Isolation and characterization of viruses infecting Acidobacteria from Arctic soil

Heli Marttila
Master's thesis
University of Helsinki
Department of Microbiology
Microbiology
October 2021

Heli Marttila
Master´s thesis
University of Helsinki
November 2021

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet/ — Faculty	Masters´s Programme
Faculty of Agriculture and Forestry	Microbiology and Microbial Biotechnology

Tekijä — Författare — Author

Heli Marttila

Työn nimi — Arbetets titel — Title

Isolation and characterization of viruses infecting Acidobacteria from Arctic soil

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Number of pages
Master's Thesis	November 2021	49 + 20

Tiivistelmä — Referat — Abstract

Global warming affects permafrost in the Arctic regions, where melting organic carbon storages will increasingly contribute to the emission of greenhouse gases. Little is known about tundra soil microbial communities, but Acidobacteria and viruses seem to have important roles there. Here, for the first time, we isolated five Acidobacteria infecting viruses from Kilpisjärvi tundra soils using host strains previously isolated from the same area. Three viruses were isolated on *Edaphobacter* sp. X5P2, one on *Edaphobacter* sp. M8UP27, and one on *Granulicella* sp. X4BP1. The viruses had circular double-stranded DNA genomes 63,196–308,711 bp in length and 51–58% GC content. From 108 to 348 putative ORFs were predicted, 54–72% of which were sequences unique to each virus. Annotations indicated that all five phages most likely have tailed virions. The diversity of viruses present in the studied soils was estimated with the metagenome analysis. Only 0.1% (627) of all assembled metagenomic contigs were phage-positive. The gene-sharing network analysis showed approximately genus-level clustering between the virus isolates and a few metagenomic viral contigs, but overall, all (except one) viral contigs clustered only with each other, not with any known viruses from the NCBI database. No taxonomical assignments could be done for the metagenomic viral contigs, highlighting overall undersampling of soil viruses. Further detailed studies on virus-host interactions are needed to understand the impact of viruses on host abundance and metabolism in Arctic soils, as well as the microbial input into biogeochemical cycles.

Avainsanat — Nyckelord — Keywords

virus, Arctic, tundra soil, Acidobacteria, metagenomics, bacteriophage

Säilytyspaikka — Förvaringsställe — Where deposited

HELDA - Digital Repository of the University of Helsinki

Muita tietoja — Övriga uppgifter — Further information

Supervisors: PhD Tatiana Demina, Docent Jenni Hultman

Heli Marttila
Maisterintutkielma
Helsingin yliopisto
Marraskuu 2021

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet/ — Faculty	Masters´s Programme
Maatalous-metsätieteellinen	Mikrobiologian ja mikrobiotekniikan maisteriohjelma

Tekijä — Författare — Author

Heli Marttila

Työn nimi — Arbetets titel — Title

Tundramaan Acidobakteereita infektoivien virusten eristäminen ja kuvaaminen

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivu määrä
Maisterintutkielma	Marraskuu 2021	49 + 20

