Journal of Structural Biology 215 (2023) 107962

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Protein repeats evolve and emerge in giant viruses

Sofía Erdozain [a], Emilia Barrionuevo [b], Lucas Ripoll [c], Pablo Mier [d], Miguel A. Andrade-Navarro [d,*]

[a] Instituto de Biotecnología y Biología Molecular, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Argentina
[b] Laboratory of Bioactive Research and Development, Faculty of Exact Sciences, National University of La Plata, Argentina
[c] Laboratory of Genetic Engineering, Cell, and Molecular Biology, National University of Quilmes, Argentina
[d] Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany

ARTICLE INFO

Edited by KAJAVA Andrey

Keywords:
Giant viruses
Homorepeats
Protein repeats
Evolution of protein repeats

ABSTRACT

Nucleocytoplasmatic large DNA viruses (NCLDVs or giant viruses) stand out because of their relatively large genomes encoding hundreds of proteins. These species give us an unprecedented opportunity to study the emergence and evolution of repeats in protein sequences. On the one hand, as viruses, these species have a restricted set of functions, which can help us better define the functional landscape of repeats. On the other hand, given the particular use of the genetic machinery of the host, it is worth asking whether this allows the variations of genetic material that lead to repeats in non-viral species. To support research in the characterization of repeat protein evolution and function, we present here an analysis focused on the repeat proteins of giant viruses, namely tandem repeats (TRs), short repeats (SRs), and homorepeats (polyX). Proteins with large and short repeats are not very frequent in non-eukaryotic organisms because of the difficulties that their folding may entail; however, their presence in giant viruses remarks their advantage for performance in the protein environment of the eukaryotic host. The heterogeneous content of these TRs, SRs and polyX in some viruses hints at diverse needs. Comparisons to homologs suggest that the mechanisms that generate these repeats are extensively used by some of these viruses, but also their capacity to adopt genes with repeats. Giant viruses could be very good models for the study of the emergence and evolution of protein repeats.

1. Introduction

Nucleocytoplasmatic large DNA viruses (NCLDVs), hereafter giant viruses, display double-stranded DNA genomes with sizes from 100 kilobases up to 2.5 megabases (Koonin and Yutin, 2019). Their large genomes provide them with a potentially complicated biological machinery, while their viral "lifestyle" imposes restricted functional requirements on them save for interaction with the host (Schulz et al., 2022). In this respect, giant viruses constitute a new group of species for comparative genomics studies, constituting an extra group separated from the four traditional sets of phyla: eukaryota, bacteria, archaea and non-giant viruses.

In particular, the study of the emergence of repeated sequences in proteins has gained from comparisons across phyla; for example, early observations of the lower frequency of certain repeat types in prokaryotic organisms were taken as an indication of their functional importance in complex eukaryotic cell organization (Marcotte et al., 1999; Andrade et al., 2001). Also, the existence of certain trinucleotide repeats that could be explained by replication slippage has long been hypothesized (Moore et al., 1999), and the high frequency of certain repetitions in simple unicellular organisms has been noted, for example the polyN repeats in *S. cerevisiae* (Stewart et al., 2021), or in *P. falciparum* (Gardner et al., 2002).

To further our understanding of the origin and evolution of repeats in protein sequences, giant viruses are a very interesting target, since they use the machinery of the host for replication and have reduced functions. We hypothesized that a survey of protein repeats in these species could help shape the functional landscape of particular repeat types, while at the same time observing the emergence and evolution of repeats within viral families or obtained by horizontal transfer from non-viral partners.

With this goal in mind, in this work we profiled the proteomes of giant viruses for three types of repeats: tandem repeats (TRs) probably forming structured but flexible assemblies, short repeats (SRs), likely

* Corresponding author.
E-mail address: andrade@uni-mainz.de (M.A. Andrade-Navarro).

https://doi.org/10.1016/j.jsb.2023.107962
Received 23 November 2022; Received in revised form 21 March 2023; Accepted 4 April 2023
Available online 7 April 2023

disordered, and homorepeats (polyX), which can adopt dynamically changing structures. We categorized the types of repeats found (or missing), their functional associations, and their evolution, and discovered an unexpectedly rich variety of repeats in numbers that in some viruses exceed those in many prokaryotic and eukaryotic organisms. Our work sets the stage for elucidating processes that giant viruses actively use to evolve repeat assemblies.

2. Results and discussion

2.1. Repeat composition of proteomes of giant viruses

We evaluated the composition of repeats in the proteomes of 74 giant viruses from 10 different families (reported in (Koonin and Yutin 2019); see Methods for details) at three levels: (i) tandem repeats forming structural ensembles (using REP2 (Kamel et al., 2021)), (ii) short repeats of few amino acids (using RES (Kamel et al., 2019)), and homorepeats (tracks of a single amino acid; using polyX2 (Mier and Andrade-Navarro 2022)).

In order to have an overview of the distribution of the repeats found for each proteome among these families, we built a phylogenetic tree with the selected species and classified them taxonomically at the family level (Fig. 1; see Methods for details). The number of repeats per proteome, biological data of each virus, and repeats per protein are shown in Supplementary Table S1 and Supplementary Table S2. Each of these results can be reproduced using the respective tools as indicated in Table 1.

Our results show the composition of repeats in each proteome. Of the total amount of fourteen tandem repeats that REP2 can look for, only six were found: Ankyrin (ANK), Leucine Rich Repeats (LRRs), WD40, Tetratricopeptide repeats (TPRs), KELCH, and RCC1. We found that Ankyrin repeats were the most frequently found among these viruses, followed by LRRs. These two types of repeats were found in all the families, while TPRs were mainly found in the *Mimiviridae* family, KELCH in *Phycondaviridae* and *Poxviridae* families, and RCC1 only in two proteomes belonging to the *Phycondaviridae* family. For the species of

