Intervirology

Intervirology 2010;53:310–320
DOI: 10.1159/000312916

Published online: June 15, 2010

Classification and Determination of Possible Origins of ORFans through Analysis of Nucleocytoplasmic Large DNA Viruses

Mickaël Boyer[a] Gregory Gimenez[a] Marie Suzan-Monti[a, b] Didier Raoult[a]

[a]URMITE, Centre National de la Recherche Scientifique UMR IRD 6236, Faculté de Médecine, Université de la Méditerranée, Marseille, and [b]INSERM U912, Torrents, France

Key Words
Genome evolution · Giant virus · Marseillevirus · Mimivirus · Nucleocytoplasmic large DNA virus · ORFan

Abstract
Objective: An important proportion of coding sequences in genomes, notably in viruses, do not match any sequences in databases and are assigned as ORFan sequences. Nucleocytoplasmic large DNA viruses (NCLDVs) harbor great numbers of ORFs with a high number consisting of ORFans. Thus, we decided to decipher the nature of ORFans in the NCLDVs. ***Methods:*** A genome-wide study was carried out to estimate the ORFan proportion in NCLDV genomes and to analyze their general features compared with non-ORFan. ***Results:*** The ORFan percentages comprised between 2.8 and 75.2% of the ORF content according to the virus lineage. We propose to classify ORFans in four categories according to their possible match with metagenomic sequences and their prevalence at different taxonomic ranks. Our results indicate that NCLDV ORFans have overall similar features with non-ORFans, except they are shorter. ***Conclusions:*** An ORFan classification scheme was proposed to decipher their origin and evolution. Most ORFans were likely labeled ORFan owing to the gap of knowledge of the sequence space. ORFans might be true functional genes with likely the same expression potential as non-ORFan genes. Part of them may also correspond to new genes formed de novo through the diverse mechanisms of gene evolution.

Introduction

In most cases the functions of proteins deduced from sequenced genomes are assigned by correlating sequences or structural features to previously characterized proteins. However, a significant number (~9% in bacterial genomes) of open reading frames (ORFs) is referred to as ORFans because they do not match any other ORF in the sequence databases [1]. 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. A striking observation is that the global proportion of ORFans continues to be stable despite the increasing number of sequenced genomes [2]. Thus, ORFans represent a great diversity of genes whose functions remain to be explored.

Some hypotheses suppose that ORFans originate from duplications of rapidly evolving genes, lateral gene transfer [3], or might correspond to de novo created genes [4]. To examine these possibilities, research has been under-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0300–5526/10/0535–0310$26.00/0

Accessible online at:
www.karger.com/int

Didier Raoult
Faculté de Médecine, Unité de Rickettsies
Université de la Méditerranée, 27 Boulevard Jean Moulin
FR–13385 Marseille Cedex 5 (France)
Tel. +33 491 324 375, Fax +33 491 387 772, E-Mail didier.raoult@gmail.com

taken to understand the origin, evolution and function of ORFans. In *Rickettsia*, genomes are exposed to reductive evolutionary processes; thus, it has been suggested that most of the ORFans result from the degradation of genes that were present in an ancestor of the modern *Rickettsia* species [4]. Further, several studies propose that ORFans represent genes of viral origin [2, 5], notably in bacteria where ORFans are apparently acquired from phages [6, 7]. Experimental studies on ORFans from some bacteria (e.g. *Escherichia coli, Buchnera* and *Halobacterium*) [8–10], as well as 3D structure predictions of ORFan products [11, 12], suggest that some portion of them correspond to true, expressed and functional proteins. The shorter ones, however, likely result from false-positive gene prediction [13].

In viral genomes, ORFans reach even greater proportions (~30%) than in other microorganisms [2]. Thus, the issues mentioned above concerning the origin and role of ORFan genes are highly relevant for viral genomic studies. In bacteriophages, the number of ORFans continues to grow as more genomes are sequenced [14]. Moreover, marine virome studies estimate that more than 91% of marine viral genes are new genes [15]. Therefore, viral metagenomic analyses showed that the great majority of viral genetic material (>70%) is uncharacterized and corresponds to ORFan sequences, possibly representing new viral genes without a significant match in current databases because the real diversity of viruses in nature has not been adequately sampled [16–18]. Moreover, owing to the critical role of lateral gene transfer in viral genome evolution [19, 20], we speculate that viral ORFans largely contribute to the vast diversity of viruses [14].

The increasing number of available viral genome sequences enables comparative genomic analyses to understand the evolutionary history of a virus. Among large DNA viruses, *Herpesviridae*, *Baculoviridae* and *Poxviridae*, as well as bacteriophage families, have each been characterized as monophyletic [21–24]. Although large DNA viruses harbor a great number of ORFs, the majority of which are described as ORFans and are exclusive to one virus family, a cladistic analysis using a small set of 31 conserved protein-encoding genes between those viruses revealed a higher-order relationship between several groups of large DNA viruses; this latter group was named nucleocytoplasmic large DNA viruses (NCLDVs) [25, 26]. The NCLDV group was therefore based upon the identification of a core gene that was putatively inherited from a common ancestor, implying a monophyletic evolution of NCLDVs. The NCLDV group was initially restricted to six families *(Poxviridae*, *Iridoviridae*, *Ascoviridae*, *Asfarviridae*, *Phycodnaviridae* and *Mimiviridae)* but has been further enlarged by the discovery of Marseillevirus [27]. Thus, investigations of giant viruses contribute to the understanding of their evolution and also show a large heterogeneity in NCLDV genome size (from 100 kb to 1.2 Mb). Genome analysis of newly discovered NCLDV members show that they often display a great number of ORFans (e.g. 70% for Mimivirus), which are exclusive to a virus family. Therefore, the origin and function of proteins encoded by this high proportion of probable coding sequences remain to be explored. In this study, we performed a genome-wide analysis to: (1) estimate the proportion of ORFans in NCLDV genomes at the species (singleton and doubleton ORFans) or at different lineage levels (Box 1), considering the current databases enriched with the metagenomic sequences, and (2) to describe the general features of ORFans in comparison with non-ORFan genes.

Box 1. Glossary of ORFans

ORFan: ORF that lacks homolog in a given database (RefSeq in this study).

MetaORFan: ORFan that has homolog(s) in environmental databases.

