t.	12 positive (2.3%)
	2003–2004	France	129	32	Adults; patients with ICU-acquired pneumonia, one-year survey	BAL	1 positive sample of 32 (3.1%)	5 positive samples of 26 (19.2%)
				50	Adults; controls (patients tested for anti-*Rickettsia* spp. antibodies)	N.t.	N.t.	All negative
				21	Adults; intubated control patients in the ICU who did not present with pneumonia	BAL	All negative	N.t.
17	2003	France	157		Adults; ICU pneumonia patients (pneumonia was com-acquired or ventilator-associated)	N.t.	N.t.	7 cases with a high level of evidence and 7 additional cases with a low level of evidence
18	2004	France	1		38-year-old laboratory technician	N.t.	N.t.	Positive
19	2006–2008	France	300		Adults; ventilated patients in the ICU with a suspicion of a ventilator-associated pneumonia and positive serology for Mimivirus (cases)	N.t.	N.t.	59 positive (19.7%)
20	2005–2006	Austria	214		Children hospitalized for respiratory tract infections; 209 were non-immunocompromised; six-month survey during the fall and winter seasons	NP aspirate samples	All negative	N.t.
21	2000–2001	Urban USA	496	124	Children < 5 y; com-acquired pneumonia cases	Nasal swabs	All negative	N.t.
	2003–2004	Rural Thailand		120	Adults, children; com-acquired pneumonia cases	NP swabs	All negative	N.t.
	2002–2004	USA		71	Geriatric; nosocomially acquired pneumonia outbreak, retirement centers	Nasal or NP swabs	All negative	N.t.
		USA		5	Adults, children; com-acquired pneumonia outbreak (familial cluster)	Lower respiratory samples	All negative	N.t.
	2001–2003	USA		87	Adults; bone marrow transplant recipients	NP aspirates, nasal wash, or NP swabs	All negative	N.t.
	2002–2003	Canada		89	Adults; lung transplant recipients	NP swabs	All negative	N.t.
22	–	Urban Italy	63 (69 BAL) specimens	30 BAL	Patients receiving mechanical ventilation for at least 48 h	BAL	All negative	N.t.
				39 BAL	Non-ventilated patients from different clinical settings, including lung transplant recipients	BAL	All negative	N.t.
23	2003–2004	Queensland, Australia	315 (477 specimen)		People with suspected acute respiratory tract infections; the subjects ranged in age from 1 day to 80.3 years (mean = 7.7 years); 79% were ≤5 years of age; an additional 81 consecutively collected summer specimens formed a secondary population	Predominantly (92.4%) nasopharyn-geal aspirates	All negative	N.t.
24	2004–2005	Stockholm, Sweden	210		Patients with respiratory tract infections	NP aspirates	Metagenomics identified 2 reads matching mimivirus	–
25	2009–2010	The Netherlands	109 (220 sputum samples)		COPD patients during stable conditions and during exacerbations of COPD, referred for pulmonary rehabilitation 115 sputum samples were collected from 84 patients during the stable phase, and 105 samples were collected from 74 patients during an acute exacerbation Mimivirus serology was performed for 118 serum samples, 88 collected during the stable phase and 30 during an exacerbation	Sputum	All negative	3 positive (2 during an acute exacerbation; 1 during the stable phase)

BAL, Bronchoalveolar lavage; Com., Community; COPD, chronic obstructive pulmonary disease; NP, naso-pharyngeal.

Table 2
Studies that identified sequences closely related to those of *Megavirales* in human and animal samples.

Reference	Mean	Enrichment in viruses	Sample	Name of virus(es) (Number of hits)	City, continent, country
41	Shotgun library (with strand displacement polymerase amplification), standard sequencing	Cesium chloride gradient	Human blood from 20 healthy donors	Cowpox virus (2)	San Diego, California, USA
42	Random primer amplification, standard sequencing	Filtration through 0.45-μm filters	Fecal samples from 12 distinct pediatric patients suffering from acute diarrhea	Mimivirus (5)	Melbourne, Australia; Seattle, USA
43	454 pyrosequencing	Cesium chloride gradient for sewage (not described for serum)	Sewage and human sera from 199 healthy volunteers and patients with acute febrile illness	*Asfarviridae* (6)	Serum: Middle East; Sewage: Barcelona, Spain
44	454 pyrosequencing	Filtration through 0.22- or 0.45-μm filters	Serum samples from 45 pairs of monozygotic twins affected and unaffected with chronic fatigue illness	*Asfarviridae* (1), *Iridoviridae* (3), *Mimiviridae* (2), *Phycodnaviridae* (2), *Poxviridae* (5)	Sweden
45	Illumina	None described	Human sera from 123 Nicaraguan patients presenting with dengue-like symptoms (but testing negative for dengue virus)	*Asfarviridae*	Nicaragua
24	454 pyrosequencing	Filtration through 0.22- and 0.45-μm filters	210 nasopharyngeal aspirate samples from patients with respiratory tract infection	Mimivirus (2), *Paramecium bursaria* Chlorella virus A1 (2), African swine fever virus (2)	Stockholm, Sweden
46	454 pyrosequencing or Illumina sequencing	None described	50 nasopharyngeal swabs and 23 plasma samples from children under 3 years of age with unexplained fever	*Asfarviridae*	Washington DC, USA
47	Retrospective study of large metagenomic datasets	–	Human gut	Virophages (65)	–
Present study	454 pyrosequencing	None	Human stools (nondiarrheic)	Marseillevirus (9), Mimivirus (44)	Senegal
48	454 pyrosequencing	Tangential flow filtration, ultrafiltration and cesium chloride gradient	Stools from 5 pigs	*Mimiviridae* (0.11% of reads), *Poxviridae* (0.52%), *Iridoviridae* (0.06%), *Phycodnaviridae* (0.58%)	Berlin, Germany
49	454 pyrosequencing	Filtration through 0.45-μm filters and cesium chloride gradient	Three mosquitoes	*Poxviridae*	San Diego, California, USA
50	454 pyrosequencing (transcriptomics)	None described	Gypsy moth (*Lymantria dispar*)-derived IPLB-Ld652Y cell line	Mimivirus, *Cafeteria roenbergensis* virus BV-PW1 *Poxviridae*,	Baltimore, USA

because the PCR primers used targeted only the Mimivirus genome, whereas currently at least 18 close relatives of Mimivirus that exhibit considerable genetic diversity have been described.[12,26,27] Besides, positive serology for the virophage of mimiviruses, initially described in 2008,[6] was recently observed in two people who experienced fever while returning from Laos.[28] In addition, a mimivirus named Lentillevirus has been isolated from contact lens storage case liquid.[29,30]During the past decade, mimiviruses and marseilleviruses have been isolated from freshwater, seawater, and soil samples in six countries located on three continents (http://maps.google.fr/maps/ms?vps=2&hl=fr&i.e.=UTF8&oe=UTF8&msa=0&msid=200914559094835369589.0004beba4af112f60dcf2), and isolation rates reached ≈20% from water samples,[31] suggesting common exposure to these viruses of humans. In addition, the currently identified hosts of these viruses are widespread in water and soil,[14] and Mimivirus-like particles have been observed using light microscopy within Acanthamoeba species in treated sewage sludge from a wastewater treatment plant in the UK.[32] Concurrently, searches for *Megavirales* sequences in multiple environmental metagenomes enabled the identification of sequences similar to those of Mimivirus[33–39] and other members of the families *Asfarviridae*, *Poxviridae*, *Phycodnaviridae* and *Iridoviridae*.[36–39] Furthermore, sequences related to *Megavirales* members as well as virophages described to infect mimiviruses[40] have also been retrieved from human and animal metagenomes (Table 2).[24,41–50]

2. Objectives

In this paper, we describe the detection of sequences related to Mimivirus and Marseillevirus in the gut microbiota from a young Senegalese man. We also searched for sequences related to *Megavirales* in human metagenomes publicly available in sequence databases.

3. Study design

3.1. Investigation of the gut microbiota from a young Senegalese man

A previous study conducted in our laboratory consisted of ultra-deep sequencing of bacterial 16S ribosomal DNA (rDNA) in the stools of a healthy 20-year-old man living in rural Senegal.[51] A new method to avoid PCR amplification bias prior to sequencing was used in addition to the classical method based on PCR amplification of the V6 region of 16S rDNA with universal primers 917F and 1391R.[51] The new sequencing method consisted of complete enzymatic digestion of the fecal sample DNA with *Eco*O190I and *Brs*GI enzymes that are able to cleave sites inside primers 917F and 1391R and are therefore able to generate fragments corresponding to the 16S rDNA V6 region (see supplementary methods). Sequencing of the products from

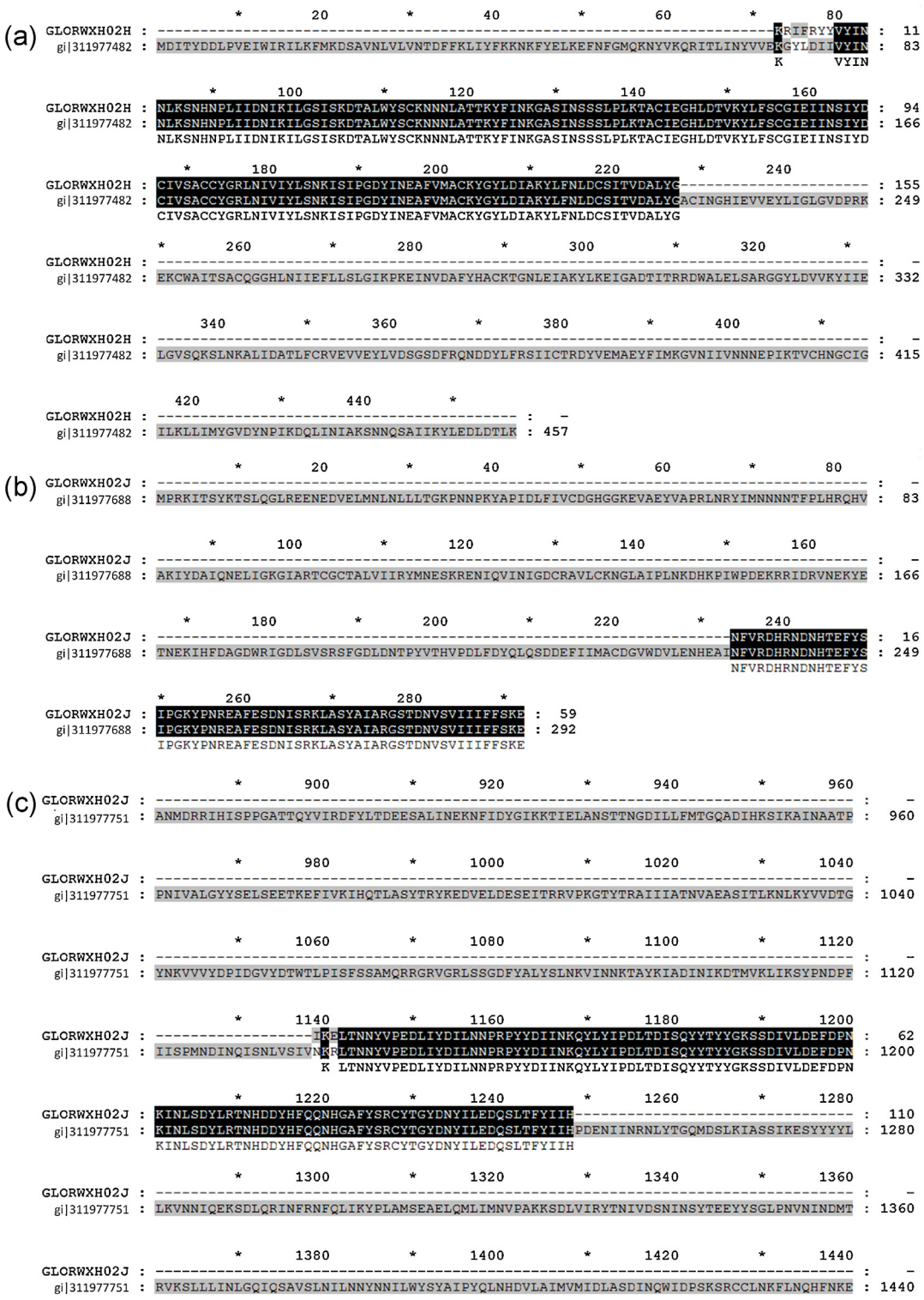


Fig. 1. Alignments of amino acid sequences of Mimivirus (a putative ankyrin repeat protein (gi|311977482) (a), a putative protein phosphatase 2C (gi|311977688) (b), and a putative ATP-dependent RNA helicase gi|311977751) (c)) and three different metagenomic reads obtained from the feces of a young Senegalese. The representation was built using the GeneDoc software (http://www.psc.edu/biomed/genedoc).

both the new and classical procedures was performed using the 454 FLX Titanium instrument (Roche, USA).[51] Sequencing products were trimmed and analyzed by BLAST searches[52] and with the QIIME pipeline.[53] Among the reads generated by enzymatic digestion, 99.9% were unrelated to bacterial 16S rDNA. These results led us to search for Mimivirus and Marseillevirus-related sequences among the initially rejected reads. The reads were mapped onto the Mimivirus and Marseillevirus genomes with CLC bio software (http://www.clcbio.com/index.php?id=479) using default parameters (50% minimum coverage, 80% minimum similarity) and tBLASTn searches were performed with the giant viruses against the reads (*e*-value threshold 1*e*−6). Senegalvirus isolation,

sequencing and assembly have been previously described.[51] Senegalvirus genome annotation is described in the supplementary methods.