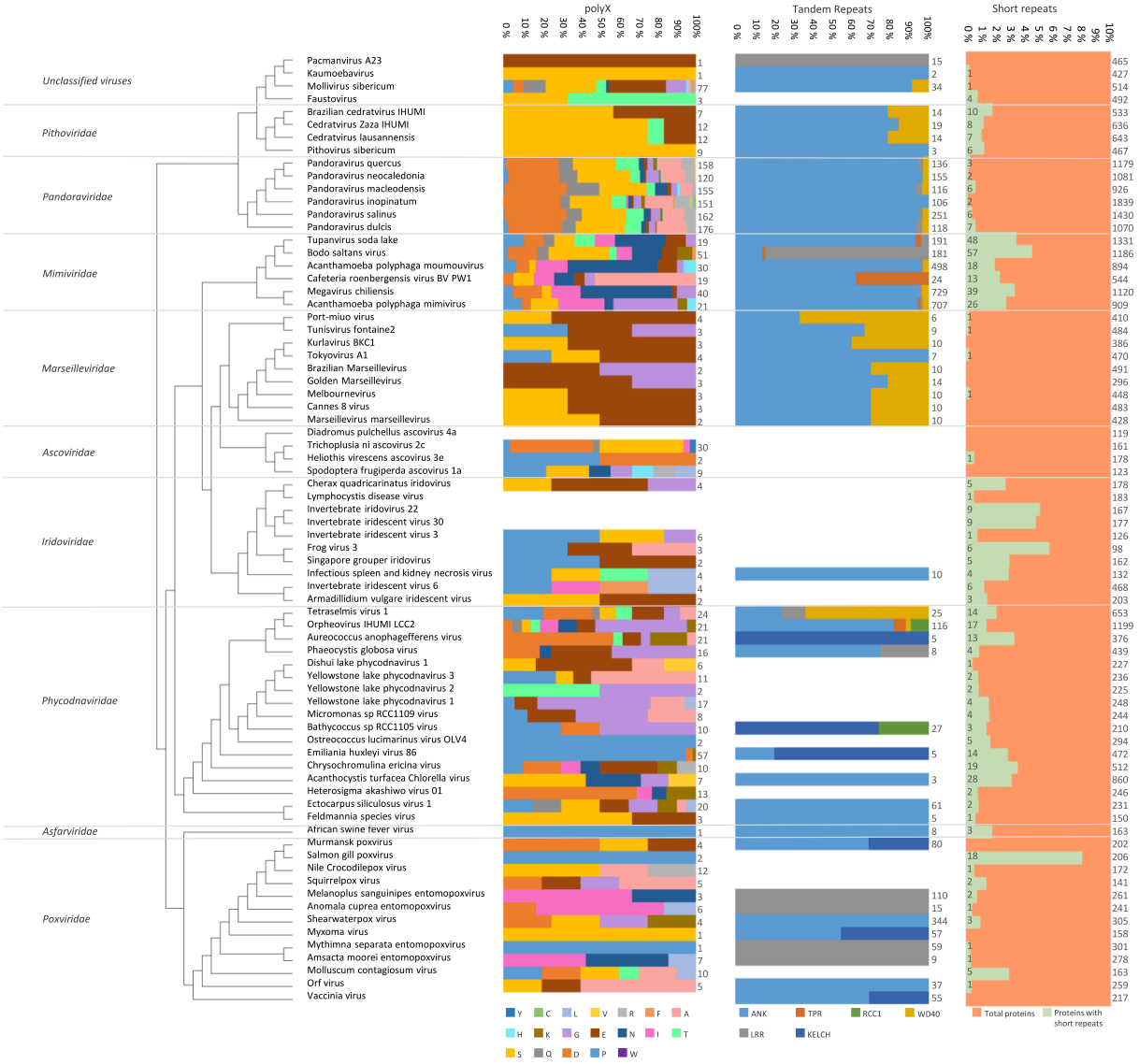

Fig. 1. Composition of proteins with repeats in 74 NCLDVs. The columns indicate, from left to right, families, phylogenetic tree of the 74 viruses (see Methods for details), virus names, percentages of polyX types found (blank for none found) with a number indicating the number of polyX found (more than one per protein is possible), percentages of types of tandem repeats found (blank for none found) with a number indicating the number of tandem repeats found, and fraction of proteins in the proteome found to have short repeats (scale is 0 % to 10 %; number of proteins indicated over the green bar), with a number (right-end) indicating the total number of proteins in the proteome. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Tools for repeat detection used in this work.

Tool	Repeat type	URL	Parameters	Reference
REP2	TRs	https://cbdm-01.zdv.uni-mainz.de/~munoz/rep/	default	(Kamel, Kastano et al. 2021)
RES	SRs	https://cbdm-01.zdv.uni-mainz.de/~munoz/res/	window size = 20 repeat length = 2 to 10 mutations = 0	(Kamel, Mier et al. 2019)
polyX2	polyX	https://cbdm-01.zdv.uni-mainz.de/~munoz/polyx2/	window length = 10 min identical residues = 8	(Mier and Andrade-Navarro 2022)

the *Ascoviridae* family, no TRs were found. While the absence of PFTA/PFTB repeats is not too surprising because they are associated with protein prenylation (Andrade et al., 2000), which is not performed by viruses, the absence of ARM and HEAT repeats, which form alpha-solenoids, is intriguing.

In the search for polyX, the ones more frequently found were polyD and polyS; no polyM or polyW were found in the proteomes analysed, which is consistent with their generally low frequency in all taxa (Mier et al., 2017). PolyS has been observed to be one of the most frequent homorepeats in Eukarya (Alba et al., 2007), but the high frequency of polyD differentiates giant viruses from all other taxa. *Pandoravirus* are responsible for these numbers, since each of them has more than 100 proteins with homorepeats (above 10 % of their proteomes). Four of the 74 proteomes analysed did not show any polyX, three of them from the *Iridoviridae* family and another from the *Ascoviridae* family (see Fig. 1 and Suppl. Table S1).

Regarding the proteins with SRs, obtained with RES, it is interesting to note that SR content does not correlate with polyX content at the proteome level. For example, while *Pandovirus* have high polyX frequency, they have low SR content, whereas *Mimiviridae* has the highest number of proteins with these repeats and have very few polyX. Among the proteomes of the *Iridoviridae* and *Marseilleviridae* families, the presence of proteins with TRs is almost nil.

We next addressed the question of whether the results obtained are shared among the members of a virus family (as described in Fig. 1). In *Pithoviridae* and *Pandoraviridae* families, we can see that the composition of polyX, TRs and SRs is mainly conserved. Composition of TRs is conserved in *Mimiviridae, Marselleviridae, Ascoviridae* and *Iridoviridae*. Viruses of the *Mimiviridae* family present mainly Ankyrin repeats, except for *Bodo saltans virus,* which presents mainly LRRs. In the *Marselleviridae* family, viral proteins have Ankyrin and WD40 repeats, except for *Tokyovirus A1*, whose proteins only have Ankyrin repeats. Proteomes analysed of *Ascoviridae* and *Iridoviridae* families do not have TRs, except for *Infectious spleen and kidney necrosis virus* (ISKNV), which have proteins with Ankyrin repeats.

Large differences can be seen between taxonomically closely related viruses. For example, *Phycodnaviridae*, *Tetraselmis virus 1* and *Orpheovirus IHUMI LCC2* are closely related but *Orpheovirus* has many more polyG proteins and a much higher frequency of proteins with TRs, particularly with Ankyrin repeats. This virus has a proteome twice as large as *Tetraselmis*, and also infects protozoa *Vermamoeba vermiformis*, an amoeba known for its gene exchange with giant viruses like the *Bodo saltans virus* (Chelkha et al., 2020). We observe a similar anomaly between two very related *Mimiviridae*: *Bodo saltans virus* and *Tupanvirus soda lake*, which infects protozoa *Acanthamoeba polyphaga*. In this case, their proteomes are similar in size, but *Bodo saltans virus* has more than twice as many proteins with polyX as *Tupanvirus*.

2.2. Comparison between different types of repeats in each proteome

The previous results suggest great variation in repeat content between highly related viruses and lack of correlation between repeat types: tandem, short or homorepeats. To investigate this in more detail, we compared the presence of different repeats in each proteome (Fig. 2). We found that a high incidence of one type of repeats often implies a low or nil incidence of the other two types. Viruses that showed unusual high quantities of repeats were further analysed: *Bodo saltans virus*, (orange in Fig. 2), with 1448 amino acids in short repeats, the six different species of the genus *Pandoravirus* with more than one hundred homorepeats in their proteomes (blue in Fig. 2), and *Acanthamoeba polyphaga mimivirus* and *Megavirus chilensis*, with 729 and 707 tandem repeats respectively