ORFan singleton: ORFan that corresponds to unique predicted gene owing to absence of significant homolog either in its residing genome or outside.

ORFan multipleton: ORFan that has one or more ORFan paralogs in its residing genome but none in other genomes. ORFan doubleton, tripleton, …, n-ton, designate ORFans that have respectively, one, two, …, n-1 paralogs in its residing genome but none outside.

Lineage ORFan: ORFan that has only homologs among a given taxonomic rank and none outside (e.g. species ORFan corresponds to ORFan set that encompasses singletons and multipletons; genus ORFan to ORFan that has only homologs among species of the same genus and none outside).

Materials and Methods

Genomic Data

NCLDV genome sequence data were downloaded from the collection of virus genomic sequences of the NCBI reference sequence (NCBI RefSeq collection) using the following accession numbers: *Acanthamoeba polyphaga* mimivirus (APMV [GenBank: AY653733]), African swine fever virus (ASFV [GenBank: U18466]), *Aedes taeniorhynchus* iridescent virus (ATIV [Gen-

Bank: DQ643392]), Invertebrate iridescent virus 6 (IIV-6 [GenBank: AF303741]), Infectious spleen and kidney necrosis virus (ISKNV [GenBank: AF371960]), Lymphocystis disease virus – isolate China (LDV-IC [GenBank: AY380826]), Singapore grouper iridovirus (SGIV [GenBank: AY521625]), Marseillevirus (MarV [GenBank: GU071086]), Canarypox virus (CPV [GenBank: AY318871]), *Melanoplus sanguinipes* entomopoxvirus (MSEV [GenBank: AF063866]), *Emiliania huxleyi* virus 86 (EhV-86 [GenBank: AJ890364]), *Ectocarpus siliculosus* virus 1 (EsV-1 [GenBank: AF204951]), *Ostreococcus tauri* virus 1 (OtV-1 [GenBank: FN386611]), *Paramecium bursaria Chlorella* virus NY-2A (PBCV-NY2A [GenBank: DQ491002]), *Heliothis virescens* ascovirus 3e (HvAV-3e [GenBank: EF133465]). To construct a phylogenetic tree of DNA polymerase (B family) sequences, multiple sequence alignment was performed using MUSCLE [28], gaps were removed automatically, and a Maximum Likelihood tree was constructed using the TreeFinder program [29] with the WAG substitution model [30].

ORF Identification

For each NCLDV genome, ORFs were searched by using GeneMark.hmm 2.0 [31] with the standard genetic code. The predicted ORFs were selected according to the following criteria: (1) longer than 150 bp, and (2) no major overlap with the adjacent ORFs. The analyses described in this study were based on the ORF set determined for each viral genome.

Homology Search and ORFans Identification

The predicted NCLDV ORF sequences were queried against the NCBI RefSeq protein sequence database (7,044,477 sequences available at NCBI in August, 2009) using BLASTX [32]. Homology searches were also performed against env_nr (environmental non-redundant) protein sequence databases (6,028,192 sequences available at NCBI in August, 2009) which encompasses sequence data from metagenomic studies. Viral ORFans were identified if their BLASTX E-value was lower than 1e-03 (for alignment lengths <80, we used 1e-05 instead). This E-value threshold has been used in previous works to define ORFans [2, 14]. Beyond each viral genome analysis, we defined metaORFans, ORFan singletons and multipletons, delineated by similarity-based clustering of ORFan protein sequences using the BLASTClust program [32]. We also defined genus or sub-family ORFans with the criteria defined above. Single or duplicated ORFans were also determined for these last categories using the BLASTClust program.

ORF Compositional Features and Statistical Analyses

For each viral genome analyzed, GC% averages of ORFans and non-ORFans were calculated over the entire genes and at each codon position. Statistical tests were conducted by using R language [33].

RNA Extraction and RT-PCR Analyses

A. polyphaga was seeded at 4×10^5 cells/ml in Page's amoebal saline [34], infected with titrated APMV at an amoeba cell:virus ratio of 1:10 and centrifuged at 1,000 *g* for 30 min. RNA was extracted from APMV-infected *A. polyphaga* at different time points postinfection (p.i.) and from uninfected *A. polyphaga* as a negative control. RNA was extracted using the commercially available RNeasy Mini Kit (Qiagen, Courtaboeuf, France), following the manufacturer's instructions. The sequences of the APMV-specific primers used are listed in table 1. The sequences of the *A. polyphaga* 18S RNA-specific primers were as previously published [35]. RT-PCR was performed using a SuperScript One-Step RT-PCR kit (Invitrogen, Carlsbad, Calif., USA), and amplified products were analyzed as previously described [35].

Results

ORFan Identification in NCLDV Representatives

To compare the ORFan proportions in each viral genome, we decided to re-evaluate the coding potential by using the same ORF prediction method for each genome. One viral representative (fig. 1) was selected in each NCLDV genus (or sub-family for *Poxviridae*) based on the largest genome size of the RefSeq genome database (fig. 1), and its genome was submitted to new ORF prediction according to the criteria described in the 'Materials and Methods'. In this way, ORF set prediction for each viral genome was normalized and did not hinge on different criteria that were defined when the annotations were individually made and submitted to GenBank. In this study, a predicted ORF was assigned as an ORFan if a BLASTX search against RefSeq displayed no significant match outside its resident genome. Thus, 38.2% of the predicted ORFs in all viral genomes showed no match against RefSeq and were classified ORFans. This proportion was slightly higher than that described from all viral genomes (30.0% in [14]) but much larger than that of bacterial ORFans (9.1% in [2]). However, ORFan percentages (table 2) showed a large range of variation [between 2.8% (PBCV-NY2A) and 75.2% (EhV-86)] according to the type of virus.