3.2. Searches for Megavirales-related sequences in metagenomes recovered from human samples

Reads annotation was automatically performed from the metagenomics RAST (MG-RAST) server (http://metagenomics.anl.gov/) for viral metagenomes from eleven published studies and obtained from human samples, including stools,[54–59] nasopharyngeal aspirates,[24,60] saliva,[61] oropharyngeal swabs,[62] and sputum[63] (supplementary Table S1) against the NCBI GenBank protein sequence database (*e*-value threshold, 1*e*−5). Only hits with alignment lengths ≥40 amino acids were considered. In addition, metagenomes from 9 of the previous studies, and from two other studies that analyzed lung samples,[64–65] were downloaded and preprocessed using several tools including Prinseq[66] for removal of duplicate reads as well as low quality and low complexity reads (supplementary Fig. S1). The remaining sequences were annotated with an in-house strategy using BLASTn searches[52] against all genomes of the *Megavirales* members and virophages available in the NCBI GenBank sequence database as well as those not yet released, available in our laboratory (supplementary Table S2). We considered only matches with ≥30% coverage and ≥90% identity and identical coverage and identity for the corresponding reads through BLAST searches against the NCBI GenBank nucleotide sequence database. BLASTn searches were also performed for all genomes of members of the order *Megavirales* and virophages (supplementary Table S2) against four human gut bacterial metagenomes (supplementary Table S3) through the CAMERA portal (http://camera.calit2.net/), using default parameters. Metagenome reads identified as matching with *Megavirales* sequences were extracted and manually tested against the NCBI nonredundant protein sequence database (nr) using BLASTx (*e*-value threshold, 1*e*−05) to determine whether a *Megavirales* sequence was among the best matches.

Finally, amino acid BLAST (BLASTp) searches were performed for the published proteomes of mimiviruses (Mimivirus, Mamavirus, *Cafeteria roenbergensis* virus (Crov; isolated from *Cafeteria roenbergensis*, a widespread marine unicellular flagellate), *Megavirus chilensis*), marseilleviruses (Marseillevirus, Lausannevirus) and the Sputnik virophage against annotated bacterial metagenomes recovered from eleven different body sites ((Supplementary Table S3; Table 5) as part of the human microbiome project (http://www.hmpdacc.org/HMGI/). Hits were considered significant based on the *e*-value threshold of 1*e*−04 and amino acid identity and coverage above 30% and 70%, respectively.

4. Results

4.1. Serendipitous identification of Mimivirus- and Marseillevirus-related sequences in the stool metagenome of a young Senegalese male

Among the metagenomic reads recovered from the feces of a young Senegalese male,[51] 44 and 9 (supplementary Table S4) could be mapped to the Mimivirus and Marseillevirus genomes, respectively, and 12 reads > 300 bp could be mapped to Mimivirus DNA with >90% identity and coverage (supplementary Table S4; Table 3; Fig. 1a–c). In addition, tBLASTn searches with the Mimivirus and Marseillevirus genomes against the stool metagenome yielded 2306 and 259 hits, respectively. These findings prompted us to inoculate one gram of the young Senegalese stools on *Acanthamoeba polyphaga*, as previously reported,[51] with the purpose of confirming the presence of the giant virus from amoeba in the human feces. Indeed, amoebal culture enabled the isolation of a new marseillevirus, named Senegalvirus (Fig. 2), and the sequencing of its genome.[51] The Senegalvirus double-stranded DNA genome is ≈372,690 base pairs (bp) in length, currently making this genome the largest among marseilleviruses; it is ≈4200 bp larger than that of Marseillevirus and ≈26,000 bp larger than that of Lausannevirus. Comparison of the 479 protein sequences of Senegalvirus predicted using GeneMarkS[67] to those of Marseillevirus or Lausannevirus by all-against-all BLASTp searches yielded bidirectional best hits for 351 and 253 Marseillevirus and Lausannevirus proteins, respectively. Thus, overall, 384 Senegalvirus proteins could be considered *bona fide* orthologs to Marseillevirus or Lausannevirus proteins, because these Senegalvirus proteins are involved in pairs of bidirectional best hits with predicted proteins of Marseillevirus and/or Lausannevirus. The mean (±standard deviation (SD)) amino acid identity between Senegalvirus and Marseillevirus protein in pairs was 97 ± 7%, whereas the mean identity for Senegalvirus and Lausannevirus protein pairs was 59 ± 16%. We detected Senegalvirus orthologs to three histone-like proteins first described in Marseillevirus, as well as proteins containing bacterial-like membrane occupation and recognition nexus (MORN) repeat domains (proteins described as mediating membrane-membrane or membrane-cytoskeleton interactions[5]) and serine/protein kinases, including a unique kinase shared by marseilleviruses, iridoviruses and ascoviruses.[5,8] Homology for the Senegalvirus proteins was greater with their Marseillevirus counterparts than their Lausannevirus counterparts. Congruently with comparative genomics, phylogeny reconstruction based on the family-B DNA polymerase showed that Senegalvirus was clustered with Marseillevirus within the family *Marseilleviridae* (Fig. 3).

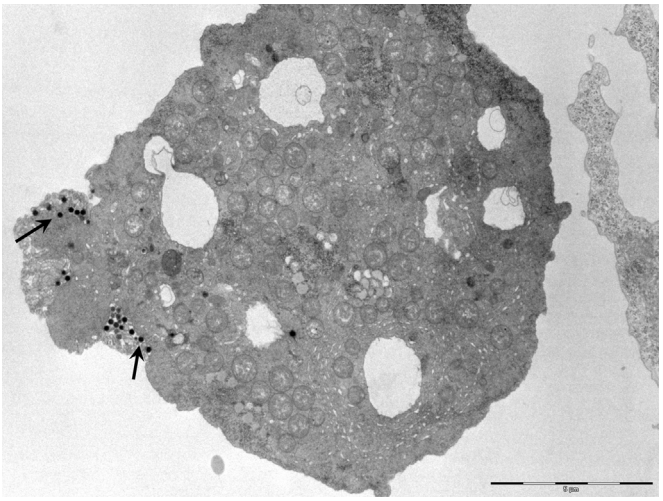


Fig. 2. Electron microscopy image of Senegalvirus in *Acanthamoeba polyphaga*. The scale bar represents 5 μm.

4.2. Blast searches for Megavirales-like sequences in metagenomes

A few reads from human stools and oropharyngeal viromes[58,61,62] available on the MG-RAST server were found to match Mimivirus sequences. They were predicted to encode a collagen-like protein 6 (MIMI_L668), an uncharacterized HNH endonuclease (MIMI_L245) and a hypothetical protein (MIMI_R892) (Supplementary Table S6). Although a BLASTx search using these metagenomic reads against the NCBI GenBank non-redundant protein sequence database did not find *Acanthamoeba polyphaga*

Table 3
Description of reads longer than 300 nucleotides that map to Mimivirus sequences.

Read length (nt)	Best matches	BLASTn (nucleotide)		BLASTp (amino acid)	
		Eval	Identities	*e*-Val	Identities
504	YP_003986602.1 (L112): putative ankyrin repeat protein	0.0	504/505 (99%)	8*e*−154	148/148 (100%)
403	YP_003986629.1 (L137): Hypothetical proteins	0.0	399/405 (99%)	2*e*−99	82/83 (99%)
334	YP_003986695 (L199): Hypothetical proteins	9*e*−157	329/339 (97%)	6*e*−82	94/119 (79%)
361	YP_003986808 (R306/R307): putative protein phosphatase 2C	0.0	361/363 (99%	2*e*−50	59/59 (100%)
350	YP_003986871.1 (R366): putative ATP-dependent RNA helicase	4*e*−180	349/350 (99%)	3*e*−109	107/107 (100%)
475*	AEJ34636.1 (R398): hypothetical protein	0.0	467/476 (98%)	6*e*−49	44/44 (100%)
397*	YP_003986902.1 (R398): putative phosphoesterase	0.0	393/398 (99%)	9*e*−37	32/32 (100%)
447	YP_003987107.1 (R592): putative helicase	0.0	442/447 (99%)	3*e*−90	88/148 (59%)
403	YP_003987195.1 (L673): putative serine/threonine-protein kinase	0.0	400/403 (99%)	1*e*−100	112/144 (78%)
331	YP_003987287.1 (R757): putative F-box protein	2*e*−165	329/332 (99%)	2*e*−52	73/73 (100%)
376	YP_003987388.1 (R857): hypothetical protein	0.0	374/376 (99%)	9*e*−41	49/49 (100%)
406	YP_003987407.1 (L872): hypothetical protein	0.0	406/407 (99%)	2*e*−131	128/128 (100%)

* Two overlapping reads; nt, nucleotide.

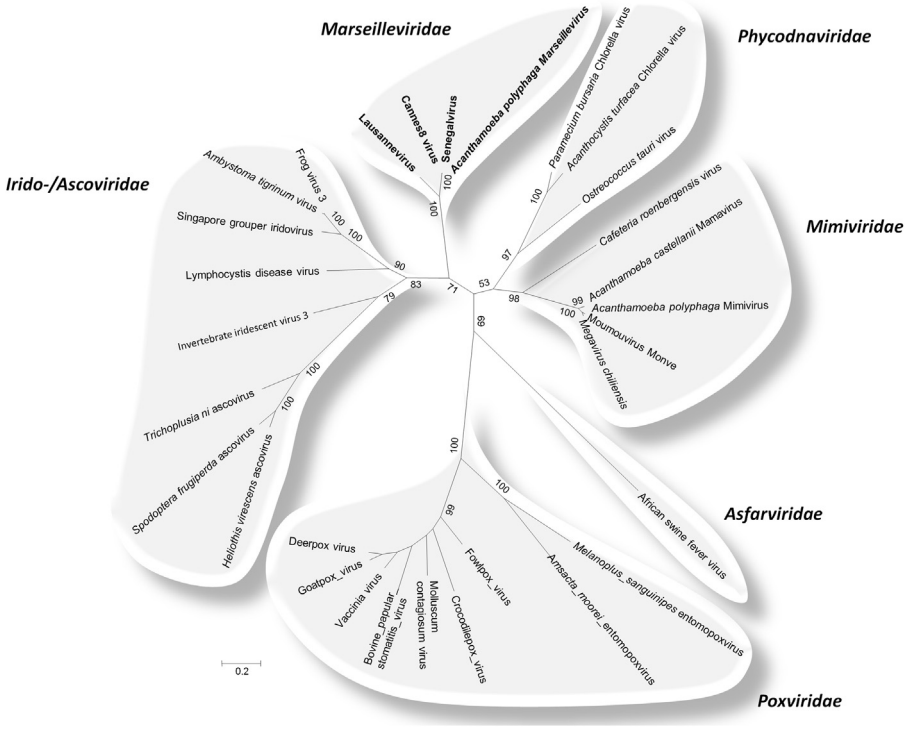


Fig. 3. Phylogeny reconstruction based on an alignment (generated by Muscle (http://www.ebi.ac.uk/Tools/msa/muscle/)) of DNA polymerase of marseilleviruses and other *Megavirales* members, using the Maximum Likelihood method with the Mega 5 software (http://www.megasoftware.net/). Probabilities are shown near branches as a percentage and are used as confidence values of tree branches. Scale bar represents the number of estimated changes per position for a unit of branch length.

