Comparative Genomics and Environmental Distribution of Large dsDNA Viruses in the Family Asfarviridae

Sangita Karki, Mohammad Moniruzzaman and Frank O. Aylward*

Department of Biological Sciences, Virginia Tech, Blacksburg, VA, United States

The family Asfarviridae is a group of nucleocytoplasmic large DNA viruses (NCLDVs) of which African swine fever virus (ASFV) is well-characterized. Recently the discovery of several Asfarviridae members other than ASFV has suggested that this family represents a diverse and cosmopolitan group of viruses, but the genomics and distribution of this family have not been studied in detail. To this end we analyzed five complete genomes and 35 metagenome-assembled genomes (MAGs) of viruses from this family to shed light on their evolutionary relationships and environmental distribution. The Asfarvirus MAGs derive from diverse marine, freshwater, and terrestrial habitats, underscoring the broad environmental distribution of this family. We present phylogenetic analyses using conserved marker genes and whole-genome comparison of pairwise average amino acid identity (AAI) values, revealing a high level of genomic divergence across disparate Asfarviruses. Further, we found that Asfarviridae genomes encode genes with diverse predicted metabolic roles and detectable sequence homology to proteins in bacteria, archaea, and eukaryotes, highlighting the genomic chimerism that is a salient feature of NCLDV. Our read mapping from Tara oceans metagenomic data also revealed that three Asfarviridae MAGs were present in multiple marine samples, indicating that they are widespread in the ocean. In one of these MAGs we identified four marker genes with > 95% AAI to genes sequenced from a virus that infects the dinoflagellate Heterocapsa circularisquama (HcDNAV). This suggests a potential host for this MAG, which would thereby represent a reference genome of a dinoflagellate-infecting giant virus. Together, these results show that Asfarviridae are ubiquitous, comprise similar sequence divergence as other NCLDV families, and include several members that are widespread in the ocean and potentially infect ecologically important protists.

Keywords: Asfarviridae, NCLDV, Megavirales, eukaryotic viruses, Nucleocytoviricota

INTRODUCTION

The nucleo-cytoplasmic large DNA viruses (NCLDVs), also called Nucleocytoviricota, comprise a phylum of dsDNA viruses that infect diverse eukaryotes (Van Etten et al., 2010b; Koonin et al., 2020). NCLDV's include the largest viruses known, both in terms of virion size and genome length, and genomes within this group often contain genes involved in metabolic pathways...
that are otherwise present only in cellular lineages (Fischer et al., 2010; Van Etten et al., 2010b; Schvarcz and Steward, 2018; Moniruzzaman et al., 2020a). Some families of NCLDV such as the Poxviridae, Asfarviridae, Iridoviridae, and Phycodnaviridae have been studied for decades, while others, such as the Pandoraviridae, Mimiviridae, and Marseilleviridae, have been discovered relatively recently (Raoult et al., 2004; Boyer et al., 2009; Philippe et al., 2013; Abergel et al., 2015). Although amoebae have been used as an effective system to cultivate many NCLDV, recent cultivation-independent studies have discovered a wide range of these viruses in diverse environments, suggesting that uncultivated members of this viral phylum are ubiquitous in the biosphere and infect diverse hosts (Monier et al., 2008; Hingamp et al., 2013; Bäckström et al., 2019; Endo et al., 2020; Moniruzzaman et al., 2020a; Schulz et al., 2020). Given the notable complexity of NCLDV and their cosmopolitan distribution, there is a need to better understand their genomic diversity and biogeography.

The Asfarviridae is a family of NCLDV for which the most well-studied member is the African swine fever virus (ASFV), an emerging pathogen that was first discovered in 1921 (Montgomery and Eustace Montgomery, 1921). Although ASFV has been extensively studied due to its high mortality rate and subsequent economic toll on livestock production, other viruses within the same family have remained relatively underexposed, and until recently ASFV was the only known member of the Asfarviridae family. In 2009, a virus infecting the marine dinoflagellate Heterocapsa circularisquama (HcDNAV) was cultivated, and partial sequencing of the DNA polymerase type B and MutS genes revealed that the virus likely belonged to the Asfarviridae (Ogata et al., 2009). Furthermore, a new amoeba virus, Faustovirus and other isolates of amoeba-infecting Asfarviruses that clustered with the Asfarviridae have also been reported (Reteno et al., 2015; Benamar et al., 2016). Using amoeba as the host, two other Asfarviridae, Kaumoebavirus and Pacmanvirus, were isolated (Bajrai et al., 2016; Andreani et al., 2017). Lastly, a culture independent study in early 2020 reported Asfar-like virus (AbalV) causing mass mortality in abalone (Matsuyama et al., 2020). Together, these studies have begun to show that the Asfarviridae are likely a diverse family of NCLDV that are globally distributed and infect both protist and metazoan hosts.

Recently, two studies (Moniruzzaman et al., 2020a; Schulz et al., 2020) reported numerous new metagenome-assembled genomes (MAGs) of NCLDV, some of which have phylogenetic affinity with the Asfarviridae family. However, the genomic characteristics of these MAGs have not been studied in detail. In this study, we leveraged five previously available Asfarvirus genomes and 35 new Asfarvirus MAGs to perform comparative genomic and biogeographic analysis of the Asfarviridae family and provide an assessment of the scale of Asfarvirus diversity in the environment. We assess the phylogenetic relationship of these new MAGs and previously discovered Asfarviruses to explore their evolutionary relationships, and we identify the potential evolutionary origins of the Asfarviridae genomic repertoires. We also report numerous genes encoding for different functions including central amino acid metabolism, nutrient homeostasis, and host infection. Moreover, we assess the distribution of marine Asfarvirus genomes in the ocean, and we identified high sequence similarity between one marine Asfarvirus MAG to marker gene sequences available from a virus known to infect the dinoflagellate Heterocapsa circularisquama, suggesting a potential host for this MAG. Our findings reveal that the Asfarviridae members are widespread in the ocean and potentially have roles in biogeochemical cycling through infection of ecologically-important protist lineages.