Tiivistelmä — Referat — Abstract

Ilmaston lämpeneminen vaikuttaa ikiroutaan arktistisilla alueilla, missä sulavat orgaanisen hiilen varastot yhä enenevässä määrin vaikuttavat kasvihuonekaasujen päästöihin. Maaperän mikrobiyhteisöiden toiminnasta tiedetään toistaiseksi vähän, mutta Acidobacteria-pääjakson bakteereilla ja viruksilla on merkittävä rooli tundran maaperässä. Tässä työssä eristimme ensimmäistä kertaa viisi Acidobakteereita infektoivaa virusta Kilpisjärven tundramaasta käyttäen aiemmin samalta alueelta eristettyjä isäntäkantoja. Kolme virusta eristettiin *Edaphobacter* sp. X5P2-kannalla, yksi *Edaphobacter* sp. M8UP27-kannalla, ja yksi *Granulicella* sp. X4BP1-kannalla. Viruksilla oli rengasmainen kaksijuosteinen DNA-genomi, pituudeltaan 63,196–308,711 emäsparia ja GC-pitoisuus oli 51–58 %. Sekvensseistä ennustettiin 108–348 mahdollisesta avointa lukukehystä, joista 54–72 % olivat uniikkeja näille viruksille. Annotoinnin perusteella kaikilla viidellä viruksella on hännällinen virioni. Virusten monimuotoisuutta tundramaassa arvioitiin metagenomi-analyysilla. Vain 0.1 % (627 kpl) metagenomin koostetuista kontigeista oli faagipositiivisia. Yhteisten geenien verkostoanalyysi osoitti sukutason ryhmittymistä eristettyjen virusten ja metagenomin suhteen muutaman kontigin välillä ja yhtä lukuun ottamatta kaikki kontigit ryhmittyivät vain keskenään eivätkä minkään NCBI-tietokannan tunnetun viruksen kanssa. Metagenomin virus-kontigeja ei pystytty luokittelemaan taksonomisesti, mikä korostaa maaperävirusten vähäistä osuutta tietokannoissa. Yksityiskohtaisia tutkimuksia virus-isäntä-vuorovaikutuksista tarvitaan, jotta voimme ymmärtää virusten vaikutuksia isäntien runsauteen ja aineenvaihduntaan arktisessa maaperässä, ja mikrobien merkitystä biogeokemiallisissa kierroissa.

Avainsanat — Nyckelord — Keywords

virus, Arktis, tundran maaperä, Acidobakteeri, metagenomiikka, bakteriofaagi

Säilytyspaikka — Förvaringsställe — Where deposited

HELDA - Helsingin yliopiston digitaalinen arkisto

Muita tietoja — Övriga uppgifter — Further information

Ohjaajat: FT Tatiana Demina, Dosentti Jenni Hultman

Abbreviations

AMG	auxiliary metabolic gene
ATP	adenosine triphosphate
ATPase	adenosine triphosphatase
BLAST	basic local alignment search tool
bp	base pair
DNA	deoxyribonucleic acid
ds	double-stranded
GC-content	guanine-cytosine content
NCBI	National Center for Biotechnology Information
ORF	open reading frame
PFU	plaque forming unit
RNA	ribonucleic acid
ss	single-stranded
vOTU	viral operational taxonomic unit

Table of contents

1. Introduction

1.1 Ecological importance of Acidobacteria in Arctic tundra soils

Human-induced global warming has been observed to accelerate during the last decades (WMO 2021). The accumulating greenhouse gases, mainly carbon dioxide and methane, play a crucial role in the climate change. In high-latitude regions, temperature rise is amplified and temperatures in the Arctic have been measured to rise even twice as fast as the global average causing permafrost to melt (Post et al., 2019). This has drawn attention to the organic carbon stored in permafrost-affected soils, which are estimated to contain twice as much carbon as is currently in the atmosphere (Tarnocai et al., 2009; Nauta et al., 2015). Permafrost thawing is anticipated to increase the rate of carbon dioxide and methane production by soil microbes (Mackelprang et al., 2011). Microbial decomposition of soil organic materials is expected to speed up, as the temperature is rising in these high-latitude areas (Post et al., 2019; WMO 2021). It is challenging to accurately estimate these effects, as little is known about the microbial communities residing in permafrost soils. A diverse range of bacterial and archaeal phyla have been found in permafrost and Arctic soils, where the most abundant bacteria belong to Acidobacteria, Actinobacteria, Proteobacteria and Chloroflexi, and archaea belong to the Euryarchaeota (Hultman et al., 2015; Woodcroft et al., 2018; Pessi et al., 2020). Acidobacteria are widely distributed in Scandinavian Arctic soils (Männistö and Häggblom 2006, Männistö et al. 2007, 2011, 2012, 2013; Emerson et al., 2018; Trubl et al., 2018; Woodcroft et al., 2018). For example, it has been shown to be the most abundant phylum in Stordalen Mire in the northern Sweden, where it is the dominant polysaccharide degrading group of bacteria. In Stordalen Mire, the highest relative abundance of Acidobacteria was observed in the bog habitat (29%), while the relative abundancies were lower in the palsa (5%) and fen (3%) habitats (Woodcroft et al., 2018). Acidobacteria have been shown to dominate in the tundra soils in Kilpisjärvi (Finnish Lapland), where Männistö et al. (2007) studied the seasonal and spatial variations in microbial communities. DNA and fatty acid profile analysis indicated similar microbial communities at various altitudes and under different vegetation types. Members of the phylum Acidobacteria were especially abundant in slightly acidic soils, and different bedrock materials causing variation in the soil pH were the major factor affecting microbial community composition in the studied Kilpisjärvi sites (Männistö et al. 2007).