mimivirus as the top hit, mimivirus proteins were among the best hits, with *e*-values ranging from 1*e*−6 to 1*e*−14 and amino acid identities ranging from 28 to 58%. In addition, when searching using tBLASTn with the three mimivirus proteins against 55 human metagenomes with the NCBI genomic BLAST tool (http://www.ncbi.nlm.nih.gov/sutils/blast_table.cgi?taxid=Environmental&taxidinf=environ_info&selectall), significant matches, with *e*-values ranging from 8*e*−24 to 3*e*−25 and amino acid identities

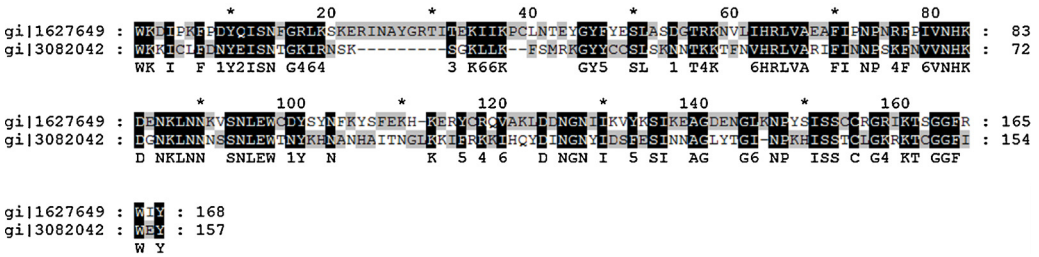


Fig. 4. Amino acid alignments for translated sequences of metagenomic reads (GenBank Accession number BAAZ01000542.1 (gi|162764931)) recovered from a human gut metagenome and the uncharacterized Mimivirus HNH endonuclease (accession no. ADO18080.1 (gi|308204279)). The representation was built using the GeneDoc software (http://www.psc.edu/biomed/genedoc).

Table 4
Number of BLASTn hits obtained for 4 gut metagenomes analyzed through the CAMERA portal.

		CAM_PROJ_Human Distal Gut Human distal gut biome project[69]	CAM_PROJ_Human Gut 13 healthy human gut metagenomes[68]	CAM_PROJ_Human Gut Diagnosis Metagenomic diagnosis of bacterial infections [70]	CAM_PROJ_Twin Study A core gut microbiome in obese and lean twins [71]
Mimiviridae	Group I – lineage A	10	40	0	18
Mimiviridae	Group I – lineage B	32	15	0	105
Mimiviridae	Group I – lineage C	4	20	0	50
Mimiviridae	Group II	0	2	0	3
Mimiviridae	Any	46	77	0	176
Phycodnaviridae		9	29	2	36
Poxviridae		8	1	0	0
Total		63	107	2	212

ranging from 36 to 41%, were found for the uncharacterized HNH endonuclease (Fig. 4) and the putative oxidoreductase against sequences recovered in fecal samples from healthy individuals in Japan.[68]

Additional searches of viral metagenomes from 11 studies (supplementary Table S1) using cleaning and trimming of reads and comparative BLASTn searches against our database of *Megavirales* members and virophages and the NCBI sequence database enabled detection of two reads that displayed significant hits with genomes of mimiviruses (Supplementary Figs. S2–S12). In particular, a 130-nucleotide-long read (no. SRR101483.9794578) recovered from a metagenomic dataset corresponding to pulmonary microbiota from patients with acute exacerbation of idiopathic pulmonary fibrosis[65] (Supplementary Fig. S12) found a fragment of a putative bifunctional dihydrofolate reductase/thymidylate synthase (GenBank Accession no. YP_003970135.1) of Crov as the best match, through BLASTn, BLASTx as well as tBLASTx searches against the NCBI sequence databases. Nucleotides 55–130 of this metagenomic read matched amino acids 214–237 of the 452 amino acid-long putative Crov protein (Fig. 5). In addition, 384 hits, including 299 for mimiviruses, 76 for phycodnaviruses and 9 for poxviruses, were found by BLASTn searches against four human gut bacterial metagenomes through the CAMERA portal (Table 4; Supplementary Table S3). The top hits obtained for the metagenomes were from mimiviruses (Pointe-Rouge2 virus, Moumouvirus and Ochan virus)[31] in three cases and *Paramecium bursaria chlorella* virus 1 in the remaining case (Fig. 6a–d); the corresponding metagenomic reads did not find a member of

Fig. 5. Amino acid alignments for translated sequences of a metagenomic read recovered from human lung[65] and the putative bifunctional dihydrofolate reductase/thymidylate synthase (Cafeteria roenbergensis virus BV-PW (Crov; accession no. YP_003970135.1)). The representation was built using the GeneDoc software (http://www.psc.edu/biomed/genedoc).

the *Megavirales* among the best matches through BLAST searches against the NCBI sequence databases. Finally, BLASTp searches with the proteomes of mimiviruses and marseilleviruses against microbial metagenomes from 11 body sites from the human microbiome project showed from 2 to 54 significant hits per virus (Table 5), including for instance a putative metalloendopeptidase protein (Crov ORF_67) and an asparaginyl-tRNA synthetase (*Megavirus chilensis* ORF_743) in a saliva metagenome, and a 70-kDa heat-shock protein and a DNA Topoisomerase IA of Mimivirus (ORF_393, ORF_221) in a vagina metagenome. Nonetheless, BLASTp searches using the corresponding annotated proteins from the metagenomes against the NCBI GenBank non-redundant protein sequence database did not find mimiviruses or marseilleviruses proteins as best hits.

Overall, *Megavirales*-related sequences were recovered through various strategies from a variety of samples such as saliva (2), oropharynx (1), lung (1) and stools (&).

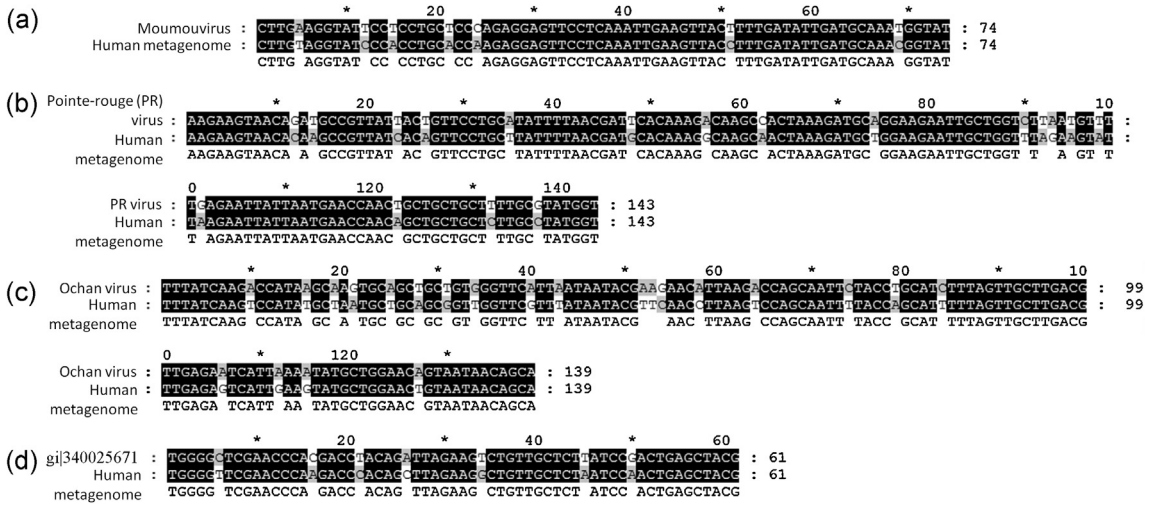


Fig. 6. Nucleotide alignments for metagenomic reads recovered from human gut and the DNA genome of Moumouvirus (5a), Pointe-Rouge2 virus (5b), Ochan virus (5c), and *Paramecium bursaria* Chlorella virus 1 (gi|340025671) (5d). The representation was built using the GeneDoc software (http://www.psc.edu/biomed/genedoc).

Table 5
Significant hits (number) for proteomes of *Megavirales* members against protein-encoding sequences from various human metagenomes of the human microbiome project (http://www.hmpdacc.org/HMGI/).

Megavirales		Human body sites (number of metagenomes)											
		Anterior nares (87)	Throat (7)	Palatine tonsils (6)	Subgingival plague (7)	Attached keratinized gingiva (6)	Saliva (3)	Right retroauricular crease (17)	Left retroauricular crease (9)	Posterior fornix (51)	Vaginal Introitus (3)	Mid vagina (2)
Mimiviridae												
Group I – lineage A	*Acanthamoeba polyphaga mimivirus*	26	32	31	33	25	11	42	34	16	11	11
	Acanthamoeba castellanii mamavirus	27	37	36	36	28	14	49	42	17	12	11
Group I – lineage B	Moumouvirus	32	36	38	40	41	21	54	45	25	14	13
Group I – lineage C	*Megavirus chilensis*	30	30	32	33	29	13	45	40	19	10	10
Group II	*Cafeteria roenbergensis* virus	20	22	24	22	21	11	35	22	13	9	10
Marseilleviridae	Marseillevirus	11	6	5	8	6	5	12	12	5	2	2
	Lausannevirus	8	7	6	8	5	4	12	14	4	4	4

5. Conclusions

In the present study, we showed association of *Megavirales* members with humans using different strategies and samples. Our attention was drawn to the presence of giant viruses in a patient stool sample through the serendipitous detection of Mimivirus- and Marseillevirus-like sequences while using a new metagenomic approach targeting bacterial DNA. Subsequently, Senegalvirus, a new member of the family *Marseilleviridae*, was isolated from this stool sample demonstrating the possibility of the presence of giant viruses in humans. In addition, we detected sequences matching DNA of *Megavirales* members in several human metagenomes, which adds to previous findings in human nasopharyngeal, fecal and blood samples by other laboratories (Table 2).[24,41–47]

Members of the order *Megavirales* represent a technical problem in the current investigation of the virome due to their size. Indeed, viruses are still usually considered small agents,[1,12] which leads the vast majority of research groups to perform viral purification by filtering samples through small-pore (0.2–0.45 μm) filters prior to metagenomic analysis, thus preventing the detection of viruses larger than the filter pores.[72–74] This biased technical approach has most likely led to considerable underestimation of the prevalence of *Megavirales* members in environmental and human samples. In addition, the present work underscores that the detection of giant viruses in humans may benefit from the concurrent use of culture and metagenomics. Accordingly, dramatic differences between the set of bacteria isolated by means of a large panel of culture conditions, the so-called culturomics approach, and the set of bacteria identified through metagenomics were recently unveiled.[51] The Senegalvirus discovery highlights that a virus may be cultured but not molecularly detected. The isolation of Senegalvirus represented the first isolation of a marseillevirus from a human sample.[51] In another report, we also described the isolation of a, Lentillevirus, from contact lens liquid.[29,30] More importantly, we have recently isolated a mimivirus, LBA111, by amoebal culture from a bronchoalveolar sample collected in a Tunisian woman presenting with pneumonia.[75] In addition, sequences related to Marseillevirus DNA and the genome of a new giant virus, named Giant Blood Marseille-like virus, were recovered from the blood of asymptomatic blood donors using high-throughput sequencing.[76]. Previous studies have also shown the presence of poxvirus- and asfarvirus-related sequences in human blood from apparently healthy persons (Table 2),[41,43] raising questions about the asymptomatic carriage of giant viruses and their role over the short and long term.

Taken together, present and previous data provide convergent evidence for the presence of mimiviruses and marseilleviruses in humans, which raises further questions about their potential pathogenicity. We recommend discarding technical procedures that are too stringent and may lead to the neglect of the study of the 'megavirome' while investigating the human virome.

Funding

Researches of LF and CD are financed through a starting grant number 242729 from the European Research Council.

Competing interests

None for all authors.

Ethical approval

None required.

Acknowledgements

None.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jcv.2013.03.018.