**MATERIALS AND METHODS**

**Comparative Analysis and Protein Annotation**

For this study, we analyzed 35 Asfarvirus MAGs generated in two previous studies (Moniruzzaman et al., 2020a; Schulz et al., 2020) and complete genomes of five Asfarviruses (Reteno et al., 2015; Silva et al., 2015; Bajrai et al., 2016; Andreani et al., 2017; Matsuyama et al., 2020). MAGs were quality-checked using ViralRecall v. 2.0 (default parameters), with results manually inspected to ensure that no large non-NCLDV contigs were present (Aylward and Moniruzzaman, 2021). We used Seqkit v0.12.0 (Shen et al., 2016) for FASTA/Q file manipulation to generate the statistics of the genomes and proteins. To predict protein and search for tRNA genes, we used Prodigal V2.6.3 (Hyatt et al., 2010) and ARAGORN v1.2.38 (Laslett, 2004), respectively, with default parameters. For the sequence similarity search, we used BLASTp against the NCBI reference sequence (RefSeq) database, version 92 (O’Leary et al., 2016). An E-value threshold of 1e-3 was used, and maximum target sequence was set to 1. Functional annotation of predicted proteins was done using hmmsearch (parameter -E 1e-5) in HMMER v3.3 (Eddy, 2011) against the EggNOG v.5.0 database (Huerta-Cepas et al., 2016) to assess the potential function of MAG-encoded proteins, and the best hits for each protein were recorded.

We calculated protein-level orthologous groups (OGs) shared between all genomes analyzed in this study using the Proteinortho tool version 6.0.14 (Lechner et al., 2011) with default parameters. The resulting matrix for the orthologous genes was used for the bipartite network analysis. A bipartite network for the 35 MAGs along with their reference genomes were constructed using igraph (Csardi and Nepusz, 2006), and selected members of Poxviridae were used as an outgroup. The network consisted of two node types, one for genomes and one for OGs. OGs that were present in at least one genome were analyzed. A Fruchterman-Reingold layout with 10,000 iteration was used for visualization purposes.

To assess the genomic diversity between Asfarviruses, we calculated amino acid identity (AAI) using the python script available at https://github.com/faylward/lastp_aai. This script uses LAST to detect bi-directional best hits to find the pairwise identity of orthologous proteins (Kielbasa et al., 2011). The results were visualized using the gplots package (Warnes et al., 2020) in the R environment.

In order to assess the sequence similarity, the raw metagenomic reads from TARA ocean samples described
previously (Sunagawa et al., 2015) were downloaded from the NCBI SRA database, and forward Illumina reads were mapped against the selected genomes using LAST (Kiełbasa et al., 2011) with default parameters. The results were visualized with fragment recruitment plots using the ggplot2 package (Wickham, 2009) in the R environment.

**Phylogenetic Reconstruction**

To generate the phylogenetic tree, we analyzed 35 MAGs and five reference genomes along with selected members of the Poxviridae as an outgroup. We used five marker genes: major capsid protein (MCP), superfamily II helicase (SFII), virus-like transcription factor (VLTF3), DNA Polymerase B (PolB), and packaging ATPase (A32), that are previously shown to be useful and used for phylogenetic analysis of NCLDV MAGs (Yutin et al., 2009; Moniruzzaman et al., 2020a). We used a python script to identify the marker genes using hmmsearch (available at github.com/faylward/ncldv_markersearch), also previously described (Moniruzzaman et al., 2020a). We used Clustal Omega v1.2.4 (Sievers et al., 2011) for alignment, and packaging ATPase (A32), that are previously shown to be highly conserved in the NCLDVs (Yutin et al., 2009; Moniruzzaman et al., 2020a). The phylogenetic analysis revealed that although the Asfarvirus MAGs formed clades with the five reference genomes (ASFV, Abalone asfarvirus, Kaumoebavirus, Faustovirus, and Pacmanvirus) in some cases, overall, the new MAGs had deep branches and were not closely related to reference viruses. The numerous deep-branching lineages in the tree underscores the high level of phylogenetic divergence between different Asfarviruses. The new MAGs were obtained from different environments, including freshwater (18), marine (12), landfill (2), non-marine saline lake (2), and mine tailing samples (1), highlighting their broad distribution. Clustering of the isolates according to the environment was also apparent in the phylogenetic tree, with several clades found only in marine or freshwater environments (Figure 1). This suggests that the broad habitat preference of many Asfarviruses may be conserved across some clades.