Molecular methods have shown that a variety of Acidobacteria are common in soil environments, but they also reside in bogs, freshwater, hot springs and waste waters (Kishimoto et al., 1991; Barns et al., 1999; Dedysh et al., 2006; Jones et al., 2009; Lee and Cho 2009; Kalam et al., 2020). Acidobacteria are widely distributed in Arctic and boreal soils, but very little is known about their functional and ecological roles in these habitats (Goulden et al., 1998, Neufeld and Mohn 2005;

Dedysh et al., 2006; Männistö et al., 2007, 2009, 2016; Lee et al., 2009; Chu et al., 2010). Even though Acidobacteria are common in various environments, only a limited number of species have been cultivated (Pankratov et al., 2008; Eichorst et al., 2011; Männistö et al., 2011, 2012). Based on the large collection of 16S rRNA gene sequences from various habitats, Acidobacteria have been divided into 26 major phylogenetic subgroups (Barns et al., 2007). Members of only seven of these subdivisions (subdivisions 1, 3, 4, 6, 8, 10, and 23) have been taxonomically described (Dedysh and Yilmaz 2018; Eichorst et al. 2018). The taxonomic classification of the phylum Acidobacteria according to the National Center for Biotechnology Information (NCBI) is shown in Table 1 (Sayers et al., 2019; Schoch et al., 2020). So far, more than 12,000 distinct phylotypes and 6,500 species-level taxonomic units have been published for Acidobacteria (Dedysh and Yilmaz 2018; Eichorst et al., 2018). When cultivated, Acidobacteria typically grow relatively slowly, and it may take up to weeks before visible colonies are developed. Cultivated acidobacterial species have been shown to be metabolically versatile and tolerate low nutrient concentrations and fluctuating conditions in soil (Rawat et al., 2012). Acidobacteria have an important role in degrading plant-derived complex carbohydrates (Eichorst et al., 2011; Pankratov et al., 2011; Rawat et al., 2012). The phylogenetic and metabolic diversity and wide distribution in a variety of habitats indicate that Acidobacteria are important in soil ecosystems (Jones et al., 2009; Faoro et al., 2010; Ganzert et al., 2011). However, the small number of isolated and characterized Acidobacteria species limits the possibility to predict their functions in soil communities, including those in permafrost-affected areas, and how they respond to the changing environmental conditions in the realm of climate change.

Table 1. Phylum Acidobacteria based on the NCBI taxonomy (Sayers et al., 2019; Schoch et al., 2020).