References

1. Raoult D, La Scola B, Birtles R. The discovery and characterization of Mimivirus, the largest known virus and putative pneumonia agent. *Clin Infect Dis* 2007;**45**(July (1)):95–102.
2. Raoult D. Giant viruses from amoeba in a post-Darwinist viral world. *Intervirology* 2010;**53**(5):251–3.
3. La Scola B, Audic S, Robert C, Jungang L, de L, Drancourt XM, et al. A giant virus in amoebae. *Science* 2003;**299**(March (5615)):2033.
4. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, et al. The 1. 2-megabase genome sequence of Mimivirus. *Science* 2004;**306**(November (5700)):1344–50.
5. Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, et al. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci USA* 2009;**106**(December (51)):21848–53.
6. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, Fournous G, et al. The virophage as a unique parasite of the giant mimivirus. *Nature* 2008;**455**(September (7209)):100–4.
7. Fischer MG, Allen MJ, Wilson WH, Suttle CA. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci USA* 2010;**107**(November (45)):19508–13.
8. Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A, et al. Lausannevirus, a giant amoebal virus encoding histone doublets. *Environ Microbiol* 2011;**13**(June (6)):1454–66.
9. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM. Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proc Natl Acad Sci USA* 2011;**108**(October (42)):17486–91.
10. Iyer LM, Balaji S, Koonin EV, Aravind L. Evolutionary genomics of nucleocytoplasmic large DNA viruses. *Virus Res* 2006;**117**(April (1)):156–84.
11. Iyer LM, Aravind L, Koonin EV. Common origin of four diverse families of large eukaryotic DNA viruses. *J Virol* 2001;**75**(December (23)):11720–34.
12. Colson P, de L, Fournous X, Raoult GD. Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. *Intervirology* 2012;**55**(April (5)):321–32.
13. La Scola B, Marrie TJ, Auffray JP, Raoult D. Mimivirus in pneumonia patients. *Emerg Infect Dis* 2005;**11**(March (3)):449–52.
14. Greub G, Raoult D. Microorganisms resistant to free-living amoebae. *Clin Microbiol Rev* 2004;**17**(April (2)):413–33.
15. Khan M, La Scola B, Lepidi H, Raoult D. Pneumonia in mice inoculated experimentally with Acanthamoeba polyphaga mimivirus. *Microb Pathog* 2007;**42**(February (2/3)):56–61.
16. Ghigo E, Kartenbeck J, Lien P, Pelkmans L, Capo C, Mege JL, et al. Ameobal pathogen mimivirus infects macrophages through phagocytosis. *PLoS Pathog* 2008;**4**(June (6)):e1000087.
17. Berger P, Papazian L, Drancourt M, La Scola B, Auffray JP, Raoult D. Ameba-associated microorganisms and diagnosis of nosocomial pneumonia. *Emerg Infect Dis* 2006;**12**(February (2)):248–55.
18. Raoult D, Renesto P, Brouqui P. Laboratory infection of a technician by mimivirus. *Ann Intern Med* 2006;**144**(May (9)):702–3.
19. Vincent A, La Scola B, Forel JM, Pauly V, Raoult D, Papazian L. Clinical significance of a positive serology for mimivirus in patients presenting a suspicion of ventilator-associated pneumonia. *Crit Care Med* 2009;**37**(January (1)):111–8.
20. Larcher C, Jeller V, Fischer H, Huemer HP. Prevalence of respiratory viruses, including newly identified viruses, in hospitalised children in Austria. *Eur J Clin Microbiol Infect Dis* 2006;**25**(November (11)):681–6.
21. Dare RK, Chittaganpitch M, Erdman DD. Screening pneumonia patients for mimivirus. *Emerg Infect Dis* 2008;**14**(March (3)):465–7.
22. Costa C, Bergallo M, Astegiano S, Terlizzi ME, Sidoti F, Solidoro P, et al. Detection of Mimivirus in bronchoalveolar lavage of ventilated and nonventilated patients. *Intervirology* 2011;**55**(November (4)):303–5.
23. Arden KE, McErlean P, Nissen MD, Sloots TP, Mackay IM. Frequent detection of human rhinoviruses, paramyxoviruses, coronaviruses, and bocavirus during acute respiratory tract infections. *J Med Virol* 2006;**78**(September (9)):1232–40.
24. Lysholm F, Wetterbom A, Lindau C, Darban H, Bjerkner A, Fahlander K, et al. Characterization of the viral microbiome in patients with severe lower respiratory tract infections, using metagenomic sequencing. *PLoS ONE* 2012;**7**(2):e30875.
25. Vanspauwen MJ, Franssen FM, Raoult D, Wouters EF, Bruggeman CA, Linssen CF. Infections with mimivirus in patients with chronic obstructive pulmonary disease. *Respir Med* 2012;**12**(September):10.
26. Colson P, Raoult D. In: Scheld WM, Grayson ML, Hughes JM, editors. *Is Acanthamoeba polyphaga Mimivirus an emerging causative agent of pneumonia?*. Washington, DC: ASM Press; 2010.
27. Vincent A, La Scola B, Papazian L. Advances in Mimivirus pathogenicity. *Intervirology* 2010;**53**(5):304–9.
28. Parola P, Renvoisé A, Botelho-Nevers E, La Scola B, Desnues C, Raoult D. Acanthamoeba polyphaga Mimivirus virophage seroconversion in patients returning from Laos. *Emerg Infect Dis* 2012;**18**(9):1500–2.
29. Cohen G, Hoffart L, La Scola B, Raoult D, Drancourt M. Ameba-associated Keratitis, France. *Emerg Infect Dis* 2011;**17**(July (7)):1306–8.
30. Desnues C, La Scola B, Yutin N, Fournous G, Robert C, Azza S, et al. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci USA* 2012;**109**(44):18078–83.
31. La Scola B, Campocasso A, N'Dong R, Fournous G, Barrassi L, Flaudrops C, et al. Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. *Intervirology* 2010;**53**(5):344–53.
32. Gaze WH, Morgan G, Zhang L, Wellington EM. Mimivirus-like particles in Acanthamoebae from sewage sludge. *Emerg Infect Dis* 2011;**17**(June (6)):1127–9.
33. Ghedin E, Claverie JM. Mimivirus relatives in the Sargasso sea. *Virol J* 2005;**2**:62.
34. Monier A, Larsen JB, Sandaa RA, Bratbak G, Claverie JM, Ogata H. Marine mimivirus relatives are probably large algal viruses. *Virol J* 2008;**5**:12.
35. Claverie JM, Grzela R, Lartigue A, Bernadac A, Nitsche S, Vacelet J, et al. Mimivirus and Mimiviridae: giant viruses with an increasing number of potential hosts, including corals and sponges. *J Invertebr Pathol* 2009;**101**(July (3)):172–80.
36. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* 2010;**18**(January (1)):11–9.
37. Correa AM, Welsh RM, Vega Thurber RL. Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. *ISME J* 2012;**7**(July (1)):13–27.
38. Monier A, Claverie JM, Ogata H. Taxonomic distribution of large DNA viruses in the sea. *Genome Biol* 2008;**9**(7):R106.
39. Williamson SJ, Allen LZ, Lorenzi HA, Fadrosh DW, Brami D, Thiagarajan M, et al. Metagenomic exploration of viruses throughout the Indian Ocean. *PLoS ONE* 2012;**7**(10):e42047.
40. Desnues C, Raoult D. Virophages question the existence of satellites. *Nat Rev Microbiol* 2012;**10**(February (3)):234–43.
41. Breitbart M, Rohwer F. Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing. *Biotechniques* 2005;**39**(November (5)):729–36.
42. Finkbeiner SR, Allred AF, Tarr PI, Klein EJ, Kirkwood CD, Wang D. Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS Pathog* 2008;**4**(February (2)):e1000011.
43. Loh J, Zhao G, Presti RM, Holtz LR, Finkbeiner SR, Droit L, et al. Detection of novel sequences related to african Swine Fever virus in human serum and sewage. *J Virol* 2009;**83**(December (24)):13019–25.
44. Sullivan PF, Allander T, Lysholm F, Goh S, Persson B, Jacks A, et al. An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. *BMC Microbiol* 2011;**11**(January (2)):2.
45. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis* 2012;**6**(2):e1485.
46. Wylie KM, Mihindukulasuriya KA, Sodergren E, Weinstock GM, Storch GA. Sequence analysis of the human virome in febrile and afebrile children. *PLoS ONE* 2012;**7**(6):e27735.
47. Zhou J, Zhang W, Yan S, Xiao J, Zhang Y, Li B, et al. Diversity of virophages in metagenomic datasets. *J Virol* 2013;**87**(April (8)):4225–36.
48. Sachsenroder J, Twardziok S, Hammerl JA, Janczyk P, Wrede P, Hertwig S, et al. Simultaneous identification of DNA and RNA viruses present in pig faeces using process-controlled deep sequencing. *PLoS ONE* 2012;**7**(4):e34631.
49. Ng TF, Willner DL, Lim YW, Schmieder R, Chau B, Nilsson C, et al. Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. *PLoS ONE* 2011;**6**(6):e20579.
50. Sparks ME, Gundersen-Rindal DE. The Lymantria dispar IPLB-Ld652Y cell line transcriptome comprises diverse virus-associated transcripts. *Viruses* 2011;**3**(November (11)):2339–50.
51. Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, et al. Microbial culturomics: paradigm shift in the human gut microbiome study. *Clin Microbiol Infect* 2012;**18**(December (12)):1185–93.
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;**215**(October (3)):403–10.
53. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;**7**(May (5)):335–6.
54. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, Felts B, et al. Viral diversity and dynamics in an infant gut. *Res Microbiol* 2008;**159**(June (5)):367–73.
55. Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, et al. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 2003;**185**(October (20)):6220–3.
56. Victoria JG, Kapoor A, Li L, Blinkova O, Slikas B, Wang C, et al. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol* 2009;**83**(May (9)):4642–51.
57. Kim MS, Park EJ, Roh SW, Bae JW. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl Environ Microbiol* 2011;**77**(November (22)):8062–70.
58. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, et al. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* 2011;**21**(October (10)):1616–25.

59. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SW, et al. RNA viral community in human feces: prevalence of plant pathogenic viruses. *PLoS Biol* 2006;**4**(January (1)):e3.
60. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* 2009;**4**(1):e4219.
61. Pride DT, Salzman J, Haynes M, Rohwer F, Davis-Long C, White III RA, et al. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *ISME J* 2012;**6**(May (5)):915–26.
62. Willner D, Furlan M, Schmieder R, Grasis JA, Pride DT, Relman DA, et al. Metagenomic detection of phage-encoded platelet-binding factors in the human oral cavity. *Proc Natl Acad Sci USA* 2011;**108**(March (Suppl. 1)):4547–53 [Epub, 2010 June].
63. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, et al. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS ONE* 2009;**4**(10):e7370.
64. Willner D, Haynes MR, Furlan M, Schmieder R, Lim YW, Rainey PB, et al. Spatial distribution of microbial communities in the cystic fibrosis lung. *ISME J* 2012;**6**(February (2)):471–4.
65. Wootton SC, Kim DS, Kondoh Y, Chen E, Lee JS, Song JW, et al. Viral infection in acute exacerbation of idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med* 2011;**183**(June(12)):1698–702.
66. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011;**27**(March (6)):863–4.
67. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 2005;**33**(July (Web Server issue)):W451–4.
68. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. *DNA Res* 2007;**14**(August (4)):169–81.
69. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic analysis of the human distal gut microbiome. *Science* 2006;**312**(June (5778)):1355–9.
70. Nakamura S, Maeda N, Miron IM, Yoh M, Izutsu K, Kataoka C, et al. Metagenomic diagnosis of bacterial infections. *Emerg Infect Dis* 2008;**14**(November (11)):1784–6.
71. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. *Nature* 2009;**457**(January (7228)):480–4.
72. Angly FE, Willner D, Prieto-Davo A, Edwards RA, Schmieder R, Vega-Thurber R, et al. The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* 2009;**5**(December (12)):e1000593.
73. Edwards RA, Rohwer F. Viral metagenomics. *Nat Rev Microbiol* 2005;**3**(June (6)):504–10.
74. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 2009;**4**(4): 470–83.
75. Saadi H, Pagnier I, Colson P, Kanoun Cherif J, Beji M, Boughalmi M, et al. First isolation of Mimivirus in a patient with pneumonia. *Clin Infect Dis* 2013; in press.
76. Popgeorgiev N, Boyer M, Fancello L, Monteil S, Robert C, Rivet R, et al. Giant Blood Marseillevirus recovered from asymptomatic blood donors. *J Infect Dis* 2013; in press.