The MAG GVMAG-S-3300005056-23 was the most basal-branching Asfarvirus genome. We compared the proteins encoded in this genome to the NCBI RefSeq database and found that 13 had best hits to Poxviruses (compared to at most 4 in the other Asfarvirus MAGs), while 37 proteins had best hits to Asfarvirus genomes in this database (Supplementary Data 1). Together with its basal placement in our phylogeny, these results suggest that GVMAG-S-3300005056-23 is either a basal branching Asfarvirus or possibly even a member of a new family of NCLDV. We chose to use Poxviruses to root our phylogeny because this family is often considered to be most closely related to the Asfarviridae (Iyer et al., 2006; Koonin and Yutin, 2018), but it remains unclear where the root of the NCLDV should be placed, and other studies have recovered topologies that place the Asfarviruses as a sister group to other NCLDV families (Guglielmini et al., 2019). For purposes of our analysis, here, we kept GVMAG-S-3300005056-23 as a basal-branching Asfarvirus, but further studies are needed to confirm the evolutionary provenance of this MAG.

**RESULTS AND DISCUSSION**

**Asfarvirus Genome Statistics**

The Asfarvirus MAG assembly sizes ranged from 120 kbp (SRX802982.1) to 580.8 kbp (GVMAG-S-3300009702-144). Among the 35 MAGs, 17 had all five core genes used for phylogenetic analysis (A32, PolB, MCP, SFII, and VLTF3) while the rest of the genomes were missing only one core gene, including three MAGs in which the highly conserved PolB marker was not identified. This suggests that the MAGs are generally high quality, although the absence of some marker genes suggests that some are only nearly complete and that MAG assembly sizes are underestimates of the complete genome sizes. The % G+C content for the new MAGs ranged from 17 to 60%, while those of reference viruses ranged from 31 to 45%. The ARAGORN software predicted three tRNA genes (Leu, Ile, and Asn) for ERX552270.16, one Ile-tRNA gene for GVMAG-M-3300013133-40, GVMAG-M-3300023174-161, GVMAG-M-3300027779-10, GVMAG-S-3300005056-23, and GVMAG-S-3300010160-169, and one Arg-tRNA gene for SRX319065.14. One tRNA gene (Ile) was also predicted in reference virus – Pacmanvirus as described previously (Andreani et al., 2017). The complete statistics for the MAGs are provided in Table 1.

**Phylogenetic Relationship Between the Asfarviruses**

To assess the phylogenetic diversity and evolutionary relationships of the new MAGs, we constructed a phylogenetic tree based on alignment of the five conserved marker genes. These marker genes have been previously described to be highly conserved in the NCLDVs (Yutin et al., 2009; Moniruzzaman et al., 2020a). The phylogenetic analysis revealed that the phylogenetic diversity and evolutionary relationships of the new MAGs, constructed a phylogenetic tree based on alignment of the five conserved marker genes. The phylogenetic analysis revealed that the Asfarvirus MAGs formed clades with the five reference genomes (ASFV, Abalone asfarvirus, Kaumoebavirus, Faustovirus, and Pacmanvirus) in some cases, overall, the new MAGs had deep branches and were not closely related to reference viruses. The numerous deep-branching lineages in the tree underscores the high level of phylogenetic divergence between different Asfarviruses. The new MAGs were obtained from different environments, including freshwater (18), marine (12), landfill (2), non-marine saline lake (2), and mine tailing samples (1), highlighting their broad distribution. Clustering of the isolates according to the environment was also apparent in the phylogenetic tree, with several clades found only in marine or freshwater environments (Figure 1). This suggests that the broad habitat preference of many Asfarviruses may be conserved across some clades.

The MAG GVMAG-S-3300005056-23 was the most basal-branching Asfarvirus genome. We compared the proteins encoded in this genome to the NCBI RefSeq database and found that 13 had best hits to Poxviruses (compared to at most 4 in the other Asfarvirus MAGs), while 37 proteins had best hits to Asfarvirus genomes in this database (Supplementary Data 1). Together with its basal placement in our phylogeny, these results suggest that GVMAG-S-3300005056-23 is either a basal branching Asfarvirus or possibly even a member of a new family of NCLDV. We chose to use Poxviruses to root our phylogeny because this family is often considered to be most closely related to the Asfarviridae (Iyer et al., 2006; Koonin and Yutin, 2018), but it remains unclear where the root of the NCLDV should be placed, and other studies have recovered topologies that place the Asfarviruses as a sister group to other NCLDV families (Guglielmini et al., 2019). For purposes of our analysis, here, we kept GVMAG-S-3300005056-23 as a basal-branching Asfarvirus, but further studies are needed to confirm the evolutionary provenance of this MAG.

In addition to phylogenetic analysis, we also performed pairwise AAI analysis to assess the genomic divergence between different Asfarviruses. Our analysis recovered pairwise AAI values ranging from 27 to 75% (Figure 2) with mean and median values of 31.7 and 31.0%, respectively. This result is consistent
TABLE 1 | General statistics of the five Asfarvirus genomes and 35 viral MAGs.