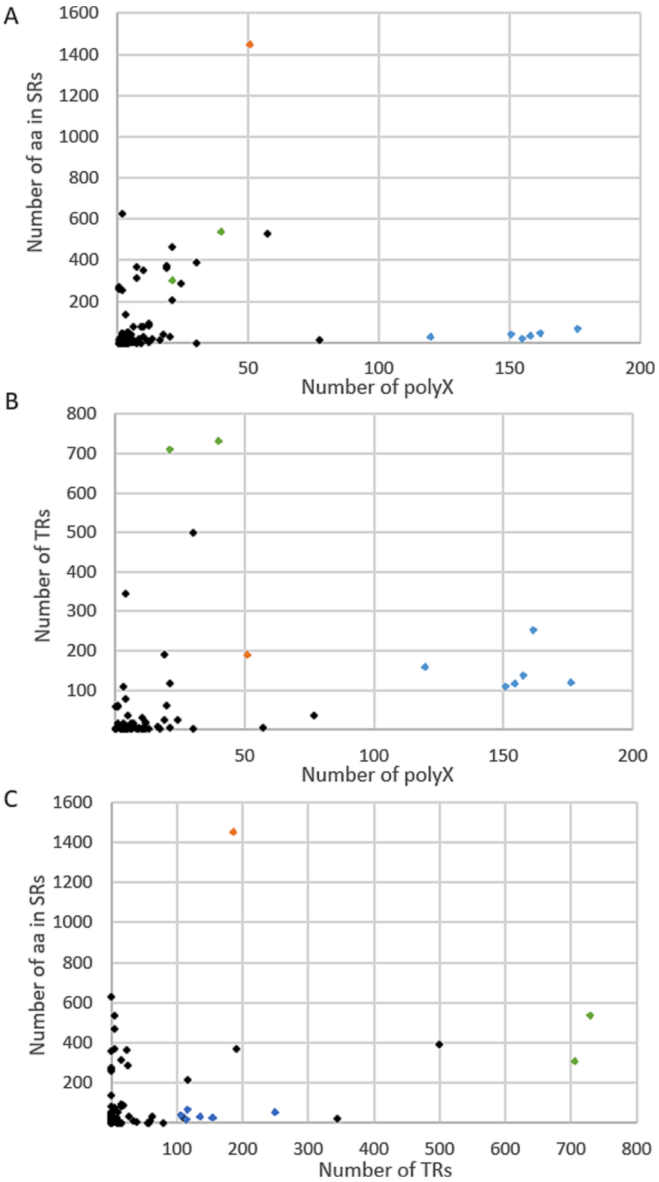

Fig. 2. Relation between numbers of repeats found in 74 NCLDVs. (A) Number of amino acids in SRs versus polyX. (B) TRs versus polyX. (C) Number of amino acids in SRs versus TRs. Pandoraviruses are highlighted in blue, Mimiviridae *Acanthamoeba polyphaga mimivirus* and *Megavirus chilensis* in green, and *Bodo saltans virus* in orange. Note that to represent abundance, we considered number of units in TRs and polyX, and number of amino acids in SRs because regions with TRs and polyX are more homogeneous in length than regions in SRs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(green in Fig. 2). We discuss the repeat containing proteins of these viruses in the following sections.

2.3. Short repeats in Bodo saltans virus

The proteome of *Bodo saltans virus* was the one that contained the highest number of proteins with SRs (Fig. 1), of the 74 proteomes analysed. This virus of the *Mimiviridae* family infects *Bodo saltans*, a free-living bacterivorous protist found worldwide in marine and freshwater habitats (Opperdoes et al., 2016). Its proteome of 1186 proteins contains 1448 amino acids in SRs (covering 0.4 % of its proteome), all within just 57 proteins. To evaluate the origin of these repeat containing proteins, we looked for homologs in other organisms excluding viruses (Supplementary Table S3).

Several proteins in the *Bodo saltans virus* proteome have over one hundred residues detected to be in perfect SRs. An extreme case is the 320 amino acid protein UniProt:A0A2H4UVV3, which is composed almost entirely of a total of 74 perfect repeats with the sequence "NYID" (positions 18–313). We did not find a homolog for this protein.

Among the proteins with homologs, UniProt:A0A2H4UTS7 shows the domain organization of a nuclear export mediator factor (NEMF) protein, a protein found in all domains of life (for example, the human homolog has 25 % identity to the *Bodo saltans virus* sequence – BLAST E-value = 1e-25). The closest homolog (NCBI:MBA43143.1) is from a bacteria of the Magnetococcales order (32 % identity – BLAST E-value = 3e-83). None of the homologs displayed the SRs (Fig. 3A), an insertion in the viral protein corresponding to the end of a coiled coil region. Interestingly, we found another four cases of virus SR proteins where the closest homolog was also from *Magnetococcales bacterium*. These bacteria are mainly isolated from freshwater sediments. We can hypothesise that horizontal gene transfer in the host or in its environment led to the presence of these proteins. A high level of horizontal gene transfer for giant viruses has already been well described (Filee et al., 2008).

Sequence similarity is not always indicative of homology and in some sequence comparisons, similarity was restricted to the SRs, which can trigger very significant (low) E-values in BLAST searches. This was the case for the 199 amino acid protein UniProt:A0A2H4UUT6, which has nine perfect repeats of eight residues "KETIAETP"; visual inspection indicates three more repeats at the start and the end of the ensemble (positions 73–168; Fig. 3B). A similar protein was found (NCBI: KAH9106337.1) in *Aphanomyces euteiches*, an oomycete (eukaryotic, water mould), which is a plant pathogen. The similarity however was restricted to repeats of sequence "AETP" and could be spurious. The AlphaFold structure model (Jumper et al., 2021) of the viral protein suggests that terminals of the repeats region overlap alpha-helices, whereas the central repeats are not structured (Fig. 3B). It was interesting to note that, like in the previous example, the sides of the repeats could be bordering helical conformation. But this is not the case in

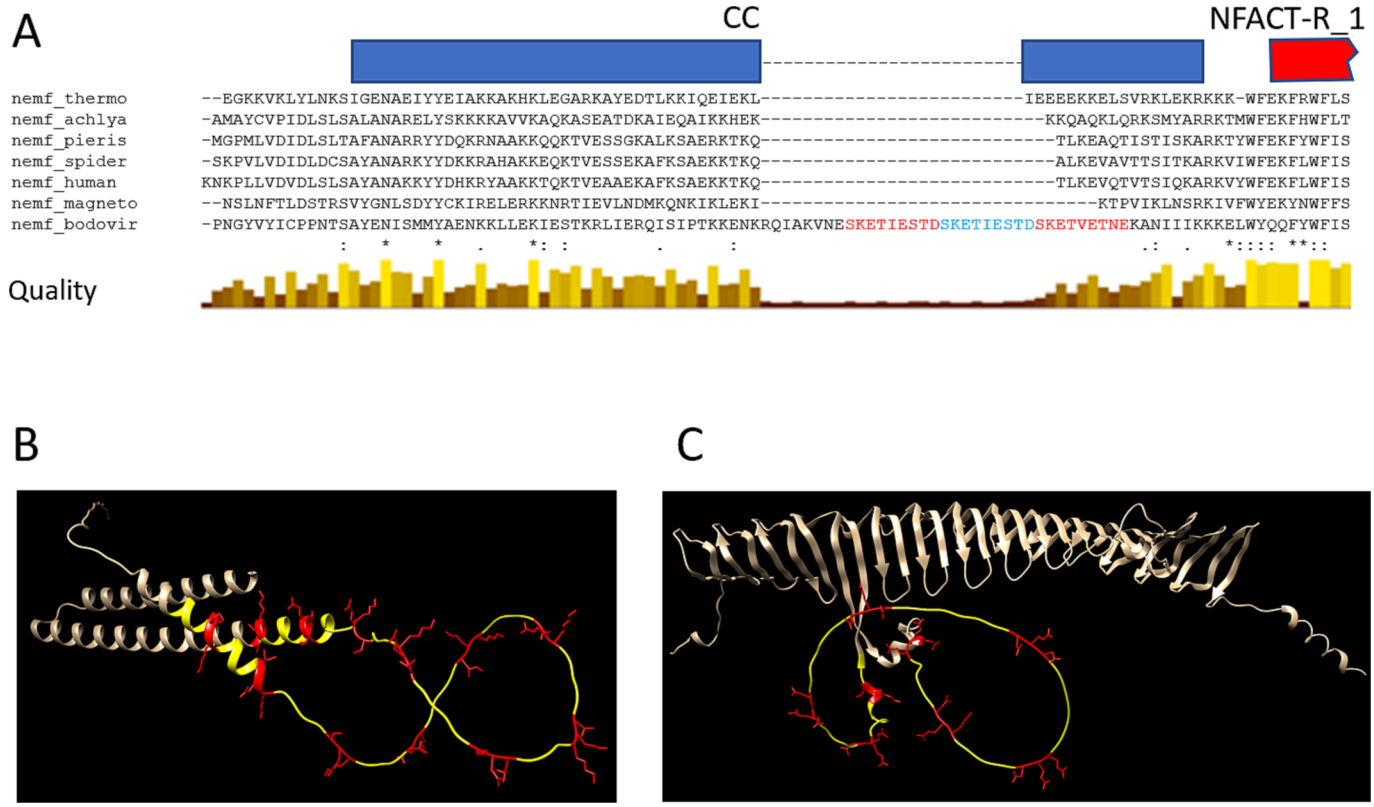

Fig. 3. Proteins with SRs in *Bodo saltans virus*. (A) Fragment of a multiple sequence alignment (MSA) of *Bodo saltans virus* NEMF protein (NCBI:ATZ80225.1, nemf_bodovir in the Figure) with homologs from (top to bottom) *Thermococcaceae archaeon* (NCBI:HII68122.1, nemf_thermo), *Achlya hypogyna* (water mould; NCBI: OQR81173.1, nemf_achlya), *Pieris brassicae* (a butterfly; NCBI:XP_045532918.1, nemf_pieris), *Trichonephila inaurata madagascariensis* (a spider; NCBI:GFY76475.1, nemf_spider) Homo sapiens (NCBI:NP_004704.3, nemf_human) and *Magnetococcales bacterium* (NCBI:MBA43143.1, nemf_magneto). Boxes represent a coiled coil in the human sequence (blue) and the start of the NFACT-R_1 domain (red). Three SRs are coloured in the *Bodo saltans virus* sequence. Conserved positions are indicated with asterisks, colon or dot, indicating full conservation, and two lower levels of decreased conservation, respectively. Column reliability is indicated using the Quality measure computed by Jalview (Waterhouse et al., 2009). (B) AlphaFold model of *Bodo saltans virus* sequence A0A2H4UUT6. The repeat region (73–168) is marked in yellow with the three first residues of each repeat (consensus "KET") represented with side chains in red. The average pLDDT score for this region is 35.4. (C) AlphaFold model of *Bodo saltans virus* sequence A0A2H4UW39. Most of the protein is composed of an ensemble of 13 MORN beta-hairpin tandem repeats. A region with the eight SRs (120–186) is inserted between two MORN repeats. This region is marked in yellow with the three first residues of each repeat ("DQQ") represented with side chains in red. The average pLDDT score for this region is 35.6. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