Conclusions and Perspectives

CONCLUSIONS AND PERSPECTIVES

During the past decade, two new viral families have emerged from the study of environmental samples, mostly water, using amoebal co-culture. Thus, the families *Mimiviridae* and *Marseilleviridae* have been established and recognized by the international committee on taxonomy of virus (ICTV). These families now encompass dozens of giant viruses, the remarkable features of which have largely moved the boundaries of the known virosphere. Particularly, mimiviruses, with genome sizes and particle sizes in the same order of magnitude than those of small bacteria and genes encoding protein from the translation apparatus, challenged the definition of viruses and fostered interest for giant DNA viruses (Colson et al. 2012; Yutin & Koonin, 2012; Yutin et al. 2013a). Comparative genomics and phylogenetic analyses of the mimivirus genomes and gene content delineated two groups. One group is composed of mimiviruses that infect *Acanthamoeba* spp., and three lineages A, B and C have been delineated that have Mimivirus, Moumouvirus and *Megavirus chiliensis* as leading members, respectively (Raoult et al. 2004; Yoosuf et al. 2012; Arslan et al. 2011). A second group is a sister group that encompasses mimiviruses of marine cellular organisms, including *Cafeteria roenbergensis* virus that infects a widespread marine dinoflagellate, and *Phaeocystis globosa* viruses and Organic lake phycodnaviruses that were recently shown being genuine mimiviruses (Fischer et al. 2010; Colson et al. 2011b; Santini et al. 2013; Yutin et al. 2013b).

The gene content of Moumouvirus expands the pan-genome of the family *Mimiviridae* and emphasizes the dynamic evolution of this mimivirus, especially extensive gene loss. The comparative analysis of the genomes of Moumouvirus with the previously described genomes of amoeba-associated mimiviruses of lineage A and C showed that Moumouvirus is much closer from *Megavirus chilensis*. The close evolutionary relationship between Moumouvirus, isolated from freshwater and *Megavirus chilensis*, isolated from the marine environment shows the ecological plasticity of these giant viruses that can survive in different habitats (Arslan et al. 2011; Yoosuf et al. 2012). The role of widespread phagocytic protists in this fact is likely considerable. Besides, we annotated the genomes of two mimviruses, Terra1 virus and Terra2 virus, which were isolated from soil, and not water as for previously described mimiviruses. The comparative analysis of Terra1 virus and Terra2 virus genomes indicated that they are new *bona fide* members of the family *Mimiviridae*, belonging to lineages C and A of mimiviruses of amoeba, respectively.

The architecture of the mimivirus genomes shows conserved collinear central regions and far less conserved extremities. This pattern seems to be a general trend among the members of the proposed order "*Megavirales*", as it was noted earlier in poxviruses and phycodnaviruses (Senkevich et al. 1997; McLysaght et al. 2003; Filée et al. 2007). In genomic comparisons, we also identified large inverted regions in the middle part of the genomes and shorter collinear regions at the extremities (Arslan et al. 2011; Yoosuf et al. 2012). These

findings suggest that the reshaping of the genomes may occur through the rearrangement of large fragments.

Acanthamoeba spp. are phagocytic protists classified in the phylum *Amoebozoa* and are predominant among the organisms in soil and water. These protists are the known hosts of mimiviruses (Barker & Brown, 1994; Moliner et al. 2010; Thomas et al. 2011). They are free living amoebae that can ingest any particle of a size greater than 0.5 μm, and graze many microorganisms including bacteria, yeasts, fungi, viruses and algae (Barker & Brown, 1994; Horn & Wagner, 2004; Rodriguez et al. 1994). Therefore, *Acanthamoeba* spp. engulf large amounts of foreign DNA. The mimiviruses living inside amoebae have a sympatric lifestyle with other viruses and microorganisms (bacteria have been mostly detected) and their eukaryotic host and the analyses of the mimiviruses, and marseilleviruses, gene repertoires have highlighted that amoeba might act as a melting pot for the microorganisms that can survive to its ingestion, which provides opportunities for gene exchanges between these different organisms (Colson & Raoult, 2010; Moliner et al. 2010; Raoult & Boyer, 2010; Thomas & Greub, 2010). Such opportunity for amoeba resisting organisms to gain genes was also significantly associated with a greater genome size for microorganisms and giant viruses that have a sympatric intra-amoebal lifestyle in comparison with those from the same phyla that have an allopatric lifestyle (Filée et al. 2007; Moliner et al. 2010; Raoult & Boyer, 2010).

In the studies conducted during my Thesis, we identified in the genome of Terra1 virus a cluster of genes, all adjacent to each other, which are orthologous to bacterial genes and have no counterpart in other viral genomes. This finding led to detect that a cluster of genes homologous to those found in the genome of Terra1 virus was present in the genomes of lineage C of mimiviruses of amoeba but absent in the two other lineages. This observation supports the hypothesis of an evolutionary scenario of gene gain by the ancestor of mimivirus of lineage C or, alternatively, a less parsimonious evolutionary scenario of gene gain by a common mimivirus A-C ancestor then gene loss in mimivirus lineages A and B. Such clusters of genes of bacterial origin were earlier noticed in Mimivirus and *Cafeteria roenbergensis* virus genomes (Filée et al. 2007; Fischer et al. 2010). The cluster of bacterial genes identified in the Terra1 virus genome encodes proteins involved in carbohydrate metabolism. Strikingly, a 38-kilobase-pair genomic fragment was identified in *Cafeteria roenbergensis* virus that encodes 34 predicted genes, among which 7 were predicted to be involved in carbohydrate metabolism among the 14 that were most similar to bacterial genes (Fischer et al. 2010).

Since the discovery, a decade ago, of the first mimivirus, the pan-genome of the family *Mimiviridae* has exhibited a 2.5-fold expansion. Our recent studies of new mimiviruses have detected far lower number of ORFans in their genome, which suggests that the pan-genome of mimiviruses might reached a plateau and could be considered as closed pan-genome, based on currently available genomes. Interesting

issues that might need to be addressed more extensively are the studies of the considerable amount of family ORFans and hypothetical proteins present in the mimivirus genomes. These predicted proteins potentially represent keys to an increased knowledge of the origin, evolution and replication of the mimiviruses. Transcriptomic and proteomic analyses that are increasingly conducted in our research unit will likely provide several information regarding these groups of genes.

Mimiviruses considerably challenge the classical definition of viruses. Indeed, from the birth of the virus concept, viruses were considered as small (and mainly non-living) entities with a very limited gene armamentarium and fully dependent of the host cell for translation and energy production. In addition, mimiviruses considerably expanded the diversity of the viral world. Importantly, the presence in mimiviruses of several proteins involved in the biosynthesis of nucleotides, transcription and translation enabled to show that the members of the *Megavirales* may compose a fourth branch of life (Boyer et al. 2010; Colson et al. 2011b). Thus, these mimiviruses enabled to point out that giant viruses were involved in early step of evolution and have an ancient origin dating back to the proto-eukaryotes. The discovery of virus of mimiviruses, so-named virophages based on their functional analogy with bacteriophages, also contributed to improve our knowledge of the viral word (Desnues et al. 2012; La Scola et al. 2008). Recently, still larger viruses (with particle and genome size as high as 1 μm and 2,5 kilobase pairs, respectively), named pandoraviruses, have been isolated from environmental samples, which infect *Acanthamoeba* spp.

These biological entities were classified as viruses based on their gene repertoire, which linked them to *Megavirales*, and the existence of an eclipse phase during their replicative cycle. Nonetheless, only 7% of the genes of these viruses have homologs in sequence databases, no gene was identified that encodes a capsid. The pandoravirus discovery is another hint that lots remain to know about biological entities and particularly viruses (Philippe et al. 2013).

Taken together, previous findings indicate that new strategies should be implemented to stalk more viruses including divergent ones from our biosphere. They may include the use of other protists than *Acanthamoeba* spp., and collection from other sources. Such approaches are currently developed at URMITE. One of the major topic regarding the further study of mimiviruses will be linked with its prevalence and potential pathogenicity in humans. There has been accumulating evidence supporting the role of mimivirus as causative agents of pneumonia (La Scola et al. 2005; Raoult et al. 2007). Recently, LBA111 was isolated at URMITE by amoebal culture from a bronchoalveolar sample collected in a Tunisian woman presenting pneumonia, and another amoeba-associated virus of lineage C, Shan, was also isolated from the feces of another Tunisian woman with pneumonia (Saadi et al. 2013a; Saadi et al. 2013b). Pan-genome analysis and comparative genomics of mimiviruses from same and different lineages may be helpful to detect virulence-associated genotypic patterns.

Annex I

"Marseilleviridae", a new family of giant viruses infecting amoebae.

Colson P, Pagnier I, Yoosuf N, Fournous G, La Scola B, Raoult D

Annex I

"Marseilleviridae", a new family of giant viruses infecting amoebae

Five years after Mimivirus, *Acanthamoeba polyphaga* marseillevirus was isolated from water collected from a cooling tower in Paris, France in 2007. Marseillevirus has an icosahedral shape with a diameter of about 250 nm. Its genome is a double-stranded circular DNA that is 368,454 base pairs in length and is predicted to encode 457 proteins. Phylogenetic reconstructions indicated clearly that it belongs to a new viral family among Nucleocytoplasmic large DNA viruses (Boyer et al. 2009). In 2011, *Acanthamoeba castellanii* lausannevirus, a Marseillevirus close relative, was isolated from river water in France (Thomas et al. 2011). After the discovery of Marseillevirus, other giant viruses were isolated from freshwater using the amoebal co-culture method. Among these new viruses, we identified Cannes8 virus which was revealed being a bona fide member of the family *Marseilleviridae* based on the phylogenetic analysis of B-family DNA polymerase gene (Aherfi et al. 2013). In addition, a close relative of *Acanthamoeba polyphaga* marseillevirus, named Senegalvirus, was recovered and isolated in our laboratory from the stool sample of a young Senegalese man (Colson et al. 2013). Marseillevirus and its close relatives exhibit remarkable features, a majority being shared with mimiviruses, and

have contributed to a considerable increase of the interest in NCLDVs. Genome mosaicism was particularly highlighted for Marseillevirus, and was linked to the sympatric lifestyle of these viruses with other micro-organisms inside *Acanthamoeba* spp., which provides opportunities to exchange genes with these microorganisms and the amoebal host (Colson & Raoult, 2010; Moliner et al. 2010; Raoult & Boyer, 2010). In addition, Marseilleviruses have been influential in proposing to reclassify these large and giant DNA viruses in a new viral order named the ‘‘Megavirales’. We concurrently proposed the family ‘‘*Marseilleviridae*’’ as a new viral family to the international committee on the taxonomy of viruses (Colson et al. 2012). The only currently identified hosts for ‘‘marseilleviruses’’ are Acanthamoeba spp. The prevalence of these viruses in the environment and in animals and humans largely remains to be determined.

Arch Virol (2013) 158:915–920
DOI 10.1007/s00705-012-1537-y

VIROLOGY DIVISION NEWS

"Marseilleviridae", a new family of giant viruses infecting amoebae

Philippe Colson · Isabelle Pagnier · Niyaz Yoosuf · Ghislain Fournous · Bernard La Scola · Didier Raoult

Received: 3 August 2012 / Accepted: 3 October 2012 / Published online: 29 November 2012

Abstract The family "Marseilleviridae" is a new proposed taxon for giant viruses that infect amoebae. Its first member, *Acanthamoeba polyphaga* marseillevirus (APMaV), was isolated in 2007 by culturing on amoebae a water sample collected from a cooling tower in Paris, France. APMaV has an icosahedral shape with a diameter of ≈250 nm. Its genome is a double-stranded circular DNA that is 368,454 base pairs (bp) in length. The genome has a GC content of 44.7 % and is predicted to encode 457 proteins. Phylogenetic reconstructions showed that APMaV belongs to a new viral family among nucleocytoplasmic large DNA viruses, a group of viruses that also includes *Acanthamoeba polyphaga* mimivirus (APMV) and the other members of the family *Mimiviridae* as well as the members of the families *Poxviridae*, *Phycodnaviridae*, *Iridoviridae*, *Ascoviridae*, and *Asfarviridae*. In 2011, *Acanthamoeba castellanii* lausannevirus (ACLaV), another close relative of APMaV, was isolated from river water in France. The ACLaV genome is 346,754 bp in size and encodes 450 genes, among which 320 have an APMaV protein as the closest homolog. Two other giant viruses closely related to APMaV and ACLaV have been recovered in our laboratory from a freshwater sample and a human stool sample using an amoebal co-culture method. The only currently identified hosts for "marseilleviruses" are *Acanthamoeba* spp. The prevalence of these viruses in the environment and in animals and humans remains to be determined.