| MAGs | No. of contigs | Genome length | GC content | No. of proteins | N50 size | tRNA genes | Missing core genes |
|------|----------------|---------------|------------|----------------|----------|------------|-------------------|
| ERX552720.16  | 10  | 262,392       | 20.82      | 229            | 25,815   | 3          | VLTF3             |
| ERX556003.45   | 13  | 246,693       | 18.51      | 215            | 19,050   | 0          | A32               |
| GVMAG-M-330000574-23 | 8  | 214,432       | 27.43      | 214            | 45,232   | 0          | MCP               |
| GVMAG-M-330000906-46 | 18 | 220,122       | 25.09      | 188            | 13,641   | 0          | –                 |
| GVMAG-M-330000943-29 | 11 | 215,977       | 17.34      | 198            | 21,496   | 0          | –                 |
| GVMAG-M-3300010160-26 | 6  | 221,296       | 29.40      | 199            | 55,434   | 0          | –                 |
| GVMAG-M-330001300-23 | 8  | 285,977       | 40.48      | 251            | 76,336   | 0          | –                 |
| GVMAG-M-330001333-40 | 15 | 171,139       | 34.27      | 158            | 11,819   | 1          | MCP               |
| GVMAG-M-3300014204-43 | 9  | 297,005       | 49.23      | 274            | 41,226   | 0          | –                 |
| GVMAG-M-3300014204-45 | 10 | 202,564       | 37.00      | 201            | 7,786    | 0          | –                 |
| GVMAG-M-3300017989-35 | 19 | 154,311       | 20.07      | 325            | 23,777   | 0          | PolB              |
| GVMAG-M-3300018416-36 | 21 | 222,130       | 46.21      | 277            | 173,012  | 0          | –                 |
| GVMAG-M-3300020564-1 | 19 | 197,061       | 20.23      | 219            | 11,922   | 0          | –                 |
| GVMAG-M-3300022309-7 | 19 | 310,111       | 20.07      | 325            | 23,777   | 0          | PolB              |
| GVMAG-M-3300022916-57 | 3  | 330,706       | 46.21      | 277            | 173,012  | 0          | –                 |
| GVMAG-M-3300023174-150 | 21 | 225,080       | 29.51      | 220            | 13,647   | 0          | MCP               |
| GVMAG-M-3300023174-161 | 21 | 229,163       | 30.10      | 206            | 15,453   | 1          | A32               |
| GVMAG-M-3300023174-165 | 29 | 353,296       | 26.17      | 398            | 12,847   | 0          | –                 |
| GVMAG-M-3300023184-110 | 14 | 474,957       | 25.61      | 372            | 44,583   | 0          | –                 |
| GVMAG-M-3300023184-186 | 10 | 496,936       | 27.06      | 417            | 55,208   | 0          | –                 |
| GVMAG-M-3300024062-1 | 14 | 349,792       | 21.91      | 377            | 56,296   | 0          | –                 |
| GVMAG-M-3300027707-33 | 21 | 337,477       | 30.37      | 299            | 20,830   | 0          | A32               |
| GVMAG-M-3300027793-10 | 14 | 349,525       | 33.53      | 405            | 33,981   | 1          | –                 |
| GVMAG-M-3300027833-19 | 11 | 297,297       | 30.60      | 324            | 73,615   | 0          | –                 |
| GVMAG-M-1035124-107 | 7   | 239,202       | 18.99      | 200            | 55,079   | 0          | –                 |
| GVMAG-M-1039444-30 | 13  | 223,700       | 59.45      | 215            | 25,128   | 0          | –                 |
| GVMAG-M-3300002466-141 | 10 | 331,811       | 39.41      | 420            | 49,921   | 0          | –                 |
| GVMAG-M-3300005056-23 | 19 | 518,885       | 29.19      | 549            | 37,879   | 1          | –                 |
| GVMAG-M-3300009702-144 | 28 | 580,795       | 28.00      | 591            | 38,326   | 0          | –                 |
| GVMAG-M-3300010160-169 | 11 | 229,474       | 33.48      | 226            | 22,741   | 1          | SFII              |
| SRX319064.32 | 8   | 154,393       | 32.72      | 145            | 27,897   | 0          | –                 |
| SRX319065.14 | 3   | 342,906       | 21.77      | 305            | 178,940  | 1          | –                 |
| SRX227722.61 | 4   | 202,406       | 45.42      | 179            | 47,579   | 0          | –                 |
| SRX029263.109 | 8   | 175,258       | 30.38      | 161            | 18,153   | 0          | –                 |
| SRX029282.1 | 6   | 120,034       | 43.81      | 131            | 18,923   | 0          | –                 |
| ASFV  | –   | 170,101       | 38.95      | –              | –        | 0          | –                 |
| Abalone asfarvirus | –   | 155,181       | 31.63      | –              | –        | 0          | –                 |
| Faustovirus | –   | 466,265       | 36.22      | –              | –        | 0          | –                 |
| Kaumoebavirus | –   | 350,731       | 43.7       | –              | –        | 0          | –                 |
| Pacmanvirus | –   | 395,405       | 33.62      | –              | –        | 1          | –                 |

Pan-Genomics of the Asfarviruses

We found 7,410 total OGs, including 6,480 that were found in one Asfarvirus genome only. The number of unique OGs for each genome ranged from 48 to 428. We observed 12 core OGs in 90% of genomes, including the MCP, VLTF3-like transcription factor, A32 packaging ATPase, DNA topoisomerase II, DNA ligase, DNA PolB, RNA polymerase subunit B, ATP dependent helicase hrpA, VV A8L-like transcription factor, and some hypothetical proteins (Figure 3). Nonetheless, the high number of genome-specific OGs highlights the genomic diversity present in the Asfarviridae family, which is consistent with the high level of variability in other families of NCLDV (Van Etten et al., 2010b).