Class	Order	Family
Acidobacteriia	Acidobacteriales	Acidobacteriaceae
	Bryobacterales	Bryobacteraceae Solibacteraceae
	Candidatus Acidoferrales	
Blastocatellia	Blastocatellales	Blastocatellaceae Pyrinomonadaceae
Holophagae	Acanthopleuribacterales	Acanthopleuribacteraceae
	Holophagales	Holophagaceae
	Thermotomaculales	Thermotomaculaceae
Thermoanaerobaculia	Thermoanaerobaculales	Thermoanaerobaculaceae
Vicinamibacteria	Vicinamibacteriales	Vicinamibacteraceae

1.2 Viruses and their key roles in soil microbial communities

Viruses are obligate intracellular parasites which require a host organism to replicate (Fierer 2017; Kuzyakov and Mason-Jones 2018). They exist in all habitats where cellular life is found and infect all life forms from microorganisms to plants and animals (Beijerinck 1898; Suttle 2007; Campos et al., 2014; Vainio et al., 2017). The interactions between host cells and viruses vary, as viral infection may be destructive to the host or remind a symbiotic relationship (Paez-Espino et al., 2016; Pradeu 2016; Jagdale and Joshi 2018). Viruses are an integral part in any microbial community, affecting its structure and functions (Narr et al., 2017; Emerson et al., 2018; Trubl et al., 2018, Emerson et al., 2019). Bacteriophages (= phages), i.e., viruses infecting bacteria, may account for a higher rate of variation in the prokaryotic community composition than abiotic factors (Zhang et al., 2017). Viruses not only control host abundance through mortality, but can metabolically reprogramme their hosts and mediate horizontal gene transfer (Suttle 2007). By interacting with their hosts, viruses play key ecological roles on a global scale, e.g., in the regulation of global carbon cycling (Suttle 2007). In oceans, viruses lyse approximately one-third of microorganisms every day, liberating large amounts of organic compounds contained in cells (Suttle 2007). The abundance of certain viral populations has been shown to reliably predict the flux of carbon from ocean surfaces to the deep sea (Guidi et al., 2016).

Soils contain more carbon than all the vegetation and the atmosphere together, 1,500–2,400 Gtn (Lehmann and Kleber 2015), and high numbers of viruses: 10^7-10^9 virus particles per gram of soil (Williamson et al., 2005; Williamson et al., 2013). Soil viruses have also been shown to affect carbon cycling (Trubl et al., 2018; Bonetti et al., 2019, 2021). For example, a direct link between viral infection rates and the production of greenhouse gases from the decomposing microbial cells in freshwater wetlands has been shown, confirming the impact of viruses on microbial biogas production in soil (Bonetti et al., 2019, 2021). However, the roles of viruses in soils are not as clear as in marine ecosystems. Although some virus-host systems isolated from soil have been successfully characterized (Cresawn et al., 2015; Sutela et al., 2019), viral diversity in soil remains largely unexplored (Paez-Espino et al., 2016; Emerson 2019). To date, the research on soil viruses has been typically limited to direct counts and microscopy, which have given some insights into their diversity: a larger variety of viral morphotypes is observed in soil than in aquatic ecosystems, and moisture, organic matter content, pH, and abundance of microbes set limits to the number of viruses in soils (Williamson et al., 2005; Männistö et al., 2007; Narr et al., 2017; Kuzyakov and Mason-Jones 2018). Not much is known about viral infection cycles in soil, but the abundance of lysogeny and the overall virus-to-bacteria ratio have been shown to change with soil depth (Liang et al., 2020).

The physical structure of soil creates separate microhabitats which support the formation of viromes with more diversity than in other ecosystems (Paez-Espino et al., 2016; Pratama and van Elsas 2018; Emerson 2019). All microorganisms in soils, including bacteria, archaea, protozoa, algae and fungi, are infected by viruses (Pratama and van Elsas 2018; Sutela et al., 2018; Emerson 2019). The most common and diverse group of these viruses is bacteriophages. Soils also contain eukaryotic viruses hosted by plants and animals. Viruses in soils exist as free particles or inside their host cells being replicated or integrated into the genome as prophages (Kimura et al., 2008; Trubl et al., 2020). So far, soil viruses have remained largely uninvestigated due to the challenges associated with their isolation from various types of soil matrices. The recent metagenomics and viromics methods have revealed a large diversity of soil viruses, but their molecular and ecological characterization remains obscure (Emerson et al., 2018; Paez-Espino et al., 2016; Pratama & van Elsas 2018; Trubl et al., 2018, 2019). Viruses in soil can be studied with different methods, e.g., viral metagenome analysis and culturing of previously unknown viruses. For more details about culture-dependent and culture-independent approaches applied for studying soil viruses, see chapters 1.4-1.6.