We then examined the sequence data from environmental DNA sampling, which includes more than 6 million predicted metagenomic proteins, to search for possible NCLDV ORFan relatives. The percentage of metaORFans (Box 1) identified in environmental sequence databases was calculated for the viral representative of each NCLDV family (fig. 1 and table 2). Analyses of ORFan homolog searches in environmental databases were carried out with the previously defined ORFan set. Therefore, this ORFan set, without metaORFans, corresponds to a species ORFan set. Overall, we noticed that the metaORFan proportion was 3.5% of the predicted ORFs in all viral genomes, but match results were quite different according to the virus. For example, 63 out of 474 APMV virus ORFans (6.9% of the APMV ORF set) had closest matches (BLASTX E-values ranging from 10^{-41} to 10^{-3}) to environmental sequences. Surprisingly,

Table 1. ORFan expression analysis in APMV

APMV ORFan[1]	Proteomic detection [41]	RT-PCR detection				
		forward/reverse primer sequences	0 h[2, 3]	2 h	4 h	8 h	16 h
MIMI_L48	+						
MIMI_L152		ggcacttcaaggttcacgat/cccatgggaattcttccttt	+[3]	+	+	+	+
MIMI_L208	+						
MIMI_L226		ggatttgaaccgactgatgaa/ttgatgtgatgcgatttcaa	+[3]	+	+	+	+
MIMI_L272		tggatttgacgaattgatgg/tgttctgttgaccggattgt	–	–	+	+	+
MIMI_L274	+						
MIMI_L283		cgaatctaaccgatccgaaa/cccaaatacaccgcgataaa	+	+	+	+	+
MIMI_L291		ccacaaatgaatccgtcaca/tcgctgagagaaggtggttt	+/–	+	+	+	+
MIMI_L309	+						
MIMI_L330	+						
MIMI_L352	+						
MIMI_L356		attccaccgtccaaatacga/ttgggaaaactgtttcttcgat	+[3]	+	+	+	+
MIMI_L389	+						
MIMI_L399	+						
MIMI_L442	+						
MIMI_L452	+						
MIMI_L485	+						
MIMI_L488	+						
MIMI_L492	+						
MIMI_L520		tgtcccaattcgaaattaaaa/atcggaaaacaacaggatca	+[3]	+	+	+	+
MIMI_L533	+						
MIMI_L550	+						
MIMI_L585	+						
MIMI_L591	+						
MIMI_L611		aacgaacaatgtttgcgtca/aattgtccgtcttccaatcg	+[3]	+	+	+	+
MIMI_L647	+						
MIMI_L688	+						
MIMI_L724	+						
MIMI_L725	+						
MIMI_L768		ttgttgatgttcaaggggtaa/cggccattgttgacatttct	+	+	+	+	+
MIMI_L769		ggaaatttcatcaatgccaaa/ttccaacatcgtgacattcc	–	+	+	+	+
MIMI_L778	+						
MIMI_L851	+						
MIMI_L872	+						
MIMI_L899	+						
MIMI_R160	+						
MIMI_R217		aaccgacagatcgtgatgaa/ttctttttcagacggaatgtga	–	–	+	+	+
MIMI_R219		gccatcaatcaagtggaaaaa/ccgatccttgtttaattgctg	–	–	+	+	+
MIMI_R326	+						
MIMI_R338		aaccttgggttgtctcgatg/attgattgaaatccgccaaa	–	+	+	+	+
MIMI_R347	+						
MIMI_R349		tgactgtggacctcgatctg/ccgcttggacgaataacaat	+[3]	+	+	+	+
MIMI_R387	+						
MIMI_R403	+	tcaatccagcagcatttcag/cgtcgcaagatgaaacaaga	+	+	+	+	+
MIMI_R457	+						
MIMI_R459	+						
MIMI_R463	+						
MIMI_R557	+						
MIMI_R584	+						
MIMI_R646	+						
MIMI_R653	+						
MIMI_R658	+						
MIMI_R679	+						

Table 1 (continued)

APMV ORFan[1]	Proteomic detection [41]	RT-PCR detection				
		forward/reverse primer sequences	0 h[2, 3]	2 h	4 h	8 h	16 h
MIMI_R691	+						
MIMI_R692	+						
MIMI_R695	+						
MIMI_R705	+						
MIMI_R710	+						
MIMI_R727	+						
MIMI_R734		cacggaacaggaactcacaa/tgggagttaatttcgggaca	+[3]	+	+	+	+
MIMI_R748		tcggttttgcgactcaaatag/tgcttgaattgattcgggtaa	+[3]/–	+	+	+	+/–
MIMI_R822		tcgcaacatcccaaatacaa/gtttcctgcagccaattcat	+	+	+	+	+
MIMI_R865		tgtgtttcggatgtcgagaa/ttcttttgcgatttggttcc	–	+/–	+	+	+
MIMI_R871		tctccggtaagacgcagatt/ccaaatttttgctccgtcat	+	+	+	+	+

[1] APMV ORFans corresponded to those identified in genome annotation [54].
[2] Time post-infection for RT-PCR transcript detection in RNA extracted from APMV-infected *A. polyphaga*.
[3] Indicates ORFan transcript associated with the viral particles.

Table 2. ORFan classification in NCLDV genome representatives

NCLDV species	Detected ORF	ORFan (%[1])	% MetaORFan (single/ duplicated)	% Genus ORFan (single/ duplicated)	Species ORFan							
					% species ORFan (singleton/ multipleton)	multipleton						
						double-ton	triple-ton	4-ton	5-ton	7-ton	9-ton	10-ton	14-ton
APMV	986	474 (48.1)	6.9 (61/2)	–	41.7 (337/74)	23	1	2	–	1	–	1	–
EhV-86	459	345 (75.2)	4.8 (22/0)	–	70.4 (296/26)	10	2	–	–	–	–	–	–
MarV	449	305 (67.9)	5.8 (24/2)	–	62.1 (221/58)	11	3	1	–	–	1	–	1
PBCV-NY2A	390	11 (2.8)	0.3 (1/0)	36.2 (129/12)	2.6 (10/0)	–	–	–	–	–	–	–	–
CPV	312	31 (9.9)	–	8.3 (24/2)	9.9 (25/6)	3	–	–	–	–	–	–	–
EsV-1	238	30 (12.6)	–	2.5 (6/0)	12.6 (21/9)	1	–	–	–	1	–	–	–
OtV-1	231	8 (3.5)	1.7 (4/0)	1.7 (4/0)	1.7 (4/0)	–	–	–	–	–	–	–	–
MSEV	224	65 (29.0)	0.9 (2/0)	11.2 (25/2)	28.1 (58/5)	1	1	–	–	–	–	–	–
IIV-6	200	86 (43)	4.5 (9/0)	–	38.5 (75/2)	1	–	–	–	–	–	–	–
HvAV-3e	165	31 (18.8)	–	6.1 (10/0)	18.8 (31/0)	–	–	–	–	–	–	–	–
SGIV	136	45 (33.1)	–	11.0 (11/4)	33.1 (29/16)	2	1	1	1	–	–	–	–
LDV-IC	135	8 (5.9)	–	6.7 (9/0)	5.9 (8/0)	–	–	–	–	–	–	–	–
ASFV	134	91 (67.9)	10.4 (14/0)	–	57.5 (60/17)	1	–	2	–	1	–	–	–
ATIV	123	45 (36.6)	4.9 (6/0)	–	31.7 (39/0)	–	–	–	–	–	–	–	–
ISKNV	111	64 (57.7)	0.9 (1/0)	–	56.8 (61/2)	1	–	–	–	–	–	–	–