In order to visualize the pattern of gene sharing, we performed bipartite network analysis using the Asfarvirus OGs, with six Poxvirus genomes used as non-Asfarvirus references. Given that virus evolution is characterized by extensive gene loss, gain, and exchange, this approach can be complementary to traditional phylogenetic analysis (Iranzo et al., 2016). The bipartite network showed some clustering of the MAGs based upon the habitat (Figure 4), although many co-clustered MAGs are also closely

with the deep-branching clades identified in the phylogenetic analysis and confirms the high genomic divergence within the Asfarviridae.
related and common gene content due to shared ancestry cannot be ruled out. The Poxviridae clustered separately in a small sub-network, indicating that their gene content is clearly distinct from the Asfarviridae. Hence, the bipartite network provides support for the phylogenetic findings we have for the Asfarviruses and depicts the gene-sharing pattern of these viruses.
**FIGURE 3** Unique and core genes shared between the MAGs and reference Asfarviruses. Here, we define “core” as all genes found in 90% or more genomes.

**FIGURE 4** Bipartite network plot for the MAGs. The larger nodes represent genomes while the smaller nodes represent OGs/gene families. Genomes were connected to the genes if they encode one. MAGs are colored based on their habitat. The size of the larger nodes represents genome size.
Genomic Chimerism of the Asfarviruses

Nucleo-cytoplasmic large DNA viruses are known to have chimeric genomes with genes that are derived from multiple sources (Boyer et al., 2009), and we therefore sought to quantify the extent of this genomic chimerism in environmental Asfarviruses by comparing the encoded proteins of the Asfarvirus MAGs to the RefSeq database (see section “Materials and Methods” for details; Supplementary Data 1). We found that between 40 and 70% of the proteins in each genome had no detectable hits to reference proteins, while 16–55% had best matches to other viruses, 5–22% to Eukaryotes, 3–15% to Bacteria, and 0–2% to Archaea (Figure 5A). We examined the proteins with best hits to Eukaryotes in more detail because this may provide some insight into host-virus gene exchange and therefore link these viruses to putative hosts. Overall, best hits to eukaryotes included matches to Animalia, Plantae, Fungi, and Protists such as Stramenopiles, Alveolata, Archaeplastida, Cryptista, Excavata, Choanomonada, Apusozoa, Porifera, and Amoebozoa (Figure 5B). The percent identity of these matches ranged from 19.4 to 93.2 (median 35.3), with only 4 greater than 90%, suggesting that, if these represent gene exchanges between NCLDV and eukaryotes, the vast majority have not occurred recently. Although recent studies have revealed a dynamic gene exchange between NCLDV and eukaryotic lineages that can be used to link viruses to their hosts (Moniruzzaman et al., 2020b; Schulz et al., 2020), our analysis did not identify any clear signatures in the Asfarvirus MAGs that could be used for this purpose. It is possible that future work examining endogenous NCLDV signatures in eukaryotic genomes may be useful to better identify virus-host relationships.

Asfarvirus Genes Involved in Manipulating Host Metabolism

To assess the potential functions of the proteins encoded by the MAGs, we performed functional annotation using HMMPER searches against the EggNOG database (all annotations available in Supplementary Data 2). As expected, in all MAGs we detected genes involved in DNA replication and repair, transcription, and post-translational modification, which is consistent with the prevalence of these functions across NCLDV (Yutin and Koonin, 2012; Figure 6). Among the proteins involved in post-translational modification, we found genes responsible for ubiquitination (KOG0802 and KOG1812) and ubiquitin dependent proteins in 26 MAGs. Ubiquitination has been found to be an important counteracting mechanism to oxidative stress response in eukaryotes that direct the unwanted proteins to proteasome for degradation (Silva et al., 2015). In Acreococcus anophagefferens giant virus (AaV), ubiquitin dependent protein-ubiquitin ligases were found to be expressed within 5 min of virus-infection and is thought to be involved in degradation of host proteins (Moniruzzaman et al., 2018). The ubiquitin protein has also been reported in Marseilleviruses, where it is thought to play an important role in host signaling (Boyer et al., 2009). A protein homologous to the ubiquitin-proteasome (UP) system has been found to be encoded by ASFV, suggesting its role during early infection and replication (Barrado-Gil et al., 2017). Hence, this suggests that ubiquitination may be a common mechanism across diverse Asfarviruses.

Genes predicted to be involved in carbohydrate metabolism were prevalent in the MAGs, consistent with previous findings that these genes are widespread in NCLDVs. We observed glycosyltransferase enzymes that are important in glycosylation of viral proteins in 15 Asfarvirus MAGs. These enzymes have been previously reported in giant viruses (Markine-Goriaynoff et al., 2004). Also, past studies have indicated the presence of glycosylating genes (Van Etten et al., 2010a; Picente et al., 2015) and other enzymes involved in carbohydrate metabolism in NCLDVs (Fischer et al., 2010). Interestingly, we found genes involved in the shikimate pathway that is linked to the biosynthesis and metabolism of carbohydrates and aromatic amino acids (phenylalanine, tryptophan, and tyrosine) in five MAGs. We found 3-deoxy-7-phosphohexulonate synthase (2QPSU) (the first enzyme in the shikimate pathway), chorismate synthase (KOG4492), and prephenate dehydrogenase (KOG2380) all in ERX556003.45 and only 3-deoxy-7-phosphohexulonate synthase in four other MAGs. The shikimate pathway is widespread in bacteria, archaea, and protists but not in metazoans (Richards et al., 2006). We also found acetolactate synthase genes (KOG4166) in three MAGs. Acetolactate synthase that are involved in the synthesis of amino acids such as leucine, isoleucine, and valine has been previously described to be present in large DNA viruses infecting green algae mainly, Prasinovirus (Weynberg et al., 2009; Moreau et al., 2010; Zhang et al., 2015). Hence, the detection of these enzymes shows the potential role of the Asfarvirus MAGs in the manipulation of amino acid metabolism in their hosts during infection.