general. For example, the 474-residue protein UniProt:A0A2H4UW39 is predicted to contain 13 Membrane Occupation and Recognition Nexus (MORN) tandem repeats. In the AlphaFold predicted structure, each of these repeats forms a beta hairpin and together they assemble as an elongated beta-sheet (in agreement with experimentally obtained structures (Sajko et al., 2020)). In this case, the region of SRs (positions 120–186) is inserted between two regions of 3 and 10 MORN repeats (positions 18–79 and 193–441) and predicted mostly to be unstructured (see Fig. 3C), save for some small helices. The MORN domain was the most frequently observed domain in this set of viral proteins with SRs. Although it has no known function, in the parasite *Toxoplasma gondii* it has been hypothesised to function as a linker protein between certain membrane regions and the cytoskeleton (Takeshima et al., 2000). These proteins might be doing a similar structural function for the virus.

2.4. Tandem repeats in Acanthamoeba polyphaga mimivirus

The proteome of *Acanthamoeba polyphaga mimivirus* was the one for which we detected the second highest number of instances of tandem repeats, second (but close) to *Megavirus chilensis*. Both viruses belong to the *Mimiviridae* family and infect the free-living amoebae *Acanthamoeba polyphaga* sp. The genome of *Acanthamoeba polyphaga mimivirus* codes 909 proteins, for which we detected 707 tandem repeats contained in 101 proteins (729 TRs in 118 proteins from *M. chilensis*), with a huge majority of Ankyrin repeats (in 93 proteins), and a few proteins with other TRs (WD40, TPR and LRR, in 5, 2 and 1 proteins, respectively).

Homologs for the repeat containing proteins were found in non-viral organisms like *Smittum angustum*, *Furculomyces boomerangus*, *Bordetella hinzii* and *Epsilonproteobacteria bacterium* (Supplementary Table S4). The fungi *S. angustum* and *F. boomerangus* have organisms of the *Tanytarsus* genus as hosts, but no information relates *Tanytarsus* with *Acanthamoeba polyphaga mimivirus*. On the other hand, *Bordetella hinzii* and *Epsilonproteobacteria bacterium* are found in diverse hosts, such as *Homo sapiens*, which has been also described as a possible host for *Acanthamoeba polyphaga mimivirus* and for *Acanthamoeba polyphaga* (Abrahao et al., 2014).

While Ankyrin repeat proteins could have been obtained via horizontal transfer from these organisms, we wanted to assess if Ankyrin repeats can duplicate within the viral lineage. A maximum number of 18 Ankyrin repeats was identified for UniProt:Q5UQI8. We investigated the

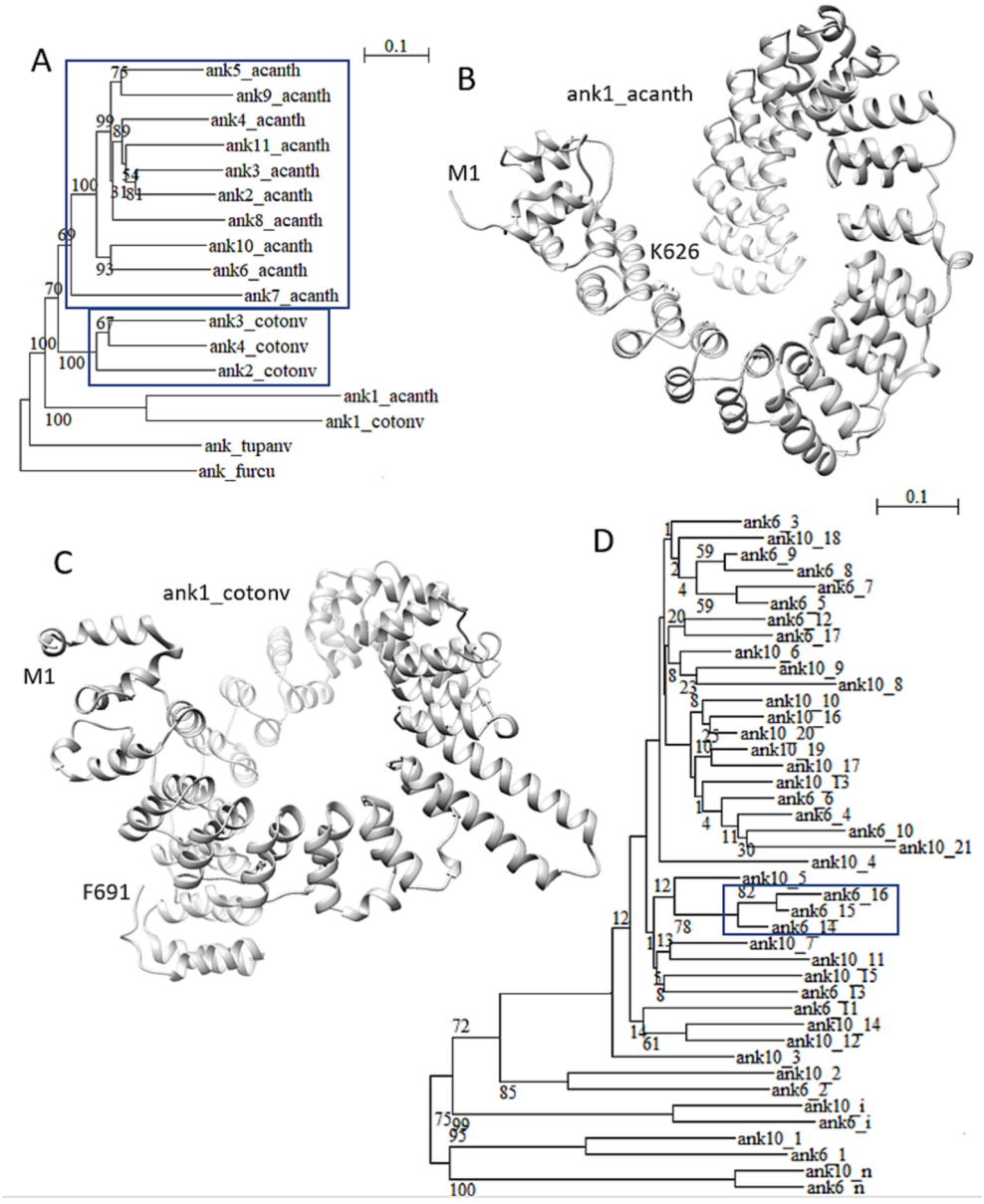

Fig. 4. Ankyrin repeats in *Acanthamoeba polyphaga mimivirus* sequence Q5UQI8 and homologs. (A) Phylogenetic tree of full sequences of Ankyrin repeats proteins Q5UQI8 with homologs in mimivirus: *Cotonvirus japonicus*, *Tupanvirus soda lake*, and with closest non-viral homolog from fungi *Furculomyces boomerangus* (used as outlier for this unrooted tree, ank_furcu). Two branches (marked with boxes) with multiple sequences from *A. polyphaga* (*_acanth) and from *C. japonicus* (*_cotonv), respectively, suggest, the independent expansion of this family in different virus. Numbers in the branches indicate branch support from 100 bootstraps. The tree was obtained excluding positions with gaps. (B) AlphaFold model of UniProt:Q5UQI8 (ank1_acanth in A). Labels indicate the most N- and C-terminal residues. Average pLDDT score = 84.0. (C) AlphaFold model of NCBI: BCS82623.1 (ank1_cotonv in A). Labels indicate the most N- and C-terminal residues. Average pLDDT score = 82.1. (D) Phylogenetic tree of individual Ankyrin repeats and other sequence fragments from two *Acanthamoeba polyphaga mimivirus* homologs of UniProt:Q5UQI8, ank6 and ank10 (NCBI: AKI79744.1 and NCBI:AKI80215.1, respectively). The labels indicate the repeat position in the corresponding protein (e.g. ank6_16 is the 16th repeat in ank6), correspondence to the N-terminal (ank6_n and ank10_n), or to an insertion after the first repeat (ank6_i and ank10_i). A branch indicating repetitions of contiguous repeats (ank6_14, ank6_15 and ank6_16) is marked with a box. Numbers in the branches indicate branch support from 100 bootstraps.

origin and evolution of Ankyrin repeats in mimivirus using this protein as an example.