P. Colson · I. Pagnier · N. Yoosuf · G. Fournous · B. La Scola · D. Raoult
URMITE UM63 CNRS 7278 IRD 198 INSERM U1905, Aix-Marseille Université, Facultés de Médecine et de Pharmacie, 27 boulevard Jean Moulin, 13385 Marseille Cedex 05, France

P. Colson · B. La Scola · D. Raoult (✉)
Pôle des Maladies Infectieuses et Tropicales Clinique et Biologique, Fédération de Bactériologie-Hygiène-Virologie, IHU Méditerranée Infection, Centre Hospitalo-Universitaire Timone, Assistance Publique - Hôpitaux de Marseille, 264 rue Saint-Pierre, 13385 Marseille Cedex 05, France
e-mail: didier.raoult@gmail.com

Introduction

Acanthamoeba polyphaga marseillevirus (APMaV) was isolated in 2007 from water collected from a cooling tower in Paris, France, using a method based on *Acanthamoeba polyphaga* culture [1]. The name of this virus originates from the name of its amoebal host and the name of the French city, Marseille, where it was discovered. APMaV was described five years after the discovery of *Acanthamoeba polyphaga* mimivirus (APMV), the first giant virus identified using an amoebal co-culture method. APMV was revealed to be the largest known virus [2, 3]. APMaV was found to be smaller than APMV with respect to the sizes of the capsid and the genome. Nonetheless, with a capsid diameter of approximately 250 nm (Fig. 1) and a genome composed of 368,454 base pairs (bp) encoding 457 genes, APMaV represents a new giant virus. After the discovery of APMaV, other giant viruses were isolated from freshwater using the amoebal co-culture method and were briefly described in 2010 [4]. Among these new viruses, Cannes 8 virus (Ca8V) is a close relative of APMaV based on the phylogeny of the B-family DNA polymerase gene [4]. In 2011, another large DNA virus, *Acanthamoeba castellanii* lausannevirus (ACLaV), was described, and this virus was determined to be a close relative of APMaV. ACLaV was isolated by culturing freshwater collected in 2005 from the Seine River in France on amoebae [5]. An additional close relative of APMaV was recently recovered in our laboratory from the stool of a young Senegalese

man, and this new virus was named Senegal virus (SNGV) (Fig. 2) [6]. The genomes of Ca8V and SNGV have been sequenced on a 454-Roche GS20 instrument (Roche, USA) as described previously [1]. Additionally, the genome of the Ca8V isolated in our laboratory was sequenced on a SOLiD instrument (Life Technologies Corporation).

Genomics of "marseilleviruses"

The APMaV genome is a circular double-stranded DNA molecule of 368,454 base pairs (Table 1) [1]. Its GC content is 44.7 %. The APMaV genome harbors 457 open reading frames (ORFs) predicted to encode proteins with a size ranging from 50 to 1,537 amino acids. These ORFs represent 89 % of the genome. APMaV was identified as representing a unique nucleocytoplasmic large DNA virus (NCLDV) family [1, 7]. NCLDVs were described in 2001 as a monophyletic group of large viruses with a DNA genome. This group of viruses comprises the families *Poxviridae*, *Asfarviridae*, *Iridoviridae* and *Phycodnaviridae*, which were grouped together based on a set of core genes shared by all of the member viruses [8]. Later, APMV and then APMaV were found to be related to this group of viruses [1, 3], for which we recently proposed reclassification in a new viral order, "Megavirales" (talk.ictvonline.org/files/proposals/taxonomy_proposals_fungal1/m/fung01/4261.aspx) [9]. All of these giant viruses share a common and very early ancestor based on phylogenetic and phyletic analysis of conserved and informational genes [7, 10, 11]. Among the NCLDVs, APMaV branched deeply with irido-/ascoviruses on the basis of the phylogenetic reconstruction of conserved

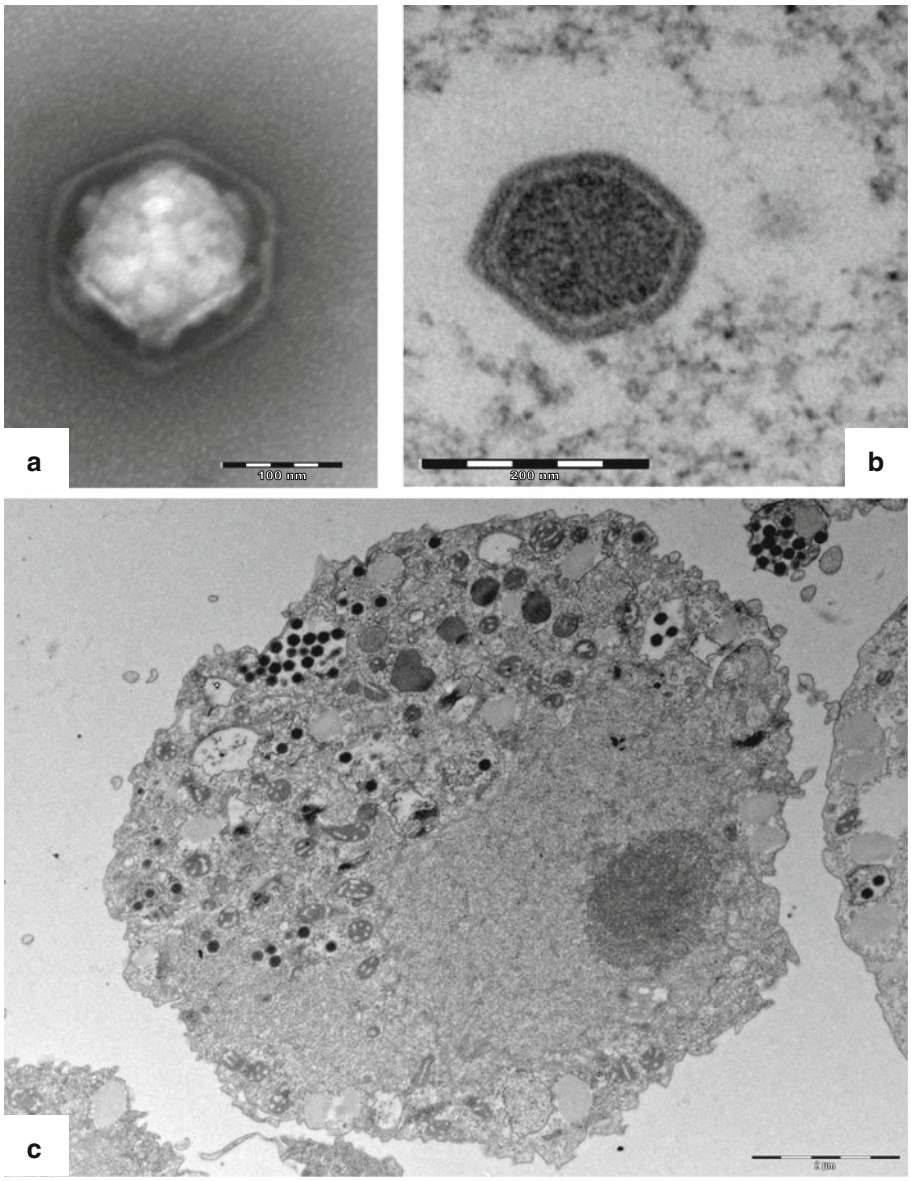


Fig. 1 Electron microscopy images of APMaV particles in a culture supernatant (*scale bar* represents 100 nm) (**a**) and in *Acanthamoeba* sp (*scale bar* represents 200 nm) (**b**), and of *Acanthamoeba* sp infected with APMaV (*scale bar* represents 2 μm) (**c**)

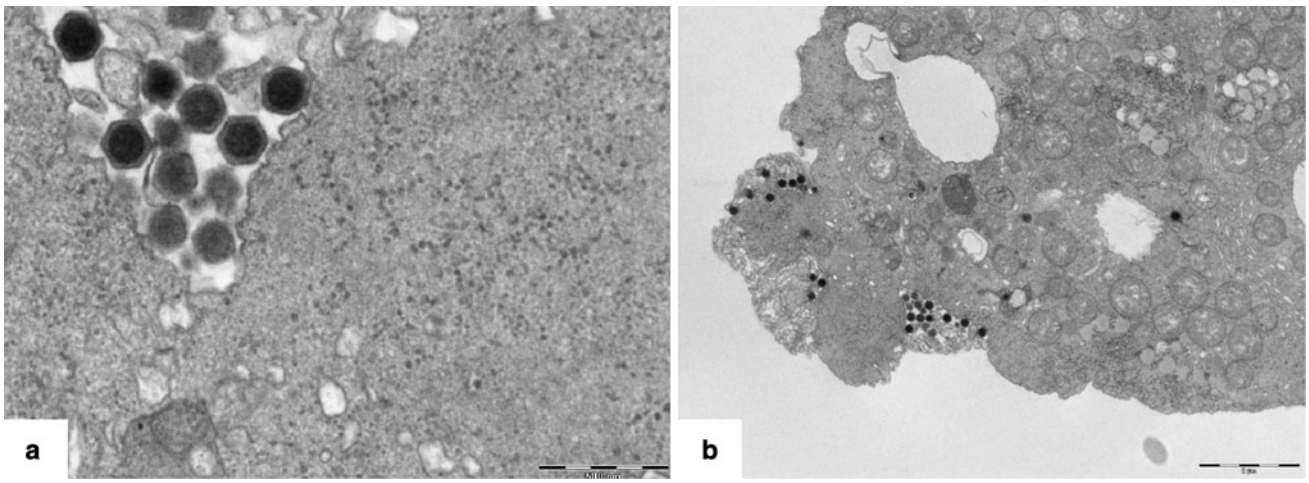


Fig. 2 Electron microscopy images of SNGV in *Acanthamoeba polyphaga*. **a** The *scale bar* represents 500 nm. **b** The *scale bar* represents 2 μm

Table 1 Description of the primary features of the "Marseilleviridae" members

Name	Source	Country/region	Capsid size (nm)	Genome GenBank accession no.	Date of creation	Genome topology	Genome size (bp)	Number of genes	References
APMaV	Cooling tower	France (Paris)	250	NC_013756	25/01/2010	Circular	368,454	457	[1]
ACLaV	River (Seine)	France	190–220 nm	NC_015326	01/04/2011	Linear/circular	346,754	450	[5]
Ca8V	Cooling tower	France (Cannes)	180	JF979175.1[a]	30/06/2012	–	374,039	–	[4]
SNGV	Human stool sample	Senegal	210	JF909596-JF909602	13/09/2011	–	372,690	–	[6]

[a] GenBank accession no. corresponds to the B-family DNA polymerase gene

genes [1, 7]. In contrast, a comparison of the NCLDV gene repertoires instead grouped APMaV with APMV and *Acanthamoeba polyphaga* mamavirus (APMV2). The analysis of the APMaV genome has highlighted its mosaicism and the role of the amoeba as a biological niche for gene acquisition and exchange between sympatric bacteria, viruses and their amoebal hosts [1, 12]. Thus, on the basis of phylogenetic analysis, the APMaV genome contains 51 genes (11 %) of probable NCLDV origin, 49 (11 %) of probable bacterial or bacteriophage origin, and 85 (19 %) of probable eukaryotic origin [1]. A total of 49 proteins have been identified in purified APMaV virions [1]. These proteins have been linked to several functional categories and include NCLDV core proteins, including the capsid protein. Of note, APMaV messenger RNAs, including transcripts encoding the DNA polymerase and the capsid, were found to be encapsidated in the virions.