[1] Percentages were calculated in comparison with total number of ORF for each species.

ORFans from viruses LDV-IC, SGIV, CPV, EsV-1 and HvAV-3e, which respectively belong to four different families *(Iridoviridae*, *Poxviridae, Phycodnaviridae* and *Ascoviridae)*, had no significant match against the environmental databases, whereas up to 14 out of 91 ASFV virus ORFans (10.4% of the ASFV ORF set) were converted to metaORFans. Thus, the presence of numerous ASFV virus ORFan homologs in the metagenomic database suggests that sequences clearly related to asfarvirus genes (BLASTX E-values ranging from 10^{-93} to 10^{-4}) are abundant in the sampled environment (or at least abundant enough to be collected by environmental sampling) [36, 37].

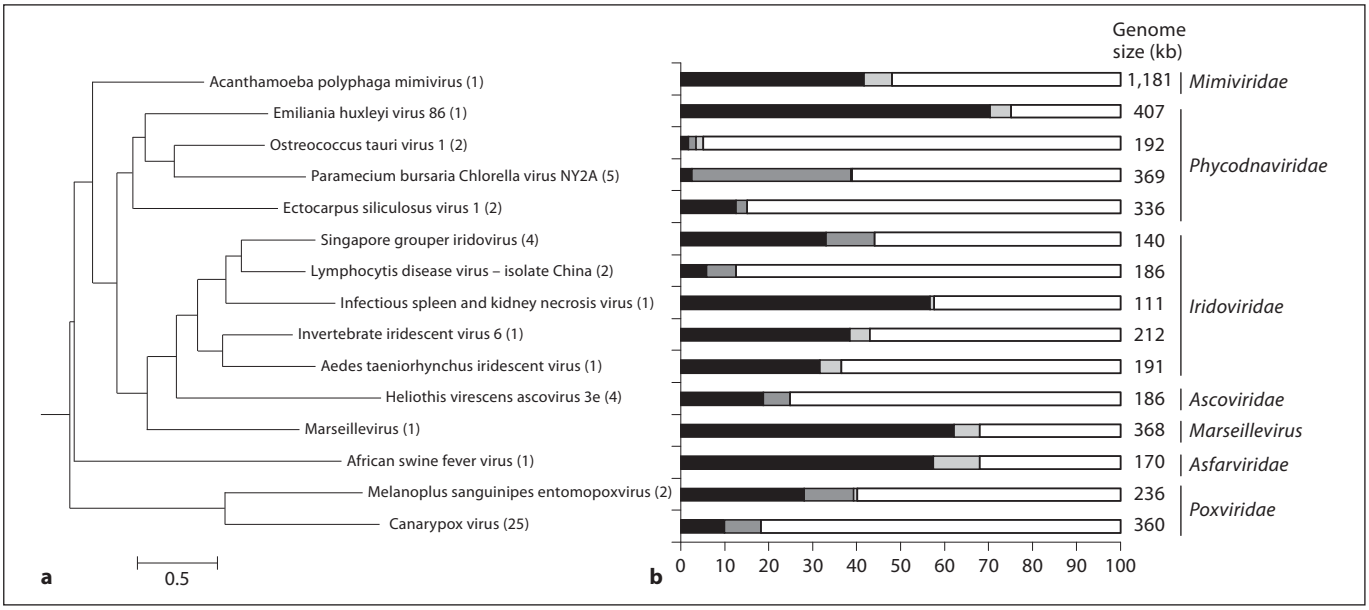

Fig. 1. Maximum-likelihood tree of NCLDV representatives and prevalence of ORFans categories within their respective genome sequences. (**a**) The maximum-likelihood tree was based on DNA polymerase (B family) conserved protein sequences of each viral species. The number of sequenced genomes in each NCLDV genus available in Refseq is also indicated in parentheses next to the names of genus representatives. (**b**) Percentage of each ORFs categories within the studied genomes: species ORFans (black bars), genus ORFans (or sub-families ORFans for *Poxviridae*) dark grey bars), metaORFans (light grey bars), and non-ORFans (white bars). Virus genome sizes and viral taxonomic families are indicated on the right.

ORFan Classification inside Viral Families

When genomes of closely related viruses are sequenced, the ORFan proportions in these genomes are low because they share a high number of homologous sequences. We noticed a clear correlation between the number of available sequenced genomes in a genus and the ORFan proportion in a genome of a related virus of the same genus (fig. 1). Thus, as with species ORFans, we also identified genus ORFans, which correspond to ORFs of a given NCLDV genome only having homologs in related viruses of the same genus. Considering this additional category, the proportion of genus ORFans could greatly change in comparison with the proportion of species ORFans (fig. 1 and table 2). For instance, the PBCV-NY2A genome contains 2.6% of ORFs that are only found in its genome (species ORFans), but 36.2% match only ORFs from species of the same genus (i.e. genus ORFans). Thus, the high proportion of genus ORFans in this species might indicate that this category of ORFans is strongly involved in viral diversity among genus representatives of the same family. Beyond lineage ORFan identification, we also determined the proportion of ORFans singletons and multipletons inside the species ORFan sets, along with the proportion of single or duplicated ORFans inside the genus and sub-family ORFan sets (table 2). We observed that duplication exerts a strong influence on ORFan evolution because the proportion of ORFan multipletons reached 14% of total ORFs as for SGIV (table 2).