Genes responsible for signal transduction were also present in some of the MAGs. KOGs representing serine/threonine protein kinase and tyrosine/threonine phosphatase were present in seven MAGs. These enzymes constitute a major form of signaling and regulation of many cellular pathways such as cell proliferation, differentiation, and cell death. Serine/threonine kinases have also been reported in Marseillevirus, Iridivirus, and Ascoviruses (Boyer et al., 2009; Piégu et al., 2015) and ASFV, suggesting that it might have a role in early infection and programmed cell death (apoptosis) (Baylis et al., 1993).

We found genes homologous to cysteine desulurase (COG1104) proteins in 21 out of 35 MAGs (Supplementary Data 2). NiF genes whose presumed functions are similar to that of cysteine desulurase are reported to be associated with ASFV, Faustovirus, and Pacmanvirus with possible involvement in host cell interactions (Andreani et al., 2017). Cysteine desulurase proteins are found in bacteria and eukaryotes and are involved in the biosynthesis of iron (Fe) – sulphur (S) clusters, thiamine, biotin, lipoic acid, molybdopterin, NAD, and thionucleosides in tRNA (Mihara and Esaki, 2002). Hence, the discovery of the enzyme cysteine desulurase adds to the viral proteins involved in electron transfer processes.

Gene encoding for cell redox homeostasis (KOG0191) and cellular response to nitrogen starvation (KOG1654) were also common among the MAGs. Nutrient limitation has the potential
to reduce viral productivity; virus reproduction mostly depends upon the intracellular nitrogen and phosphorous pool during early infection while they might depend upon the extracellular nitrogen availability as infection proceeds (Zimmerman et al., 2020). Genes involved in responding to nutrient starvation can influence the nutrient uptake and replication in these viruses. Overall, these results demonstrate that in addition to universal genes that play a role in host invasion and viral replication,
FIGURE 6 | Protein annotation for MAGs. The x-axis represents the MAGs while y-axis represents the COG category. The number inside the bubble represents the number of genes present in that MAG that had the annotated function.

FIGURE 7 | (A) Distribution of Asfarvirus matching metagenomic reads from the TARA ocean project. (B–D) Fragment recruitment plot for metagenomic reads to ERX552270.16, ERX556003.45, and GVMAG-M-3300027833-19, respectively. The x-axis of the recruitment plot shows position of the metagenomic reads along the genome length and y-axis represents the percent identity.
Asfarviruses also contain genes involved in metabolism, hence, capable of reprogramming cells into virocells during infection (Moniruzzaman et al., 2020a).

### Biogeography of Marine Asfarviruses

While ASFV is a terrestrial pathogen and most cultured Asfarviruses were isolated from sewage samples, various metagenomic studies have revealed that NCLDV are highly diverse and abundant in aquatic environments (Monier et al., 2008; Hingamp et al., 2013), and one recent study noted that Asfarviruses are prevalent in some marine samples (Endo et al., 2020). To examine the biogeography of the Asfarvirus MAGs in more detail we conducted a fragment recruitment analysis using reads from the Tara oceans expedition (Sunagawa et al., 2015).

We examined 28 diverse metagenomic samples from surface and deep chlorophyll maxima (DCM) oceanic regions. The Asfarvirus MAG ERX552270.16 was present in eight metagenomic samples (from five different TARA stations), ERX556003.45 was found in 19, and GVMAG-M-3300027833-19 was found in one, revealing that some Asfarvirus are globally distributed in the ocean (Figure 7A). The fragment recruitment plots revealed that the MAGs had consistent coverage of reads with 100% nucleic acid identity matches to the metagenomic reads (Figures 7B–D and Supplementary Figures 1, 2), demonstrating high similarity of these viruses across long distances. Few gaps were visible in the recruitment plots, indicating the absence of readily-identifiable genomic islands in these viruses.

Previous studies have shown that the virus HcDNAV infects the marine dinoflagellate *Heterocapsa circularisquama*, which is responsible for harmful algal blooms in the marine environment (Tarutani et al., 2001; Nagasaki et al., 2003). This is notable since very few viruses that infect dinoflagellates have been characterized, and of these HcDNAV is the only large DNA virus (Nagasaki, 2008). Although a complete genome of HcDNAV is not available, several marker genes from this virus have been sequenced, are available in NCBI and have been previously reported (Ogata et al., 2009). We found that the MAG ERX552270.16 bore high sequence similarity to the HcDNAV marker genes, indicating that this MAG represents a closely related virus that potentially infects the same host. The Family B Polymerase (YP_009507841.1), HNH endonuclease (YP_009507839.1), DNA directed RNA Polymerase (BAI48199.1), and DNA mismatch repair protein (mutS) (BAJ49801.1) of HcDNAV all had 95.8 to 99% AAI to homologs in ERX552270.16 (Table 2). The PolB enzyme of

---

**TABLE 2** Amino acid identity between the HcDNAV genes (only genes available at NCBI) and MAG ERX552270.16 as analyzed by blastp.