The ACLaV genome is 346,754 bp in length, and its GC content is 42.9 % [5]. It can be circular molecule or a linear DNA molecule with terminal repeats. This genome harbors 450 ORFs that cover 93 % of the genome and have a mean length of 716 bp. ACLaV encodes homologs for all of the NCLDV core genes detected in APMaV. The phylogenetic analyses published previously showed that APMaV and ACLaV make up a new viral family among the nucleocytoplasmic large DNA viruses (NCLDVs) [1, 5, 7, 10]. This family structure has been well established in several studies using several conserved proteins, including those encoded by NCLDV core genes. Although comparative genomics and phylogenetic reconstructions have shown that ACLaV is a close relative of APMaV and that both viruses belong to the same family [5], the genomes of these two giant viruses display considerable differences [5] (Figs. 3, 4). Indeed, a total of 332 ACLaV proteins (73.8 % of the putative proteome) display significant similarity to proteins in the NCBI non-redundant sequence database, and among those proteins, only 320 (71.1 %) have an APMaV protein as the best BLASTp hit. In addition, comparative analysis of the ACLaV and APMaV genomes revealed a 150-kb region with poor synteny with many hypothetical proteins, followed by a 200-kb region with a higher level of synteny (Figs. 3, 4), [5].

Only two-thirds of the ACLaV and APMaV proteins share a best reciprocal BLAST hit.

Another giant virus, Cannes 8 virus (Ca8V), has been isolated in our laboratory from a freshwater sample using amoebal culture, and this virus has been found to be closely related to APMaV and ACLaV based on phylogenetic reconstructions (Table 1) [4]. Moreover, we obtained the first isolate of a giant virus infecting amoebae from a human sample, a stool sample from a young Senegalese man [6]. The genome (accession numbers JF909596-JF909601) of this giant virus, named Senegal virus (SNGV), has a size of approximately 373 kbp (in the same range as those of APMaV and ACLaV). The analysis of the genomes of SNGV and Ca8V demonstrated that they are bona fide new members of the proposed family "Marseilleviridae" (talk.ictvonline.org/files/proposals/taxonomy_proposals_fungal1/m/fung01/4262.aspx). Nonetheless, the genomes of SNGV and Ca8V display some differences compared with the genomes of APMaV and ACLaV (Fig. 4). The number of bidirectional best hits for APMaV and other members of the family "Marseilleviridae" tentatively ranges from 300 to 399. Overall, the ranges in size and in the number of genes for these new members of the family "Marseilleviridae" are similar to those of APMaV and ACLaV. At the present time, we propose defining only one genus, named "Marseillevirus". The species "Marseillevirus marseillevirus" is assigned to this genus and has one member, APMaV, while the "marseilleviruses" ACLaV, SNGV and Ca8V remain presently unassigned until additional "marseilleviruses" are described.

Members of the family "Marseilleviridae" (the "marseilleviruses"), like those of the family *Mimiviridae* (the mimiviruses) and other NCLDV families, do not meet the usual criteria quoted by Lwoff to define viruses [13], and the outstanding characteristics of these viruses led us recently to propose a new order made up of these giant viruses [9].

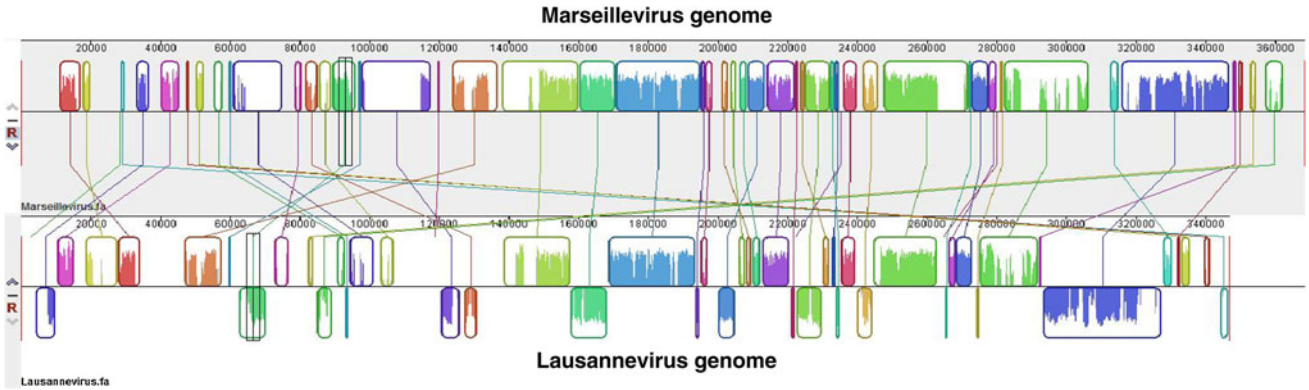


Fig. 3 Comparison and gene alignment of the genomes of APMaV and ACLaV using Mauve software [14]. Colored outlined blocks surround regions of the genome sequence that aligned to part of the other genome. The *colored bars* inside the blocks are related to the level of sequence similarity. Lines link blocks with homology between two genomes. Regions that are inverted relative to the other genome are shifted below a genome's center axis

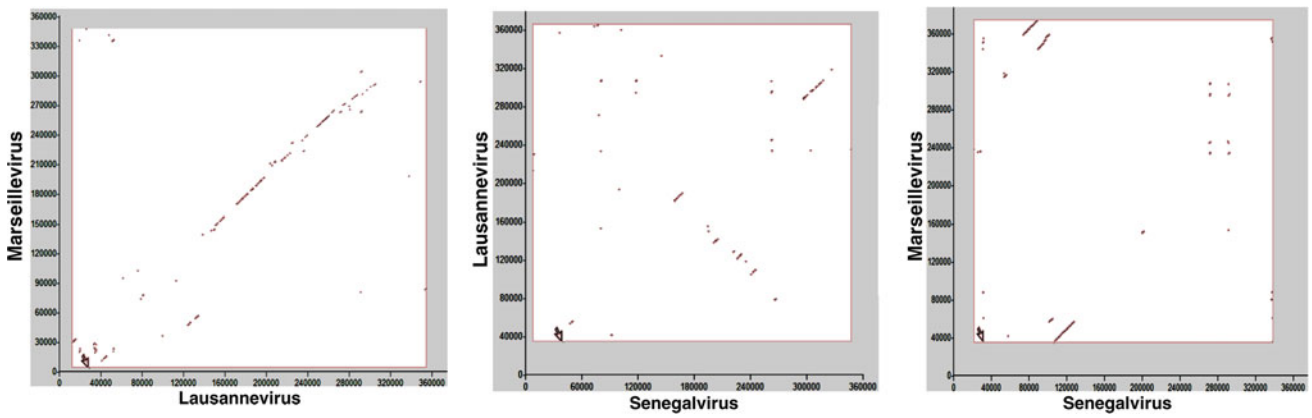


Fig. 4 *Dot plots* for the comparisons of the APMaV, ACLaV, and SNGV genomes using Owen software [15]

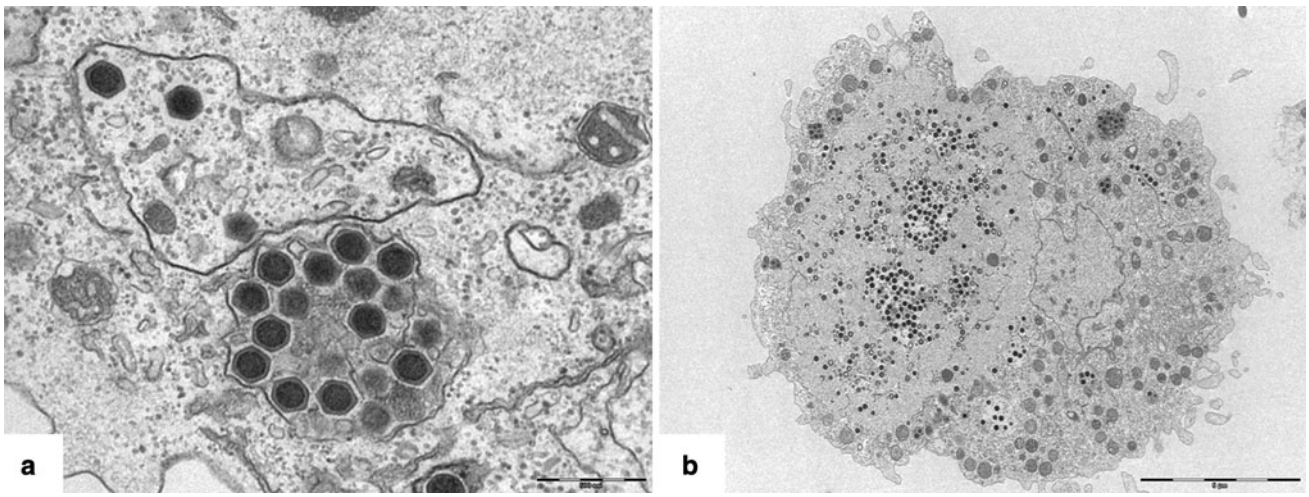


Fig. 5 Electron microscopy images of Ca8V in *Acanthamoeba polyphaga* (*scale bar* represents 500 nm) (**a**) and of *Acanthamoeba polyphaga* infected with Ca8V (*scale bar* represents 5 μm) (**b**)

Morphological properties

APMaV, ACLaV, SNGV and Ca8V share similar morphological features, including the size of their capsids, which ranges from 190 to 250 nm (Fig. 1, 2, 5; Table 1). APMaV has an icosahedral shape and a diameter of ≈250 nm (Table 1; Fig. 1) [1]. The capsid shell has a thickness of ≈10 nm, and 12-nm-long fibers with globular ends are present at the viral surface. A membrane may surround the nucleocapsid, which is separated from the capsid shell by a gap of ≈52 nm. For all members of the family "Marseilleviridae", viral factories can be observed during the replication cycle. These viral factories have different appearances than those observed for APMV and APMV2, tending to be more widely distributed in the amoebal cytoplasm.

Properties in culture

All of the currently identified members of the family "Marseilleviridae" have *Acanthamoeba* spp. as their hosts and were isolated by culturing samples on these amoebae. In amoebal culture, APMaV enters the amoeba 30-60 min post-infection (p.i.) [1]. Later, a viral factory appears close to the nucleus of the amoeba. Capsid assembly and viral genome encapsidation are observed simultaneously in these viral factories, leading to mature and immature APMaV particles. The replication cycle is complete at 5 h p.i, which is a short period of time compared to that observed for APMV. The morphology of the host-cell nucleus changes considerably between 30 min and 2.5 h p.i. Regarding ACLaV, a few viral particles are present 30 min p.i [5]. After an eclipse phase, viruses can be observed again, in large vesicles, at 4 h p.i., and they fill the entire amoeba at 8 h p.i. before amoebal lysis at 16 h p.i.

Prevalence, host, and pathogenicity

The prevalence of "marseilleviruses" in environmental samples is currently unknown. Of note, four of these viruses were recently recovered from 103 water samples [4]. The only currently identified hosts for "marseilleviruses" are *Acanthamoeba* spp. [1, 4, 5]. No data are currently available on the prevalence of "marseilleviruses" in human or animal samples, and no pathogenic role has been demonstrated to date, but one virus belonging to the family "Marseilleviridae", SNGV, has been isolated from a human stool sample [6].

Conclusion

Acanthamoeba polyphaga marseillevirus (APMaV) and its close relatives exhibit remarkable features that are shared by mimiviruses and have contributed to a considerable increase in the interest in NCLDVs and to the better delineation of this group of giant viruses, for which we have recently proposed a new viral order named "Megavirales" [1, 5, 9]. The family "Marseilleviridae" would be included in the order "Megavirales". Further isolates will most likely be described that will be closely related to APMaV. We are performing comparative genomics analysis of the genomes of new putative "marseilleviruses" and will submit these new genomes to sequence databases. We believe that these viruses should be linked to a viral family.

Conflict of interest All of the authors declare that they have no potential conflict of interest.