We obtained homologs of this protein from other viruses and identified the closest non-viral homolog (from Harpellales fungus *Furculomyces boomerangus*). In all, we collected 11 homologs from *Acanthamoeba polyphaga mimivirus*, and five from other mimivirus (one from *Tupanvirus soda lake*, and four from *Cotonvirus japonicus*). The phylogenetic tree built from the multiple sequence alignment of these sequences (Fig. 4A) suggests the expansion of this family within the viral taxa from a common ancestral sequence. We hypothesize that it could have been transferred from an ancestral fungal sequence given its similarity to the *F. boomerangus* sequence (see Fig. 4A) and to other homologs in fungi (from *Smittium angustum* and *S. culicis*, which are also of order Harpellales). All sequences seem to be mostly composed of Ankyrin repeats except for short N- and C-terminal regions. Two AlphaFold models (Jumper et al., 2021) of UniProt:Q5UQI8 (ank1_acanth in Fig. 4A) and the close homolog in *C. japonicus* NCBI:BCS82623.1 (ank1_cotonv in Fig. 4A) support this (Fig. 4B-C). The complexity of the evolution of these proteins is high.

To illustrate this point, we chose two close homologs from *A. polyphaga mimivirus* NCBI:AKI79744.1 (ank6_acanth in Fig. 4A) and NCBI:AKI80215.1 (ank10_acanth in Fig. 4A), which occupy their own branch in the phylogenetic tree of the family (Fig. 4A) and made a multiple sequence alignment of fragments of their sequences considering individual repeats and insertions. Although they have a similar number of repeats, and share homology between their N-termini (ank6_n and ank10_n in Fig. 4D) and at an insertion after the first repeat (ank6_i and ank10_i in Fig. 4D), the similarity between the repeats is complex. For example, one can see tight branches of repeats from one of the sequences (like ank6_14, ank6_15 and ank6_16; box in Fig. 4D) with high sequence identity (the last two differ by only two residues), hinting at very recent events of tandem repeat duplication within *A. polyphaga mimivirus*. Taken together, our results indicate that mimivirus has taken Ankyrin repeat proteins by horizontal transfer and that events of gene duplication and tandem repeat duplication happen very actively within their lineage.

To study the complete set of proteins with Ankyrin repeats in *A. polyphaga mimivirus*, and to see if there are different subfamilies using variants of the repeat, we tried a different annotation system that considers multiple profiles for Ankyrin repeats. We used the PFAM domain database profile searches for Ank, Ank_2, Ank_3, Ank_4 and Ank_5 (Mistry et al., 2021). We found 79 sequences that matched some of these profiles, but mostly one of them (Ank_2). We take this result as indicative that all repeats belong to a similar type of ankyrin repeat.

From each of these sequences, we extracted the three consecutive repeats with the strongest HMM score and constructed a multiple sequence alignment (sequence identifiers and amino acid positions are indicated in Figure S1). The identity between the set of three-repeat sequences ranged from higher values around 79 % (e.g. between UniProt:YL483_MIMIV and UniProt:YR911_MIMIV) to lower values like 19 % (e.g. between UniProt:YR096_MIMIV and UniProt:YR580_MIMIV). We were surprised not to find pairs of Ank repeats with higher identity; this suggests that they were acquired a long time ago and that they are not constrained by selection or might even diversify faster due to some adaptive pressures. The phylogenetic tree suggested a large group of 62 sequences, with separate branches for 11, 4 and 2 sequences (Figure S1). This analysis of the complete set of proteins with ankyrin repeats indicates large rates of gene duplication within this family. We did not see evidence of subfamilies of repeats.

Together, we take our results as an indication that proteins with ankyrin repeats in *A. polyphaga mimivirus* evolve by gene duplication and tandem repeat gain and loss with profusion, but we cannot appreciate subfamilies. We propose that the dynamics of the entire set responds to a similar functional requirement.

2.5. Homorepeats in the Pandoraviridae family

We found the highest content of homorepeats in the Pandoravirus proteomes. Those we have studied have between 926 and 1839 proteins, and we found around 160 homorepeats per proteome, distributed in a total amount of proteins between 97 and 128 proteins (data on Supplementary Table S1).

The viruses analysed from the *Pandoraviridae* family infect *Acanthamoeba* sp. The host for *Pandoravirus dulcis* and *Pandoravirus salinus* is *Acanthamoeba castellanii,* a pathogen species that causes different diseases (Soto-Arredondo et al., 2014); we found homologs of Pandoravirus proteins with homorepeats in this species (Supplementary Table S5).

The levels of functional annotation of the proteins of giant viruses are very low (Brandes and Linial 2019), and accordingly most of these proteins with homorepeats have no known function. However, we observed some with nucleic acid binding, ATP-binding, kinase, hydrolase, and helicase activity, as deduced from their content in predicted domains: PFAM domains P-kinase, PK_Tyr_Ser_Thr, MORN and Collagen were observed.

To assess the significance of the enrichment of functional annotations in these and in other proteins with repeats in giant viruses, we performed a functional enrichment analysis, which is described in the next section.

2.6. Functional enrichment analysis of proteins with repeats in giant viruses

To better understand the function of repeats in giant viruses, we performed a Gene Ontology (GO) over-representation analysis. Subsets of proteins were built for each of the 74 studied proteomes depending on whether they contain TRs, SRs or homorepeats. These subsets were then compared to their respective proteomes to elucidate if the presence of different repeats was associated with enriched molecular functions, biological processes, or cellular components.

Only proteins containing homorepeats showed significant results, either taking the complete set of homorepeat proteins, or when considering particular types of homorepeats (Table 2; detailed terms and protein names are available in Suppl. Table S6). No enrichment was observed for sets of proteins with TRs or SRs.

The presence of homorepeats was significantly associated with binding functions and with several metabolic biological processes in three of the six Pandoravirus species (*P. dulcis*, *P. salinus* and *P. neocaledonia*; Fig. 5A).

The individual study of different homorepeats revealed enriched GO-

Table 2
Gene Ontology enrichment analysis of viral proteins with homorepeats.