| HcDNAV genes                                  | MAG genes         | AAI %  |
|-----------------------------------------------|-------------------|--------|
| HNH endonuclease (YP_009507839.1)             | contig_19878_13   | 95.88  |
| DNA mismatch repair protein (BAJ49801.1)      | contig_19878_12   | 98.37  |
| Type B DNA polymerase (YP_009507841.1)        | contig_19878_19   | 98.97  |
| DNA-directed RNA polymerase subunit (BAI48199.1) | contig_19878_28   | 98.97  |

![Phylogenetic tree reconstruction based on DNA polymerase B gene (New reference virus HcDNAV has been added). The size of the black dot represents the bootstrap values. Only bootstrap values greater than 0.5 are shown.](image)
ERX552270.16 also contained the notable YSDTDS motif that was previously found in HcDNNAV (Ogata et al., 2009). Moreover, we constructed a PolB phylogeny of the Asfarviridae that confirmed that these viruses cluster closely together (Figure 8). Our fragment recruitment analysis from Tara Oceans data confirmed that ERX552270.16 is widespread in the ocean, especially in costal environments (Figure 7 and Supplementary Figure 1), consistent with the hypothesis that it is a marine virus that also infects Heterocapsa circularisquama or a closely related dinoflagellate. Given these similarities, ERX552270.16 may be a useful reference genome for exploring the genomics and distribution of close relatives of HcDNNAV, though further work will be necessary to confirm the host of ERX552270.16.

CONCLUSION

While ASFV was the only known member of Asfarviridae for many years, recent work has identified numerous additional members of this viral family. In this study, we provide a robust phylogenetic and comparative genomic analysis of this viral family. Our results highlight the high level of genomic and phylogenetic divergence between disparate members of the Asfarviridae, and homology searches suggest that many genes within this viral group are potentially the product of ancient horizontal transfers from cellular lineages. Moreover, we provide fragment recruitment plots that confirm that some Asfarviruses are ubiquitous in the ocean, where they may infect ecologically important protists such as bloom forming dinoflagellates. These findings suggest that diverse Asfarviruses are broadly distributed in the environment and play important roles in numerous ecosystems.

REFERENCES

Abergel, C., Legendre, M., and Claverie, J.-M. (2015). The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol. Rev.* 39, 779–796. doi: 10.1093/femsre/fuv037

Andreani, J., Khalil, J. Y. B., Sevvana, M., Benamar, S., Di Pinto, F., Bitam, I., Abergel, C., Legendre, M., and Claverie, J.-M. (2015). The rapidly expanding universe of giant viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus. *FEMS Microbiol. Rev.* 39, 779–796. doi: 10.1093/femsre/fuv037

Benamar, S., Reteno, D. G. I., Bandaly, V., Labas, N., Raoult, D., and La Scola, B. (2016). Faustoviruses: comparative genomics of new megavirales family members. *Front. Microbiol.* 7:3. doi: 10.3389/fmicb.2016.00003

Boyer, M., Yutin, N., Pagnier, I., Barrasi, L., Fournous, G., Espinosa, L., et al. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl Acad. Sci. U.S.A.* 106, 21848–21853. doi: 10.1073/pnas.0911354106

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Cerda, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Int. J. Complex Syst.* 1695, 1–9.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Endo, H., Blanc-Mathieu, R., Li, Y., Salazar, G., Henry, N., Labadie, K., et al. (2020). Biogeography of marine giant viruses reveals their interchange with eukaryotes and ecological functions. *Nat. Ecol. Evol.* 4, 1639–1649. doi: 10.1038/s41559-020-01288-w

Fischer, M. G., Allen, M. J., Wilson, W. H., and Suttle, C. A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc. Natl Acad. Sci. U.S.A.* 107, 19508–19513. doi: 10.1073/pnas.1007615107

Guglielmini, J., Woo, A., Krupovic, M., Forterre, P., and Gaia, M. (2019). Pacmanvirus, a new giant icosahedral virus at the crossroads between Asfarviridae and Faustoviruses. *J. Virol.* 91, e212–17. doi: 10.1128/JVI.00212-17

Karki et al.

**DATA AVAILABILITY STATEMENT**

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/.

**AUTHOR CONTRIBUTIONS**

FA designed the study. SK and MM performed the experiment. SK and FA wrote the manuscript. All authors contributed to the article and approved the submitted version.

**FUNDING**

This research was funded by a Simons Foundation Early Career Award in Marine Microbial Ecology and Evolution and an NSF IIBR award 1918271 to FA.

**ACKNOWLEDGMENTS**

We acknowledge the use of the Virginia Tech Advanced Research Computing Center for bioinformatic analyses performed in this study. We are thankful to the members of Aylward Lab for their helpful suggestions.

**SUPPLEMENTARY MATERIAL**

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.657471/full#supplementary-material

Benamar, S., Reteno, D. G. I., Bandaly, V., Labas, N., Raoult, D., and La Scola, B. (2016). Faustoviruses: comparative genomics of new megavirales family members. *Front. Microbiol.* 7:3. doi: 10.3389/fmicb.2016.00003

Boyer, M., Yutin, N., Pagnier, I., Barrasi, L., Fournous, G., Espinosa, L., et al. (2009). Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc. Natl Acad. Sci. U.S.A.* 106, 21848–21853. doi: 10.1073/pnas.0911354106

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Cerda, G., and Nepusz, T. (2006). The igraph software package for complex network research. *Int. J. Complex Syst.* 1695, 1–9.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Endo, H., Blanc-Mathieu, R., Li, Y., Salazar, G., Henry, N., Labadie, K., et al. (2020). Biogeography of marine giant viruses reveals their interchange with eukaryotes and ecological functions. *Nat. Ecol. Evol.* 4, 1639–1649. doi: 10.1038/s41559-020-01288-w