Features of the Species and Genus ORFans

We then investigate whether both species and genus viral ORFans identified in our NCLDV representatives harbor specific characteristics compared to non-ORFans. We considered both categories of ORFans because they will likely remain ORFan in the future, regardless of the number of sequenced genomes in the genus. Indeed, their chance of being converted to non-ORFan remains low since these both categories of ORFans, lacking homologs in the majority of currently known genomes, appear to be specific to a given genus. Previous studies show that ORF length is a prominent feature that differs between the ORFs and the ORFans [1]. For all of the NCLDV genomes analyzed here, ORFan length was smaller than that of non-ORFans. On average, ORFan length (mean value = 587 bp) was significantly shorter than that for non-ORFans (mean value = 1,149 bp;

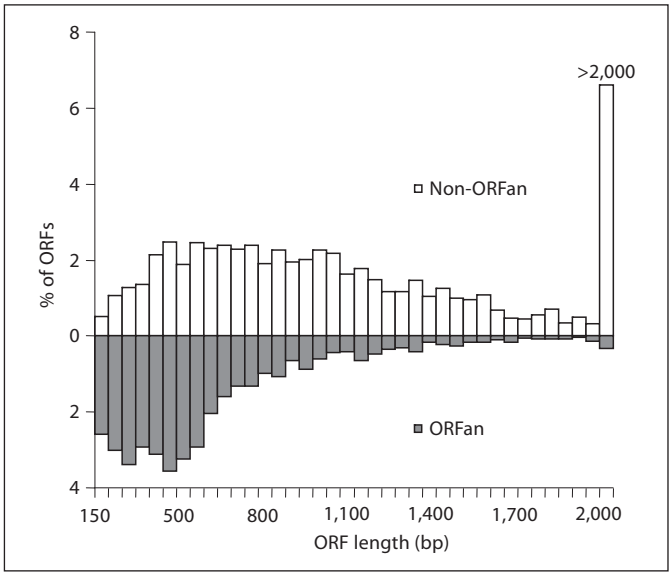

Fig. 2. Length distribution of both species and genus ORFans (grey bars) and non-ORFans (white bars) among all predicted ORFs in the studied NCLDV genomes.

$p < 2.2 \times 10^{-16}$, Wilcoxon test; fig. 2). Therefore, analysis of ORF length distribution indicated that ORFans are over-represented among the shorter ORFs in NCLDV genomes.

In addition, we observed that ORFans and non-ORFans exhibit a similar nucleotide composition pattern (considering GC content; fig. 3a). However, we hypothesized that ORFans, being in essence non-conserved ORFs, are under weaker selection constraints than non-ORFans and would thus more easily tolerate non-synonymous mutations. To test this hypothesis, we conducted GC content analysis at the three codon positions of both ORFans and non-ORFans. Variation in GC content between the three positions in a codon is a general feature often described in genomic studies [38, 39]. Usually, the third synonymous codon position shows the largest variation range, and the second codon position exhibits the lowest variation (a change at the second position is always non-synonymous). The level of variation is correlated to the degree of selective pressure exerted at a specific position. Our results clearly indicated high selective constraints at the first/second position in comparison with the third position for both NCLDV ORFans and non-ORFans, since the measure of GC content showed high dispersion at the third codon position but lowest at the first and second positions (fig. 3b, c). This analysis gives a measure of selective constraints acting on the different positions. Indeed, mutational changes at the third position are silent at the protein level. The slopes of linear regressions at the second position, which is subject to the strongest selective constraints, were not significantly different between ORFans and non-ORFans (p value = 0.6259, χ^2 test), indicating that similar selective pressure was exerted on the two ORF classes.

Discussion

A Proposition for ORFan Classification

ORFan definition depends on several parameters, including E-value cutoff, species representation in databases and the size of databases. Indeed, each new sequenced genome added to a database could modify the status ORFan. Thus, the increase of sequenced genomes in a viral family could lead to conversion of labeled ORFans in a genome of this family to non-ORFans with possible assignment to a known protein family, particularly if closely related genomes have been sequenced. We now propose that ORFans be classified into four categories according to their possible match with metagenomic sequences (metaORFans), their prevalence at the species level (ORFan singletons and ORFan multipletons) and at higher taxonomic ranks (lineage ORFans) (Box 1 and table 3).

Homology searches against metagenomes should be investigated for each new sequenced genome; thus, the introduction and use of the term 'metaORFan' would easily identify ORFs having only homolog(s) in environmental databases. At the species level, we proposed to differentiate ORFan singletons, which correspond to unique predicted genes in the current databases, from ORFan multipletons, which are likely derived from one or more duplications inside a given genome. We have also noticed that the ORFan proportion in each genome clearly depends upon the number of sequenced genomes belonging to species in the same genus. We have therefore considered genus ORFans in our analysis, which allowed a better comparison of ORFans proportions between genomes of different viral species. ORFan classification could be a way to further study conservation and evolution of ORFs having no known homologs outside a limited taxonomic group.

ORFan Proportions in NCLDV Lineages

ORFan assignment in the different categories described above could help to further characterize sequenced genomes, notably parts of them without any known matches. In a recent study [37], analysis of marine metagenomes supposedly dominated by prokaryotes re-