1.3 Viruses of Acidobacteria

Metagenome and genome analyses have shown that there are viruses infecting Acidobacteria and acidobacterial genomes contain proviruses (Eichorst et al., 2018; Emerson et al., 2018; Paez-Espino et al., 2016; Trubl et al., 2018). Paez-Espino et al. (2016) analyzed global distribution, phylogenetic diversity, and host specificity of viruses from over 5 Tb of metagenomic sequence data representing 3,042 samples from ten habitat types classified as marine, freshwater, non-marine saline and alkaline, thermal springs, terrestrial soil, terrestrial others (e.g., deep subsurface samples), host-associated human, host-associated plants, host-associated others (e.g., host animal-associated other than human), and engineered (e.g., bioreactor). CRISPR spacers and transfer RNA matches were used to link viral groups to their microbial hosts. The analysis identified 9,992 putative host–virus associations. From these massive data, only one metagenomic viral contig was assigned to the acidobacterial host (Paez-Espino et al., 2016).

Trubl et al. (2018) described viral populations from the Stordalen Mire site in northern Sweden and compared their ecology along the permafrost thaw gradient. Viruses were characterized from viromes derived from separated viral particles. The method used for virus particle extraction and purification had been specifically optimized for the acidic peat soils rich in phenolic compounds by Trubl et al. (2018). They used a gene-sharing network method indicating similar gene contents (Lima-Mendez et al., 2008) for taxonomic classification of the viral sequences. Seventeen of the 53 described viral taxonomic units (vOTU, corresponding to approximately species-level taxonomy) were linked to

microbial hosts. Four of the tentative microbial host species were identified and two of them belonged to Acidobacteria, namely *Acidobacterium* (the host for seven putative viruses) and *Candidatus* Solibacter usitatus (the host for three putative viruses). The analyses showed habitat specificity of the soil viruses along the thaw gradient, infection of key C-cycling microbes and the carriage of host metabolic genes (Trubl et al. 2018).

Emerson et al. (2018) analyzed 197 bulk metagenomes along the same permafrost thaw gradient in Stordalen Mire across three types of peat soils and recovered the total of 1,907 viral populations from them. In that work, 1,529 bacterial and archaeal genomes from the same metagenomes were screened for genomic features to link viruses and their hosts. The analysis showed 230 viruses putatively linked to Acidobacteria. As these viral populations were from bulk soil DNA, they were presumed to represent free viruses, proviruses and/or actively infecting viruses. Emerson and colleagues (2018) assessed potential viral effects on host ecology by analyzing how viral infection dynamics for specific host lineages varied across the three permafrost thaw habitats: palsa, bog and fen. No progressive patterns across all three habitats were found, only some general ones. A decline in virus/host abundance ratio with increasing thaw was observed, as viruses of the Solibacteres were less abundant than their hosts in palsa and bog and more abundant in fen. For Acidobacteriia, the virus/host abundance ratios were invariable among the three habitats along the thaw gradient, as both virus and host abundances increased from palsa to bog. The virus/host abundancies for Acidobacteriia significantly correlated with pH and especially with dissolved organic carbon concentrations. This is in accordance with the function of Acidobacteriia as the primary degraders of large polysaccharides in palsa and bog habitats (Woodcroft et al., 2018; Kalam et al., 2020).