References

1. Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann MG, Suzan-Monti M, La Scola B, Koonin EV, Raoult D (2009) Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc Natl Acad Sci USA 106:21848–21853
2. La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, Drancourt M, Birtles R, Claverie JM, Raoult D (2003) A giant virus in amoebae. Science 299:2033
3. Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM (2004) The 1.2-megabase genome sequence of Mimivirus. Science 306:1344–1350
4. La Scola B, Campocasso A, N'Dong R, Fournous G, Barrassi L, Flaudrops C, Raoult D (2010) Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. Intervirology 53:344–353
5. Thomas V, Bertelli C, Collyn F, Casson N, Telenti A, Goesmann A, Croxatto A, Greub G (2011) Lausannevirus, a giant amoebal virus encoding histone doublets. Environ Microbiol 13:1454–1466
6. Lagier JC, Armougom F, Million M, Hugon P, Pagnier I, Robert C, Bittar F, Fournous G, Gimenez G, Maraninchi M, Trape JF, Koonin E, Koonin EV, La Scola B, Raoult D (2012) Microbial culturomics: paradigm shift in the human gut microbiome study. Clin Microbiol Infect (in press)
7. Koonin EV, Yutin N (2010) Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. Intervirology 53:284–292
8. Iyer LM, Aravind L, Koonin EV (2001) Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 75:11720–11734
9. Colson P, de Lamballerie X, Fournous G, Raoult D (2012) Reclassification of giant viruses composing a fourth domain of life in the new order Megavirales. Intervirology 55(5):321–332
10. Yutin N, Wolf YI, Raoult D, Koonin EV (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6:223
11. Boyer M, Madoui MA, Gimenez G, La Scola B, Raoult D (2010) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. PLoS One 5:e15530
12. Raoult D, Boyer M (2010) Amoebae as genitors and reservoirs of giant viruses. Intervirology 53:321–329
13. Lwoff A (1957) The concept of virus. J Gen Microbiol 17:239–253
14. Darling AC, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394–1403
15. Ogurtsov AY, Roytberg MA, Shabalina SA, Kondrashov AS (2002) OWEN: aligning long collinear regions of genomes. Bioinformatics 18:1703–1704

REFERENCES

Aherfi, S., Pagnier, I., Fournous, G., Raoult, D., La Scola, B., Colson, P. (2013). Complete genome sequence of Cannes 8 virus, a new member of the proposed family "Marseilleviridae." *Virus genes*, 47, 550-555.

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., Claverie, J. M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences*, *108*, 1–6.

Barker, J., Brown, M. (1994). Trojan horses of the microbial world: protozoa and the survival of bacterial pathogens in the environment. *Mircobiology*, *140*(6), 1253–1259.

Boughalmi, M., Saadi, H., Pagnier, I., Colson, P., Fournous, G., Raoult, D., La Scola, B. (2013). High-throughput isolation of giant viruses of the Mimiviridae and Marseilleviridae families in the Tunisian environment. *Environmental Microbiology*, *15*, 2000–7.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M. G., Suzan-Monti, M., La Scola, B., Koonin, E. V., Raoult, D. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 21848–21853.

Boyer, M., Madoui, M. A., Gimenez, G., La Scola, B., Raoult, D. (2010). Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. *PLoS ONE*, *5*, 8.

Colson, P., Raoult, D. (2010). Gene repertoire of amoeba-associated giant viruses. *Intervirology*, *53*, 330–343.

Colson, P., Yutin, N., Shabalina, S. A., Robert, C., Fournous, G., La Scola, B., Raoult, D., Koonin, E. V. (2011a). Viruses with More Than 1,000 Genes: Mamavirus, a New Acanthamoeba polyphaga mimivirus Strain, and Reannotation of Mimivirus Genes. *Genome biology and evolution*, *3*, 737–742.

Colson, P., Gimenez, G., Boyer, M., Fournous, G., Raoult, D. (2011b). The giant Cafeteria roenbergensis virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of Life. *PLoS ONE*, 6(4):e18935.

Colson, P., De Lamballerie, X., Fournous, G., Raoult, D. (2012). Reclassification of Giant Viruses Composing a Fourth Domain of Life in the New Order Megavirales. *Intervirology*, *55*, 321–332.

Colson, P., Fancello, L., Gimenez, G., Armougom, F., Desnues, C., Fournous, G., Yoosuf, N., Million, M., La Scola, B., Raoult, D. (2013). Evidence of the megavirome in humans. *Journal of clinical virology, 57*(3), 191–200.

Desnues, C., Boyer, M., Raoult, D. (2012). Sputnik, a virophage infecting the viral domain of life. *Advances in virus research*, *82*, 63–89.

Filée, J., Pouget, N., Chandler, M. (2008). Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. *BMC Evolutionary Biology*, *8*, 320.

Fischer, M. G., Allen, M. J., Wilson, W. H., Suttle, C. A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. *Proceedings of the National Academy of Sciences*, *107*, 19508–13.

Ghigo, E., Kartenbeck, J., Lien, P., Pelkmans, L., Capo, C., Mege, J. L., Raoult, D. (2008). Ameobal pathogen mimivirus infects macrophages through phagocytosis. *PLoS pathogens*, *4*, e1000087.

Horn, M., Wagner, M. (2004). Bacterial endosymbionts of free living Amoebae. *J.Eukaryot.Microbiol.*, *51*, 509–514.

Iyer, L. M., Aravind, L., Koonin, E. V. (2001). Common origin of four diverse families of large eukaryotic DNA viruses. J.Virol. 75, 11720-11734.

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

Khan, M., La Scola, B., Lepidi, H., Raoult, D. (2007). Pneumonia in mice inoculated experimentally with Acanthamoeba polyphaga mimivirus. *Microbial Pathogenesis*, *42*(2-3), 56–61.

La Scola, B., Audic, S., Robert, C., Jungang, L., De Lamballerie, X., Drancourt, M., Birtles, R., Claverie, J. M., Raoult, D. (2003). *A giant virus in amoebae*. *Science,* 299: 2033.

La Scola, B., Marrie, T. J., Auffray, J. P., Raoult, D. (2005). Mimivirus in pneumonia patients. *Emerging infectious diseases*, *11*, 449–452.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G., Merchat, M., Suzan-Monti, M., Forterre, P., Koonin, E. V., Raoult, D. (2008). The Virophage as a Unique Parasite of Giant Mimivirus. *Nature*, *455*, 100-104.

La Scola, B., Campocasso, A., N'Dong, R., Fournous, G., Barrassi, L., Flaudrops, C., Raoult, D. (2010). Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. *Intervirology*, *53*, 344–353.

McLysaght, A., Baldi, P. F., Gaut, B. S. (2003). Extensive gene gain associated with adaptive evolution of poxviruses. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 15655–60.

Moliner, C., Fournier, P. E., Raoult, D. (2010). Genome analysis of microorganisms living in amoebae reveals a melting pot of evolution. *FEMS Microbiology Reviews*, *34*, 281–294.

Philippe, N., Legendre, M., Doutre, G., Couté, Y., Poirot, O., Lescot, M., Arslan, D., Seltzer, V., Bertaux, L., Bruley, C., Garin, J., Claverie, J. M., Abergel, C. (2013). Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, *341*, 281–286.

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

Raoult, D., La Scola, B., Birtles, R. (2007). The discovery and characterization of Mimivirus, the largest known virus and putative pneumonia agent. *Clinical infectious diseases*, *45*, 95–102.

Raoult, D., Boyer, M. (2010). Amoebae as genitors and reservoirs of giant viruses. *Intervirology*, *53*, 321–329.

Rodriguez, J. M., Yañez, R. J., Pan, R., Rodriguez, J. F., Salas, M. L., Viñuela, E. (1990). Multigene families in African swine fever virus: family 505. *Journal of Virology*, *68*, 2064–2072.

Saadi, H., Pagnier, I., Colson, P., Cherif, J. K., Beji, M., Boughalmi, M., Azza, S., Armstrong, N., Robert, C., Fournous, G., La Scola, B., Raoult, D. (2013a). First isolation of Mimivirus in a patient with pneumonia. *Clinical infectious disease*, *57*, e127–34.

Saadi, H., Reteno Ikanga, D., Colson, P., Aherfi, S., Minodier, P., Pagnier, I., Raoult, D., La Scola, B. (2013b). Shan virus, isolation of a new Mimivirus from the stool of a Tunisian patient with pneumonia. *Intervirology*. 56(6):424-9.

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V; Wommack, K. E; Noordeloos, A. A; Brussaard, C. P; Claverie, J. M. (2013). Genome of Phaeocystis globosa virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proceedings of the National Academy of Sciences*, 110(26):10800-5.

Senkevich, T. G., Koonin, E. V, Bugert, J. J., Darai, G., Moss, B. (1997). The genome of molluscum contagiosum virus: analysis and comparison with other poxviruses. *Virology*, *233*, 19–42.

Thomas, V., Greub, G. (2010). Amoeba/amoebal symbiont genetic transfers: lessons from giant virus neighbours. *Intervirology*, *53*, 254–267.

Thomas, V., Bertelli, C., Collyn, F., Casson, N., Telenti, A., Goesmann, A., Croxatto, A., Greub, G. (2011). Lausannevirus, a giant amoebal virus encoding histone doublets. *Environmental Microbiology*, *13*, 1454–1466.

Vincent, A., La Scola, B., Papazian, L. (2010). Advances in Mimivirus pathogenicity. *Intervirology*, *53*, 304–309.

Yutin, N., Koonin, E. V. (2009). Evolution of DNA ligases of nucleo-cytoplasmic large DNA viruses of eukaryotes: a case of hidden complexity. *Biology direct*, *4*, 51.

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

Yutin, N., Koonin, E. V. (2012). Hidden evolutionary complexity of Nucleo-Cytoplasmic Large DNA viruses of eukaryotes. *Virology journal*, *9*, 161.

Yutin, N., Raoult, D., Koonin, E. V. (2013a). Virophages, polintons, and transpovirons: a complex evolutionary network of diverse selfish genetic elements with different reproduction strategies. *Virology Journal*, *10*, 158.

Yutin, N., Colson, P., Raoult, D., Koonin, E. V. (2013b). Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virology Journal*, *10*(106).

ACKNOWLEDGEMENTS

I believe, writing thesis would not be so hard as acknowledging people who stood with me in completing my thesis. I would not have been able to complete my thesis without the help and support of countless people over the past three years.

I must express my sincere gratitude to my guide Professor Philippe Colson for his guidance, support and encouragement. In the field of genomics, when I was naive at the start of my thesis, he taught me the various skills and methods needed. He not only gave me the feedbacks, but many a times, he helped me in understanding the problems and also in writing the manuscripts. Without his guidance and constant feed backs, this PhD would not have been achievable.

I would like to pay my heartiest gratitude to my lab director Professor Didier Raoult, for his constant suggestions and guidance through work in progress. I also would like to thank him for creating a scientific environment at URMITE to learn and improve my skills and also I would like to thank him for providing me the financial help (AP-HM) to make my life easier in France.

I am indebted to express my thanks to Professor Eugene V Koonin and Natalya Yutin for the collaborated work. The interactions with professor Koonin and his highly valuable inputs helped me in bettering my thesis studies.

I would like to thank the reviewers of my thesis, Professor Bruno Pozzetto and Dr Hervé Lecoq for their scientific advises and detailed review during the preparation of my thesis. There sincere suggestions indeed helped me to improve my thesis. I thank Professor Jean-Marc Rolain for his support and honoring me by acting as the president of my thesis jury.

I am indeed thankful to the core bioinformatics team for helping me in solving various technical issues. I express my hearty thanks to Ghislain, Gregory, Fabrice and Olivier for their constant support. My thesis completion would have been harder without these guys.

I owe a special thanks to Catherine Robert and her team, especially Titti for teaching me molecular biology techniques. I also express my thanks to Nicholas Armstrong. I remember here their time and patience.

I am thankful to Francine Simula, Valerie Filosa and Sylvain Buffet for their administrative support and their constant help.

My friends in France, India and other parts of the world were my sources of laughter, joy, happiness and support. I am happy that, in many cases, my friendships with you have extended well beyond our shared times. I owe a special thanks to all those guys for keeping me spirited.

I need to give a special thanks to Prajakta, who stood with me on my happy and difficult times.

Last but not least, I would like to express my sincere gratitude to my mother Rahila, father Yoosuf, brother Malik Mohamed, Sister Seema, brother in law Shihas, my nephew and niece Saahil and Siya, and my grandma Asma for their unconditional love and support. I dedicate my thesis to mummy who is my first teacher, my inspiration, encouragement and support.