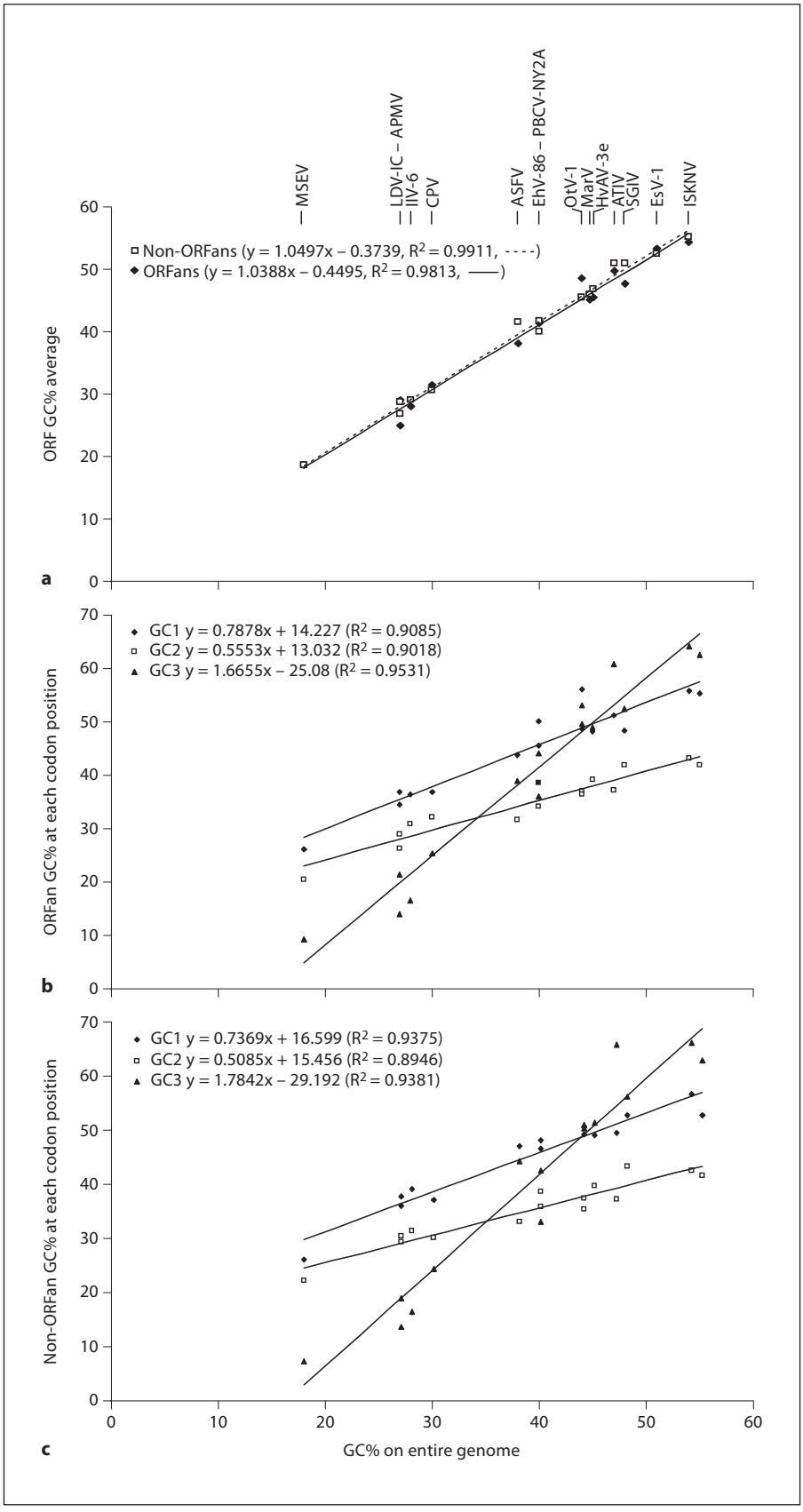

Fig. 3. Correlation between genomic GC content and the GC content of the species and genus ORFan or non-ORFan sequences (**a**) and at each codon position (**b**, **c**) within the NCLDV representatives. Linear regression slopes with their corresponding equation are also shown. Abbreviation names for each species are positioned above (**a**) according to the respective GC content of the entire genome. GC1, GC2 and GC3 correspond to the GC content at the first, second and third codon position, respectively.

Table 3. Comparison of ORFans classification with previous studies

Studied genomes	ORFan categories				Reference
NCLDV	ORFan singleton	ORFan multipleton	Lineage ORFan	MetaORFan	This study
Virus	ORFan	ORFan	–	–	[14]
NCLDV	NORF	NORF	NORF	–	[34]
Bacteria	Clade-specific ORFan	Clade-specific ORFan	Clade-specific ORFan	–	[5]
Rickettsia	ORFan	ORFan	Species- or genus-specific ORFan	–	[4]
Bacteria	Singleton ORFan	Paralogous ORFan	Orthologous ORFan	–	[2, 11, 13, 55]
Bacteria	ORFan	ORFan paralogous families	–	–	[1]

NORF = Narrow taxonomic distribution ORF.

vealed a variety of sequences homologous to conserved genes of NCLDVs. In our investigation, we found that the ORFans of some NCLDV representatives had significant matches in metagenome databases and correspond to metaORFans. This suggests that the remaining ORFans without significant matches in the current databases are likely labeled as ORFans owing to a gap of knowledge in sequence space. However, the tremendous diversity of viral sequences in metagenomes [15] may explain why metaORFan proportions remain low in the different lineages. Moreover, we observed that major contributors to the discovery of new sequences containing new genes were viruses from novel viral families (e.g. MarV showing a high ORFan proportion). A high level of species ORFans in a genome might be representative of high genetic diversity between sampled sequenced viruses. We also found that the proportion of both species and genus ORFans is clearly correlated to the viral genome sampling but not to the genome size (data not shown). Thus, a large number of species ORFans can actually be converted to genus ORFans when other genomes of the same genus are sequenced. A high genus ORFan proportion (e.g. PBCV-NY2A) indicates well-conserved ORFans among species belonging to the same genus. Thus, genus ORFans are probably vertically inherited from a putative common ancestor and may be conserved among different species during evolution. Alternatively, genus ORFans might have been laterally transferred between viruses, for examples when they co-infect the same hosts, or might have been independently recruited from their hosts. Poxviruses, constituting a large viral family with well conserved genomes (many of which are sequenced) [40, 41], contain genomes displaying few new genes, which is correlated to the low proportion of both sub-family and species ORFans in Canarypox virus.

Are ORFans Functional Genes?

The high proportion of genus ORFans, being in essence conserved with at least one species of the same genus, suggests that at least a part of the ORFans pool corresponds to functional proteins. Similar to previous results obtained on several NCLDV members [42], our results indicated that NCLDV species and genus ORFans share overall similar features (similar GC% and similar GC% at each codon position) with non-ORFans except their shorter size. Thus, a percentage of ORFans might be true functional genes with the same expression potential as non-ORFans, undergoing an equal evolution rate. This conclusion is consistent with different reports showing that 95, 96, 97 and 99% of predicted coding sequences are respectively expressed in Vaccinia, Red Sea bream iridovirus, SGIV, and *Paramecium bursaria Chlorella* virus-1 at some time during the virus infectious cycle [43–47]. Thus, ORFans represent a high proportion of potentially coding sequences in the current databases, but very few viral ORFans have been experimentally characterized [48]. Studies on the APMV proteome revealed that 6–7% of ORFans encode proteins, suggesting that viral ORFans correspond to *bona fide* expressed proteins (table 1) [49]. Preliminary transcriptomic analysis revealed that at least 20 APMV virus ORFans are transcribed at some time in the virus infectious cycle (only 20 ORFans have been tested here) and that 8 of them were also found packaged within the virus particles (table 1). Furthermore, 3D models have enabled confident prediction of structures and functions for 21 and 6 APMV virus ORFans, respectively [48]. In actuality, most ORFans likely encode proteins without predicted biological functions and might play critical roles during viral infection processes.