Class	Repeat	Ontology	GO terms	Proteins	Organisms
PolyX	All	MF, BP	48	69	*Pandoravirus dulcis* *Pandoravirus salinus* *Pandoravirus neocaledonia*
PolyX	PolyD	MF	5	8	*Pandoravirus salinus*
PolyX	PolyG	CC	4	6	*Mollivirus sibericum*
PolyX	PolyI	CC	9	15	*Acanthamoeba polyphaga mimivirus* *Acanthamoeba polyphaga moumouvirus* *Megavirus chilensis*
PolyX	PolyP	CC	4	11	*Emiliania huxleyi virus 86*
PolyX	PolyQ	BP	3	3	*Pandoravirus salinus*
PolyX	PolyS	MF	51	48	*Pandoravirus dulcis* *Pandoravirus salinus* *Pandoravirus neocaledonia* *Pandoravirus quercus*
PolyX	PolyT	MF	1	2	*Pandoravirus salinus*

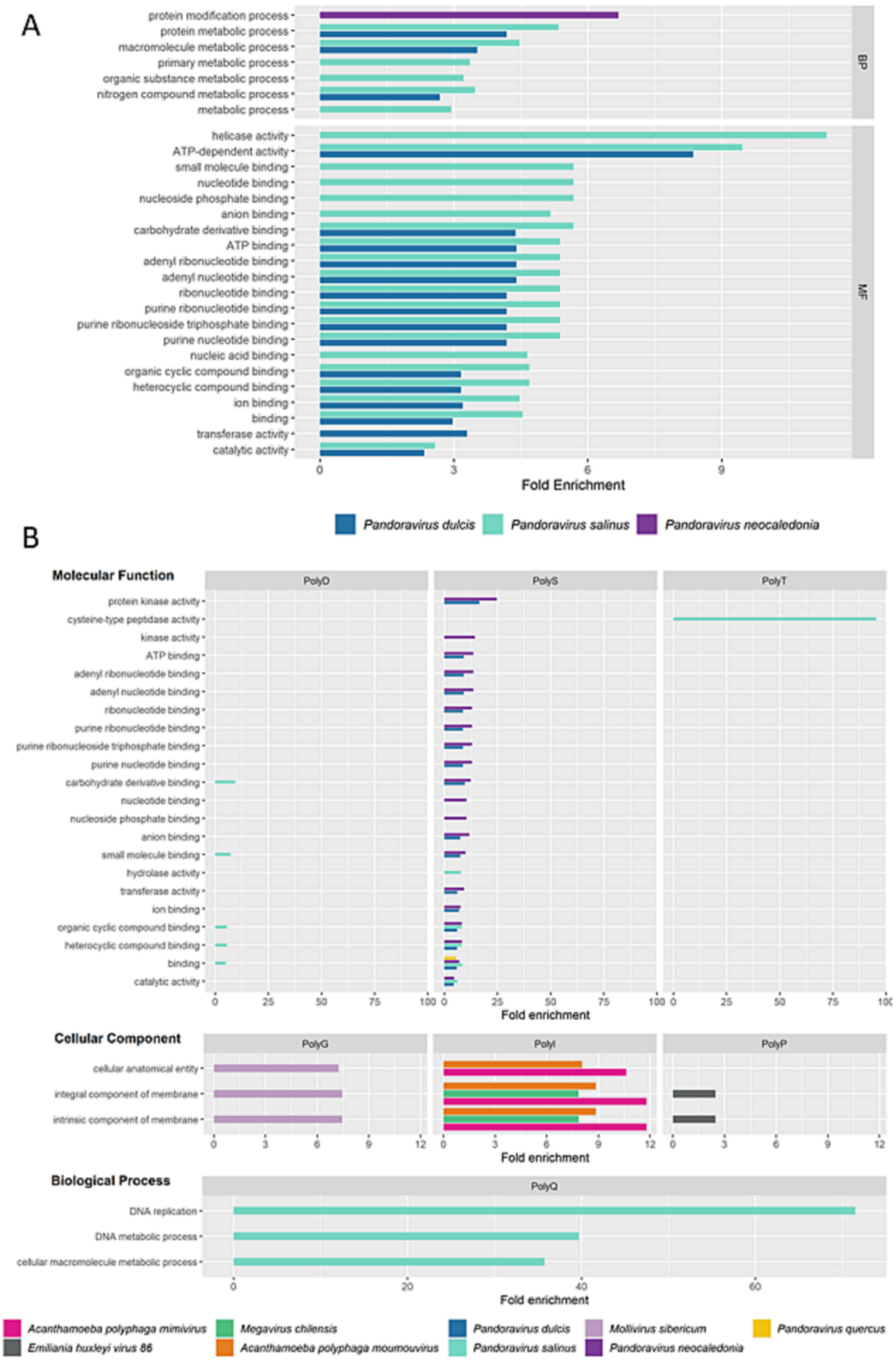

Fig. 5. Enriched GO terms in proteins with homorepeats. (A) Fold enrichment is displayed for GO terms in proteins with homorepeats from three *Pandoravirus* species (p-value < 0.05). (B) Fold enrichment is displayed for GO terms in proteins with homorepeats by type for nine species.

terms for nine different organisms. Isoleucine, proline and glycine homorepeats are associated with cellular component of membranes; glutamine repeats with DNA replication biological processes; serine, threonine and aspartic acid with several molecular functions (Fig. 5B; detailed terms and protein names are available in Suppl. Table S6).

For polyI, three organisms from the Mimiviridae family (*Acanthamoeba polyphaga mimivirus*, *Acanthamoeba polyphaga moumouvirus* and *Megavirus chilensis*) were found to have significantly enriched Cellular Component annotations (membrane location related), and are present in 15 of the 16 polyI-containing proteins in these organisms (Suppl. Table S7). A closer look at the sequences of these 15 proteins showed that the isoleucine homorepeat makes up for a large part of a predicted transmembrane (TM) helix that would anchor the protein to the viral membranes. These results agree with previous observations (Mier et al., 2017).

Isoleucine homorepeats have an uneven taxonomic distribution; their presence in viral proteins is usual, but this is not the case for any of the other taxonomic domains (Mier et al., 2017). Alignments of these proteins to homologs show an abrupt difference between viral and non-viral homologs: when aligned, the non-viral homologs present a set of hydrophobic amino acids without a significant content of isoleucine, whereas viral homologs show higher proportions of isoleucine (identified as a polyI when more than eight Isoleucines are found in a window of ten residues; Figure S2). These results suggest that, at least in the viral species mentioned, there is selection pressure to generate polyI regions within TMs, a feature that is unique to viral species. The process of generation of polyI can be by substitution (Fig. S2A), but we also observed an example where the TM-containing polyI is appended to the protein C-terminal in the viral lineage (Fig. S2B). Multiplicity of means of generation of a polyX (particularly, substitution versus insertion) has been observed and discussed for polyQ (Mier and Andrade-Navarro 2020).

We only observed three proteins with polyQ, all of them uncharacterized proteins from *Pandoravirus salinus*. Interestingly, in UniProt: S4W0W6 the polyQ was conserved in multiple viral homologs but non-viral homologs were not found (Fig. 6A); this was followed by a P-rich region, which is characteristic of many eukaryotic polyQ (Schaefer et al., 2012), hinting that it could perform a similar function in the modulation of a coiled coil protein–protein interaction. For ribonucleoside-diphosphate reductase (UniProt:S4W654, 1820 residues long), we found shorter non-viral homologs (e.g. ribonucleotide reductase from *Tribonema minus*, 821 residues long), which share a number of common domains in the same order but lack the viral insertions. In this case, the N-terminal extension of S4W654 and close homologs contains a variable polyQ region (Fig. 6B). The third case is DNA polymerase UniProt: A0A291ATN4, which displays an internal polyQ in a region that is not shared with any viral or non-viral homolog (not shown).

We note that the glutamine residues of the polyQ in S4W0W6 and S4W654 are entirely coded by CAA codons, whereas for A0A291ATN4 the majority of codons are CAG and not CAA (7 and 2, respectively) (Fig. 6C). The homogeneous use of one codon suggests mechanisms of slippage producing these homorepeats, but use of CAG or CAA seems to be possible.

We observed 48 proteins with GO terms enriched in four species of the Pandoravirus family (Suppl. Table S8). Many of them are annotated as Serine/Threonine protein kinases. Some of these proteins have very high sequence similarity and align full length to a protein from the host, *Acanthamoeba* (Suppl. Table S5), which suggests recent horizontal transfer. In *Pandoravirus dulcis*, we find one example of a large multi-domain protein, UniProt:S4VV57 (1921 amino acids long), which includes a protein kinase domain (InterPro domain:IPR000719 Protein kinase domain) followed by a DNA cyclase domain (InterPro homologous superfamily:IPR029787). Between them there is a short disordered region (according to UniProt annotations) with two polyS (with 9 serines in a window of 11, and with eight consecutive serines, respectively). This protein has a full-length homolog in *Acanthamoeba castellanii* (NCBI:XP_004334716.1), with 85 % coverage and 35 % identity (E-value 5.00E-139) of length 1682 amino acids and similar domain distribution. However, this disordered region between the kinase and cyclase domains is much shorter and lacks the polyS in the host protein.