Fischer, M. G., Allen, M. J., Wilson, W. H., and Suttle, C. A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc. Natl Acad. Sci. U.S.A.* 107, 19508–19513. doi: 10.1073/pnas.1007615107

Guglielmini, J., Woo, A., Krupovic, M., Forterre, P., and Gaia, M. (2019). Pacmanvirus, a new giant icosahedral virus at the crossroads between Asfarviridae and Faustoviruses. *J. Virol.* 91, e212–17. doi: 10.1128/JVI.00212-17

Aylward, F. O., and Moniruzzaman, M. (2021). ViralRecall: a flexible command-line tool for the detection of giant virus signatures in omic data. *Viruses* 13:150. doi: 10.3390/v13020150

Bäckström, D., Yutin, N., Jørgensen, S. L., Dharamshi, J., Homa, F., Zaremba-Niedzwiedzka, K., et al. (2019). Virus genomes from deep sea sediments expand the ocean megavirome and support independent origins of viral gigantism. *MBio* 10, e2497–18. doi: 10.1128/mBio.02497-18

Bajrai, L. H., Benamar, S., Azhar, E. I., Robert, C., Levasseur, A., Raoult, D., et al. (2016). Kaumoebavirus, a new virus that clusters with faustoviruses and Asfarviridae. *Viruses* 8:278. doi: 10.3390/v8110278

Barrado-Gil, L., Galindo, I., Martinez-Alonso, D., Viedma, S., and Alonso, C. (2017). The ubiquitin-proteasome system is required for African swine fever replication. *PLoS One* 12:e0189741. doi: 10.1371/journal.pone.0189741

Baylis, S. A., Banham, A. H., Vydelingum, S., Dixon, L. K., and Smith, G. L. (1993). African swine fever virus encodes a serine protein kinase which is packaged into virions. *J. Virol.* 67, 4549–4556. doi: 10.1128/VI.67.8.4549-4556.1993

FISH - Frontiers in Microbiology | www.frontiersin.org

March 2021 | Volume 12 | Article 657471

11
Hingamp, P., Grimsley, N., Acinas, S. G., Clerissi, C., Subirana, L., Poulain, J., et al. (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. ISME J. 7, 1678–1695. doi: 10.1038/ismej.2013.59

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 35, 518–522. doi: 10.1093/molbev/msx281

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., et al. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44, D286–D293.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, L. M., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. doi: 10.1186/1471-2105-11-119

Irango, J., Krupovic, M., and Koonin, E. V. (2016). The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. MBio 7, e0798–e16. doi: 10.1128/mBio.00798–16

Iyer, L. M., Balaji, S., Koonin, E. V., and Aravind, L. (2006). Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res. 117, 156–184. doi: 10.1016/j.virusres.2006.01.009

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimations. Nat. Methods 14, 587–589. doi: 10.1038/nmeth.4285

Kielbasa, S. M., Wan, R., Sato, K., Hortom, P., and Frith, M. C. (2011). Adaptive seeds tame genome sequence comparison. Genome Res. 21, 487–493. doi: 10.1101/gr.113985.110

Koonin, E. V., Dolja, V. V., Krupovic, M.,Varsani, A., Wolf, Y. I., Yutin, N., et al. (2020). Global organization and proposed megat Pax. Microbiol. Mol. Biol. Rev. 84, e0061–19. doi: 10.1128/MMBR.00061-19

Koonin, E. V., and Yutin, N. (2018). Multiple evolutionary origins of giant viruses. Mol. Phylogenet. Evol. 117, 156–184. doi: 10.1016/j.ympev.2018.12.013

Krause, T., Schult, F., Roux, S., Paez-Espino, D., Jungbluth, S., Walsh, D. A., Denef, V. J., et al. (2020). Giant virus diversity and host interactions through global metagenomics. Nature 578, 432–436. doi: 10.1038/s41586-020-1957-x

Schvarcz, C. R., and Steward, G. F. (2018). A giant virus infecting green algae encodes key fermentation genes. Virology 518, 433–433. doi: 10.1016/j.viroi.2018.03.010

Sen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: a cross-platform and ultrafast journal.pone.0163962. doi: 10.1371/....
Weynberg, K. D., Allen, M. J., Ashelford, K., Scanlan, D. J., and Wilson, W. H. (2009). From small hosts come big viruses: the complete genome of a second Ostreococcus taurivirus, OtV-1. *Environ. Microbiol.* 11, 2821–2839. doi: 10.1111/j.1462-2920.2009.01991.x

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Berlin: Springer Science & Business Media.

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

Yutin, N., Wolf, Y. I., Raoult, D., and Koonin, E. V. (2009). Eukaryotic large nucleocytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol. J.* 6:223. doi: 10.1186/1743-422x-6-223

Zhang, W., Zhou, J., Liu, T., Yu, Y., Pan, Y., Yan, S., et al. (2015). Four novel algal virus genomes discovered from Yellowstone Lake metagenomes. *Sci. Rep.* 5:15131.

Zimmerman, A. E., Howard-Varona, C., Needham, D. M., John, S. G., Worden, A. Z., Sullivan, M. B., et al. (2020). Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat. Rev. Microbiol.* 18, 21–34. doi: 10.1038/s41579-019-0270-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Karki, Moniruzzaman and Aylward. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.