What Are ORFans?

According to our results, viral ORFans harbor similar features as non-ORFans, suggesting that they are basically 'false ORFans', i.e. their current ORFan status is directly linked to limited knowledge of the sequence space and our ability to detect homologies. Aside from this, the shorter size of ORFans in comparison with non-ORFans may be explained by some methodological difficulties in detecting homologies for short sequences, as previously suggested [42]. This analysis, including homology searches against sequences from environmental samples, provided evidence that the number of identified ORFans per genome is strongly linked to the wealth and diversity of current sequence databases. Thus, a high ORFan number is still the result of insufficient sampling of NCLDV sequences. Furthermore, several studies have shown that lateral gene transfer strongly affects viral genome evolution, notably in phages [14, 50]. Viruses could exchange genes with their host and also with sympatric microorganisms living within the same host; this is notably evidenced in phagocytic protists like amoeba, where multiples parasites and symbionts, both viral and/or bacterial, co-exist [27, 51]. Thus, the genes repertoires of viral genomes could have diverse origins, which may explain the mosaic structure of some viral genomes [27, 52]. The high gene mixing in viral genomes could also contribute to the difficulties in identifying viral ORF homologs in databases because many ORFs could originate from organisms without sequenced genomes.

Finally, we therefore assume that ORFans correspond to newly detected genes with potential coding ability [49]. However, as well as having insufficient sampling, homolog searches and further ORFan characterization can be made more difficult if ORFans correspond to coding sequences de novo formed through gene fusion or gene degradation/extinction [3, 4]. Thus, gene modifications through these evolutionary processes could lead to the formation of small gene remnants or pseudogenes, which are classified as ORFans owing to their inconspicuous features. In the same way, ORFans can originate from the duplication of existing genes [53]. Then, duplicated sequences could diverge through fast evolutionary processes, and one of the two copies may either degenerate into a non-functional gene or gain a new function conferring a selective advantage.

The ORFan proportion in viruses is higher than bacteria, as exemplified here with NCLDVs. Thus, this ORFans pool emphasizes the high diversity in viral genomes and in the diversity of viral species. In NCLDVs whose genome sizes are similar to that of some intracellular bacteria, ORFans may actually correspond to a large number of genes of unknown function. Analysis of expression patterns of predicted coding sequences by extended proteomic studies, structural protein characterization and more structural genomic projects, including carefully conserved domain searches, will further elucidate more information about the functional status of ORFan genes [54]. The large ORFan pool of NCLDV genomes underlines the complexity of the viral world; the function of this high diversity of genes should be explored to decipher their role in infectious processes.

Acknowledgment

This work was funded by the Centre National de la Recherche Scientifique (CNRS, credits récurrents).

References

1 Fischer D, Eisenberg D: Finding families for genomic ORFans. Bioinformatics 1999;15:759–762.
2 Yin Y, Fischer D: On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. BMC Evol Biol 2006;6:63.
3 Davids W, Fuxelius HH, Andersson SG: The journey to smORFland. Comp Funct Genomics 2003;4:537–541.
4 Amiri H, Davids W, Andersson SG: Birth and death of orphan genes in *Rickettsia*. Mol Biol Evol 2003;20:1575–1587.
5 Daubin V, Ochman H: Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. Genome Res 2004;14:1036–1042.
6 Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WR Jr, Hendrix RW, Hatfull GF: Origins of highly mosaic mycobacteriophage genomes. Cell 2003;113:171–182.
7 Cortez D, Forterre P, Gribaldo S: A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol 2009;10:R65.
8 Alimi JP, Poirot O, Lopez F, Claverie JM: Reverse transcriptase-polymerase chain reaction validation of 25 'orphan' genes from *Escherichia coli* K-12 MG1655. Genome Res 2000;10:959–966.
9 Shimomura S, Shigenobu S, Morioka M, Ishikawa H: An experimental validation of orphan genes of *Buchnera*, a symbiont of aphids. Biochem Biophys Res Commun 2002;292:263–267.
10 Shmuely H, Dinitz E, Dahan I, Eichler J, Fischer D, Shaanan B: Poorly conserved ORFs in the genome of the archaea *Halobacterium* sp. NRC-1 correspond to expressed proteins. Bioinformatics 2004;20:1248–1253.
11 Siew N, Fischer D: Structural biology sheds light on the puzzle of genomic ORFans. J Mol Biol 2004;342:369–373.