In contrast, another protein with a protein kinase domain, UniProt: A0A291AU60 (664 amino acids long), has its best homolog in *Actinomycetia bacterium* at 28 % identity NCBI:MBC8390743.1, with a significant E-value (1.00E-13), but this is a much shorter protein (237 amino acids). Their inferred homology is due to having a conserved domain (InterPro domain:IPR000719 Protein kinase domain) that covers practically the entire bacterial homolog. The N-terminal part of the viral protein (containing a polyS of 24 amino acids with 20 serines) does not exist in the shorter bacterial homolog. This suggests that this family might have emerged within the viral lineage. The over-representation of protein kinase function in polyS containing proteins from Pandoravirus could indicate that polyS has a viral-specific function, which is not originating in proteins of the host. As far as we know, polyS has not been described in association to protein kinases in non-viral species or in general.

Our observations suggest that polyX can emerge within viral lineages, either by substitution or by insertion of new sequences, and that it might be associated with particular protein structures and functions.

3. Conclusion

We have profiled the proteomes of a collection of giant viruses by their content of different types of protein repeats: homorepeats, short tandem repeats, and longer structured protein repeats. We found almost all types of repeats in the families analysed, and for many of them we found cases where sequence evidence indicates that the repeat emerged and evolved within the viral lineage. While we found some consistency between TR content and phylogeny (e.g., see the similar distributions of polyX types within *Mimiviridae*, or the almost total absence of proteins with SRs in *Marseilleviridae*; Fig. 1), there are also outliers (e.g., *Bodo saltans virus* with more than twice the number of amino acids in short repeats than any other of the giant virus species analysed in this work; Fig. 2A).

The low level of functional annotation of proteins in giant viruses (Brandes and Linial 2019) has likely reduced our chances of associating particular types of repeats with functions. For example, we can only point to the high frequency of short repeats in 57 proteins of the proteome of *Bodo saltans virus* (1186 proteins), with an average of 25 repeats per protein, without being able to say if these proteins have a particular functional requirement to have these repeats or if their presence is arbitrary. Regardless, automated domain and feature annotations allowed us to find the association of homorepeats polyI and polyS with transmembrane proteins in the Mimiviridae family, and to protein kinases in the Pandoravirus family, respectively. These results give us hope that with future increase in annotations and virus sequencing further associations might be found.

It has been noted that the Pandoravirus family holds a number of genes that cannot be explained by horizontal transfer, suggesting that many of their genes emerged within the viral lineage (Legendre et al., 2018). Our results confirm this trend in other families of giant viruses and expand our understanding of the genetic mechanisms employed by viruses to gain functionality: most types of repeats found in eukaryotic and prokaryotic organisms emerge and evolve in giant viruses. These results agree with the observed lack of association between viruses and their host organisms in terms of tandem repeats content (Delucchi et al., 2020), and contribute to the idea that giant viruses have evolved from simpler genomes by addition of (in our case, duplicated) genetic material (Filee et al., 2008).

Giant viruses have their own way of adopting or generating proteins with tandem repeats, sometimes independently of non-viral organisms, leading to a diversity of proteins with tandem repeats with a rich heterogeneity. While certainly not all types of tandem repeats and

C

```
S4W0W6
595   G   P   Q   Q   Q   Q   Q   Q   Q   Q   K   P   606
1783  ggc ccg CAA CAA CAA CAA CAA CAA CAA CAA aag cct 1818

S4W654
4   L   R   Q   Q   Q   Q   Q   Q   R   E   Q   Q   T   G   17
10  ctt cga CAA CAA CAA CAA CAA CAA cgg gaa CAA CAA acg ggc 51

A0A291ATN4
432  D   A   Q   Q   Q   Q   Q   Q   Q   E   Q   Q   P   R   445
1294 gac gct CAG CAA CAG CAG CAG CAG CAA GAA CAG CAG ccg cgt 1335
```

Fig. 6. PolyQ regions in viral proteins aligned to homologs. (A) The MSA includes an internal polyQ from an uncharacterized protein (UniProt:S4W0W6) from *Pandoravirus salinus*. All proteins are viral. (B) The MSA includes the N-terminal polyQ from ribonucleoside-diphosphate reductase (UniProt:S4W654) from *Pandoravirus salinus*. The viral homologs are marked with a grey box. The numbering in the alignment views (using JalView; (Waterhouse et al., 2009)) is relative to the lead sequence. Column reliability is indicated using the Quality measure computed by Jalview. (C) Protein sequence and corresponding codons for the proteins and positions indicated (UniProtKB identifiers). Glutamine codons are represented with capital letters.

homorepeats are found in giant viruses, many non-viral taxa show also similar specificities. In this respect, we cannot say that giant viruses are more or less restricted in their capacity to use tandem repeats than non-viral organisms. In fact, several of the giant viruses analysed exceed the repeat content of many non-viral organisms. For example, *Megavirus chilensis* has more ANK repeats (693 units) than a sample of 27 prokaryotic genomes we previously analysed (Kamel et al., 2021), and more than the majority of eukaryotic species analysed in that work (31 out of 51 eukaryotic genomes).

Particular specific trends exist, like the presence of polyI in the Mimiviridae family, and the absence of HEAT and ARM TRs, which remains a puzzle in the light of the abundance of other alpha-solenoid forming TRs such as Ankyrin and LRRs. The sequencing of new viral genomes might confirm these observations, and suggest other particularities. Re-sequencing might also help confirming that none of these repeats could be due to repeats in DNA, which can cause assembly errors (Torresen et al., 2019).

Our study employed three tools, each profiling a different type of repeats (SRs, TRs and polyX). Using other tools will complete our results and could confirm or refute some of the biases we find. We believe, however, that our analysis already demonstrates that giant viruses have access to a highly dynamic set of genetic mechanisms that make them comparable to non-viral species in their ability to accumulate repeated sequences in proteins. From a research perspective, this is good news as giant viruses emerge as a new set of very simple model organisms to study the mechanisms producing gene duplication and modifying gene length by expansion and contraction of tandem repeats. It would be very interesting to study how TRs change within a viral population, particularly if these were involved in proteins related to pathological functions.

We believe that our results motivate the discovery and sequencing of more families and members of this type of virus, which might shed light on the mechanisms that expand the genetic material.

4. Materials and Methods

The proteomes of 74 selected species of Nucleocytoplasmatic large DNA viruses (NCLDVs; reported in Fig. 1 from (Koonin and Yutin 2019)) were retrieved from the UniProtKB database. These collections of proteins were analysed for repeats. For this goal we made use of three bioinformatic tools.

The REP2 tool was employed to look for tandem repeats (TR) in the proteomes. This server allows the detection in amino acid sequences of eleven different TRs with known structures (Kamel et al., 2021): Ankyrin (ANK), Armadillo (ARM), HAT, HEAT, KELCH, LRR, PFTA, PFTB, RCC1, TPR and WD40, and three variations of the HEAT Tandem Repeat (HEAT_AAA, HEAT_ADB and HEAT_IMB). The tool was used via API access. No TRs of categories ARM or any of the HEAT variants were found. The RES server allows the detection of the repeatability in a protein sequence (Kamel et al., 2019). This tool was used setting restrictive parameters that led us to search for short repeats in the proteomes of interest. We set a window size of 20, with a repeat length of 2 to 10, and 0 mutations allowed, which report only perfect repeats. For searching homorepeats within the proteins in our dataset, we used the polyX2 tool with default parameters (window length of 10 amino acids, and 8 the minimum number of identical residues required in the window to consider it to be part of a homorepeat) (Mier and Andrade-Navarro 2022).

The phylogenetic tree in Fig. 1 was made with the NCBI tool Common Taxonomy Tree, and the Molecular Evolutionary Genetic Analysis 11 (MEGA11) software (Kumar et al., 2018) was employed to visualize and edit the tree obtained. The search for homologs was made employing BLASTp on the NCBI web page (Johnson et al., 2008), with default parameters and excluding viral taxa. To determine a protein as homolog, we set a threshold value of 50 % for coverage and 30 % for identity. The PFAM database was used for searching domains present in proteins of interest (Mistry et al., 2021). Multiple sequence alignment and phylogenetic trees were made using COBALT (Papadopoulos and Agarwala 2007) from the NCBI server (using default parameters; Figs. S1, S2A-B and 6A-B), or MUSCLE (Edgar 2004) from EBI server (Madeira et al., 2022) and Clustal Omega (Sievers et al., 2011) (using default parameters; Fig. 3A and 4). The fragments of MSAs with regions of low complexity shown in Figures S2 and 6 were taken from full sequence MSAs anchored in regions with normal composition. Gene Ontology enrichment analyses were computed using the Python library GOATOOLS (Klopfenstein et al., 2018) (using default parameters, Fisher's exact test with Bonferroni correction and significance p-value cut-off 0.05, and corresponding viral proteome as background).