A large-scale genome analysis of Acidobacteria (24 genomes representing subdivisions 1, 3, 4, 6, 8 and 23) was performed to explain the wide occurrence of Acidobacteria in soils and to understand their ecophysiology (Eichorst et al., 2018). Mobile genetic elements are expected to mediate horizontal gene transfer and thus aid in the evolution and ecological success of Acidobacteria by introducing new metabolism-relevant genes across the species (Summers et al., 2005; Challacombe and Kuske 2012). Therefore, bacteriophage integration events were one of the traits specially studied in the genome analysis (Eichorst et al., 2018). The identification of sequences as prophages was based on high concentration of unknown genes and a genome organization consistent with a phage genome. The analysis by Eichorst and colleagues (2018) identified 35 putative prophages in 19 of the 24 acidobacterial genomes studied. Most of these genomes originated from soils. Prophages were not detected in Acidobacteriaceae bacterium KBS 146 (subdivision 1) and strains isolated from extreme environments, such as geothermal soils, hot springs and microbial mats. Twenty-nine of the putative

prophages had at least one virion-associated gene, which was taken to indicate that they still have the potential to complete a lytic cycle. According to this definition, *Granulicella tundricola* MP5ACTX9 contained one prophage, while *Granulicella mallensis* MP5ACTX8 and *Terriglobus saanensis* SP1PR4 both contained two prophages that were likely active (Eichorst et al., 2018). These three strains have been isolated from acidic tundra soils in Kilpisjärvi by Männistö et al. (2011, 2012), and genomic analyses have confirmed their significant role in organic carbon processing (Rawat et al., 2012). The study by Eichorst et al. (2018) indicated a high level of polylysogeny, as several bacteria had more than one prophage in the genome. Eight acidobacterial genomes had multiple genes associated with virions. Seven genomes, including *Granulicella mallensis* MP5ACTX8 and *Terriglobus saanensis* SP1PR4, had genes likely to provide resistance to superinfection, i.e., infection of cells already infected by another virus. The high level of polylysogeny and conservation of the superinfection-preventing genes indicate an intense viral pressure on soil Acidobacteria (Eichorst et al., 2018).

The 35 putative prophages identified by Eichorst et al. (2018) were not similar to any known phages. Based on the capsid-related genes identified, the prophages were classified into the order *Caudovirales*, which contains tailed double-stranded DNA bacterial and archaeal viruses. The prophages were further analyzed by clustering with the gene-content based classification method of Lima-Mendez et al. (2008). The clustering showed 12 prophages as singletons, 8 prophages clustered only with other acidobacterial prophages, and 15 clustered also with previously identified prophages from publicly available microbial genomes (Roux et al., 2015b) or soil metagenomes (Paez-Espino et al., 2016). The observed two clusters were of approximately subfamily level. One of these clusters consisted of prophages from five different Acidobacteria genomes, including *Granulicella tundricola* MP5ACTX9 and *G. mallensis* MP5ACTX8, as well as prophages from an *Alphaproteobacterium* and *Chloroflexi*, and from Iowa native prairie soil metagenome. The other cluster had prophages from six acidobacterial genomes, including *Terriglobus saanensis* SP1PR4 and other *Terriglobus* species, as well as prophages from Arctic peat soil metagenome from Alaska, and bog forest soil metagenome from Canada (Paez-Espino et al., 2016; Eichorst et al., 2018).

Despite the molecular studies showing that there are viruses infecting different species of Acidobacteria, no acidobacterial viruses have been isolated so far (Paez-Espino et al., 2016; Emerson et al., 2018; Trubl et al. 2018; Eichorst et al., 2018). The isolation and detailed characterization of such viruses would provide valuable insights into viral diversity and virus-host interactions in soil, as well as relations between different viral groups in general.