12 Shin DH, Hou J, Chandonia JM, Das D, Choi IG, Kim R, Kim SH: Structure-based inference of molecular functions of proteins of unknown function from Berkeley Structural Genomics Center. J Struct Funct Genomics 2007;8:99–105.
13 Siew N, Fischer D: Analysis of singleton ORFans in fully sequenced microbial genomes. Proteins 2003;53:241–251.
14 Yin Y, Fischer D: Identification and investigation of ORFans in the viral world. BMC Genomics 2008;9:24.
15 Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F: The marine viromes of four oceanic regions. PLoS Biol 2006;4:e368.
16 Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F: Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA 2002;99:14250–14255.
17 Breitbart M, Hewson I, Felts B, Mahaffy JM, Nulton J, Salamon P, Rohwer F: Metagenomic analyses of an uncultured viral community from human feces. J Bacteriol 2003;185: 6220–6223.
18 Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, Haynes M, Liu H, Furlan M, Wegley L, Chau B, Ruan Y, Hall D, Angly FE, Edwards RA, Li L, Thurber RV, Reid RP, Siefert J, Souza V, Valentine DL, Swan BK, Breitbart M, Rohwer F: Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 2008;452: 340–343.
19 Casjens SR: Comparative genomics and evolution of the tailed-bacteriophages. Curr Opin Microbiol 2005;8:451–458.
20 Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW: Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol 2000;299:27–51.
21 Hendrix RW, Hatfull GF, Smith MC: Bacteriophages with tails: chasing their origins and evolution. Res Microbiol 2003;154:253–257.
22 Davison AJ, Stow ND: New genes from old: redeployment of dUTPase by herpesviruses. J Virol 2005;79:12880–12892.
23 Hughes AL, Friedman R: Poxvirus genome evolution by gene gain and loss. Mol Phylogenet Evol 2005;35:186–195.
24 Lauzon HA, Jamieson PB, Krell PJ, Arif BM: Gene organization and sequencing of the *Choristoneura fumiferana* defective nucleopolyhedrovirus genome. J Gen Virol 2005; 86:945–961.
25 Iyer LM, Aravind L, Koonin EV: Common origin of four diverse families of large eukaryotic DNA viruses. J Virol 2001;75: 11720–11734.
26 Iyer LM, Balaji S, Koonin EV, Aravind L: Evolutionary genomics of nucleocytoplasmic large DNA viruses. Virus Res 2006;117: 156–184.
27 Boyer M, Yutin N, Pagnier I, Barrassi L, Fournous G, Espinosa L, Robert C, Azza S, Sun S, Rossmann M, Suzan-Monti M, La Scola B, Koonin E, Raoult D: Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimaeric microorganisms. Proc Natl Acad Sci USA 2009; 106:21848–21853.
28 Edgar RC: MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32: 1792–1797.
29 Jobb G, von Haeseler A, Strimmer K: TreeFinder: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol Biol 2004;4:18.
30 Whelan S, Goldman N: A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 2001;18: 691–699.
31 Lukashin AV, Borodovsky M: GeneMark. hmm: new solutions for gene finding. Nucleic Acids Res 1998;26:1107–1115.
32 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;25:3389–3402.
33 Ihaka R, Gentleman R: R: a language for data analysis and graphics. J Comput Graph Stat 1996;5:299–314.
34 Rowbotham TJ: Preliminary report on the pathogenicity of *Legionella pneumophila* for freshwater and soil amoebae. J Clin Pathol 1980;33:1179–1183.
35 Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM: The 1.2-megabase genome sequence of Mimivirus. Science 2004;306: 1344–1350.
36 Loh J, Zhao G, Presti RM, Holtz LR, Finkbeiner SR, Droit L, Villasana Z, Todd C, Pipas JM, Calgua B, Girones R, Wang D, Virgin HW: Detection of novel sequences related to African swine fever virus in human serum and sewage. J Virol 2009;83:13019–13025.
37 Kristensen DM, Mushegian AR, Dolja VV, Koonin EV: New dimensions of the virus world discovered through metagenomics. Trends Microbiol 2010;18:11–19.
38 Muto A, Osawa S: The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 1987;84:166–169.
39 Lawrence JG, Ochman H: Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol 1997;44:383–397.
40 Lefkowitz EJ, Wang C, Upton C: Poxviruses: past, present and future. Virus Res 2006;117: 105–118.
41 Hughes AL, Irausquin S, Friedman R: The evolutionary biology of poxviruses. Infect Genet Evol 2010;10:50–59..
42 Ogata H, Claverie JM: Unique genes in giant viruses: regular substitution pattern and anomalously short size. Genome Res 2007; 17:1353–1361.
43 Lua DT, Yasuike M, Hirono I, Aoki T: Transcription program of red sea bream iridovirus as revealed by DNA microarrays. J Virol 2005;79:15151–15164.
44 Assarsson E, Greenbaum JA, Sundstrom M, Schaffer L, Hammond JA, Pasquetto V, Oseroff C, Hendrickson RC, Lefkowitz EJ, Tscharke DC, Sidney J, Grey HM, Head SR, Peters B, Sette A: Kinetic analysis of a complete poxvirus transcriptome reveals an immediate-early class of genes. Proc Natl Acad Sci USA 2008;105:2140–2145.
45 Rubins KH, Hensley LE, Bell GW, Wang C, Lefkowitz EJ, Brown PO, Relman DA: Comparative analysis of viral gene expression programs during poxvirus infection: a transcriptional map of the vaccinia and monkeypox genomes. PLoS One 2008;3:e2628.
46 Teng Y, Hou Z, Gong J, Liu H, Xie X, Zhang L, Chen X, Qin QW: Wholegenome transcriptional profiles of a novel marine fish iridovirus, Singapore grouper iridovirus (SGIV) in virus-infected grouper spleen cell cultures and in orange-spotted grouper, *Epinephulus coioides*. Virology 2008;377:39–48.
47 Yanai-Balser GM, Duncan GA, Eudy JD, Wang D, Li X, Agarkova IV, Dunigan DD, Van Etten JL: Microarray analysis of *Paramecium bursaria Chlorella* virus 1 transcription. J Virol 2010;84:532–542.
48 Saini HK, Fischer D: Structural and functional insights into Mimivirus ORFans. BMC Genomics 2007;8:115.
49 Renesto P, Abergel C, Decloquement P, Moinier D, Azza S, Ogata H, Fourquet P, Gorvel JP, Claverie JM: Mimivirus giant particles incorporate a large fraction of anonymous and unique gene products. J Virol 2006;80:11678–11685.
50 Hendrix RW, Lawrence JG, Hatfull GF, Casjens S: The origins and ongoing evolution of viruses. Trends Microbiol 2000;8:504–508.
51 Filee J, Pouget N, Chandler M: Phylogenetic evidence for extensive lateral acquisition of cellular genes by nucleocytoplasmic large DNA viruses. BMC Evol Biol 2008;8:320.
52 Hatfull GF: Bacteriophage genomics. Curr Opin Microbiol 2008;11:447–453.
53 Suhre K: Gene and genome duplication in *Acanthamoeba polyphaga* Mimivirus. J Virol 2005;79:14095–14101.
54 Monne M, Robinson AJ, Boes C, Harbour ME, Fearnley IM, Kunji ER: The mimivirus genome encodes a mitochondrial carrier that transports dATP and dTTP. J Virol 2007;81: 3181–3186.