AlphaFold (Jumper et al., 2021) (v2.2.1) was used with MSA databases: Uniref90 release March 2021, MGnify clusters release May 2019 and smallbfd (BFD's first non-consensus sequences) release 2019; running on Colab Pro, with a CPU Xeon 2.2 GHz and a GPU Tesla T4. Chimera was used for protein structure representation (Pettersen et al., 2004).

Funding

European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie [823886]. Funding for open access charge: REFRACT (Marie Skłodowska-Curie) [823886].

CRediT authorship contribution statement

Sofía Erdozain: Formal analysis, Writing – original draft, Writing – review & editing. **Emilia Barrionuevo:** Formal analysis, Writing – original draft, Writing – review & editing. **Lucas Ripoll:** Formal analysis, Writing – original draft, Writing – review & editing. **Pablo Mier:** Formal analysis, Writing – original draft, Writing – review & editing. **Miguel A. Andrade-Navarro:** Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jsb.2023.107962.

References

Abrahao, J.S., Dornas, F.P., Silva, L.C., Almeida, G.M., Boratto, P.V., Colson, P., La Scola, B., Kroon, E.G., 2014. Acanthamoeba polyphaga mimivirus and other giant viruses: an open field to outstanding discoveries. Virol. J. 11, 120.

Alba, M.M., Tompa, P., Veitia, R.A., 2007. Amino acid repeats and the structure and evolution of proteins. Genome Dyn. 3, 119–130.

Andrade, M.A., Perez-Iratxeta, C., Ponting, C.P., 2001. Protein repeats: structures, functions, and evolution. J. Struct. Biol. 134 (2–3), 117–131.

Andrade, M.A., Ponting, C.P., Gibson, T.J., Bork, P., 2000. Homology-based method for identification of protein repeats using statistical significance estimates. J. Mol. Biol. 298 (3), 521–537.

Brandes, N., Linial, M., 2019. Giant Viruses-Big Surprises. Viruses 11 (5).

Chelkha, N., Hasni, I., Louazani, A.C., Levasseur, A., La Scola, B., Colson, P., 2020. Vermamoeba vermiformis CDC-19 draft genome sequence reveals considerable gene trafficking including with candidate phyla radiation and giant viruses. Sci. Rep. 10 (1), 5928.

Delucchi, M., Schaper, E., Sachenkova, O., Elofsson, A., Anisimova, M., 2020. A New Census of Protein Tandem Repeats and Their Relationship with Intrinsic Disorder. Genes (Basel) 11 (4).

Edgar, R.C., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinf. 5, 113.
Filee, J., Pouget, N., Chandler, M., 2008. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. BMC Evol. Biol. 8, 320.
Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig, A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D. S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. Nature 419 (6906), 498–511.
Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis and T. L. Madden (2008). "NCBI BLAST: a better web interface." Nucleic Acids Res **36**(Web Server issue): W5-9.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589.
Kamel, M., Kastano, K., Mier, P., Andrade-Navarro, M.A., 2021. REP2: A Web Server to Detect Common Tandem Repeats in Protein Sequences. J. Mol. Biol. 433 (11), 166895.
Kamel, M., Mier, P., Tari, A., Andrade-Navarro, M.A., 2019. Repeatability in protein sequences. J. Struct. Biol. 208 (2), 86–91.
Klopfenstein, D.V., Zhang, L., Pedersen, B.S., Ramirez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C.J., Yunes, J.M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., Tang, H., 2018. GOATOOLS: A Python library for Gene Ontology analyses. Sci. Rep. 8 (1), 10872.
Koonin, E.V., Yutin, N., 2019. Evolution of the Large Nucleocytoplasmic DNA Viruses of Eukaryotes and Convergent Origins of Viral Gigantism. Adv. Virus Res. 103, 167–202.
Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 35 (6), 1547–1549.
Legendre, M., Fabre, E., Poirot, O., Jeudy, S., Lartigue, A., Alempic, J.M., Beucher, L., Philippe, N., Bertaux, L., Christo-Foroux, E., Labadie, K., Coute, Y., Abergel, C., Claverie, J.M., 2018. Diversity and evolution of the emerging Pandoraviridae family. Nat. Commun. 9 (1), 2285.
Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., Lopez, R., 2022. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res.
Marcotte, E.M., Pellegrini, M., Yeates, T.O., Eisenberg, D., 1999. A census of protein repeats. J. Mol. Biol. 293 (1), 151–160.
Mier, P., Alanis-Lobato, G., Andrade-Navarro, M.A., 2017. Context characterization of amino acid homorepeats using evolution, position, and order. Proteins 85 (4), 709–719.
Mier, P., Andrade-Navarro, M.A., 2020. The features of polyglutamine regions depend on their evolutionary stability. BMC Evol. Biol. 20 (1), 59.
Mier, P., Andrade-Navarro, M.A., 2022. PolyX2: Fast Detection of Homorepeats in Large Protein Datasets. Genes (Basel) 13 (5).
Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., Finn, R.D., Bateman, A., 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49 (D1), D412–D419.
Moore, H., Greenwell, P.W., Liu, C.P., Arnheim, N., Petes, T.D., 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. PNAS 96 (4), 1504–1509.
Opperdoes, F.R., Butenko, A., Flegontov, P., Yurchenko, V., Lukes, J., 2016. Comparative Metabolism of Free-living Bodo saltans and Parasitic Trypanosomatids. J. Eukaryot. Microbiol. 63 (5), 657–678.
Papadopoulos, J.S., Agarwala, R., 2007. COBALT: constraint-based alignment tool for multiple protein sequences. Bioinformatics 23 (9), 1073–1079.
Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. 25 (13), 1605–1612.
Sajko, S., Grishkovskaya, I., Kostan, J., Graewert, M., Setiawan, K., Trubestein, L., Niedermuller, K., Gehin, C., Sponga, A., Puchinger, M., Gavin, A.C., Leonard, T.A., Svergun, D.I., Smith, T.K., Morriswood, B., Djinovic-Carugo, K., 2020. Structures of three MORN repeat proteins and a re-evaluation of the proposed lipid-binding properties of MORN repeats. PLoS One 15 (12), e0242677.
Schaefer, M.H., Wanker, E.E., Andrade-Navarro, M.A., 2012. Evolution and function of CAG/polyglutamine repeats in protein-protein interaction networks. Nucleic Acids Res. 40 (10), 4273–4287.
Schulz, F., Abergel, C., Woyke, T., 2022. Giant virus biology and diversity in the era of genome-resolved metagenomics. Nat. Rev. Microbiol. 20 (12), 721–736.
Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539.
Soto-Arredondo, K.J., Flores-Villavicencio, L.L., Serrano-Luna, J.J., Shibayama, M., Sabanero-Lopez, M., 2014. Biochemical and cellular mechanisms regulating Acanthamoeba castellanii adherence to host cells. Parasitology 141 (4), 531–541.
Stewart, T., Wolfe, B.E., Fuchs, S.M., 2021. Defining the role of the polyasparagine repeat domain of the S. cerevisiae transcription factor Azf1p. PLoS One 16 (5), e0247285.
Takeshima, H., Komazaki, S., Nishi, M., Iino, M., Kangawa, K., 2000. Junctophilins: a novel family of junctional membrane complex proteins. Mol. Cell 6 (1), 11–22.
Torresen, O.K., Star, B., Mier, P., Andrade-Navarro, M.A., Bateman, A., Jarnot, P., Gruca, A., Grynberg, M., Kajava, A.V., Promponas, V.J., Anisimova, M., Jakobsen, K. S., Linke, D., 2019. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 47 (21), 10994–11006.
Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., Barton, G.J., 2009. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25 (9), 1189–1191.