1.4 Methods for virus isolation from soil

A wide variety of methods have been used for virus isolation from soil depending on the aim of the research (Kimura et al., 2008; Göller et al., 2020; Trubl et al., 2020). The isolation of a virus is limited by the ability to grow the host in pure culture, while uncultured microbes dominate in all the diverse environments studied (Lloyd et al., 2018). Moreover, the initial isolation host is not always the most optimal for the virus, as the parameters of viral infection cycle in different host strains may vary (Howard-Varona et al., 2017; Enav et al., 2018). Extraction can produce intact virus particles that are inactivated, e.g., by losing the tail (Williamson et al., 2012). For plaque formation, both an infective virus and a suitable host culture are needed, whereas inactivated viruses can be used for metavirome analysis (see below), if their genomes are intact (Trubl et al., 2020). Culture-independent methods are used, e.g., to study the diversity of viruses in different habitats or spatial and seasonal changes in the viral of communities (Nakayama et al., 2007; Williamson et al., 2013; Zablocki et al., 2014; Ballaud et al., 2015; Trubl et al., 2018; Göller et al., 2020). Cultured virus-host systems can be used for various analyses that are impossible to perform in a comparable detail by culture-independent methods: studying viral infection cycle, molecular details of viral replication, stability of virus infectivity under different conditions, detailed structural studies of virion organization, experimental determination of gene functions and molecular functions that characterize virus physiology and virus-host relationships (Trubl et al., 2020).

The soil type is a major factor affecting the choice of the isolation method (Kimura et al., 2008; Paez-Espino et al., 2016; Williamson et al., 2017; Pratama and van Elsas 2018; Trubl et al., 2018). Several studies have shown that the extraction of virus particles needs to be optimized for each soil type: e.g., the amended citrate buffer developed for peat soils by Trubl et al., (2016) gave only 5% phage recovery with the sandy agricultural soil samples used by Göller et al., (2020), while the best phage recovery (67%) in the latter study was obtained with 10% beef extract. The soil type affects adsorption of viruses to the soil particles, as more than 90% of soil viruses are estimated to be adsorbed to the soil matrix (Hurst et al., 1980; Kimura et al., 2008). Viruses have a pH-dependent surface charge in water and other polar media, which determines their mobility and sorption behaviour (Michen and Graule 2010). To desorb viruses from soil particles, various chemical reagents and physical dispersal methods have been used (Williamson et al., 2003, Zablocki et al., 2014, Pratama and van Elsas 2018; Trubl et al., 2019). Williamson et al. (2003) and Göller et al. (2020) have used pure cultures of viruses added to the soil samples to test the extraction efficiency of different buffers. The efficiency of extraction varied from 0.5% to 66.7% for the different phages and depended largely on the extraction buffer and characteristics of the phages used. Most of the specific isolation methods have been

developed for bacteriophages, as they are a common group of viruses in soils (Williamson et al., 2017; Pratama and van Elsas 2018).

Enrichment is one of the basic methods for phage isolation (Hyman 2019). In this approach, bacteria are mixed with the environmental sample and the mixture is incubated for some time. After incubation, the remaining bacteria are removed by centrifugation and/or filtration, and phages are collected from the resulting suspension. Viruses have also been separated directly from wet soils: Ballaud et al. (2015) analyzed spatial and seasonal changes in bog virome by pressing pore water from the soil samples, concentrating the viruses with polyethylene glycol and purifying them by filtration. A number of different solutions have been used for the extraction of viruses from soils: deionized water (Zablocki et al. 2014), saline magnesium buffer (Narr et al. 2017, Göller et al. 2020), 10% beef extract (Williamson et al., 2003), glycine (Williamson et al., 2003, 2005), 1% potassium citrate (Williamson et al., 2003) and amended 1% potassium citrate (Trubl et al., 2016; Göller et al., 2020), as well as different phosphate buffers (Williamson et al., 2003; Quiros and Muniesa 2017). The salts in the buffers are needed to stabilize pH and the virus particles, whereas amino acid or protein (bovine serum albumin or beef extract) are added to bind viruses and disrupt their interactions with soil (Trubl et al., 2016). Various mechanical treatments like vortexing (Williamson et al., 2003; Trubl et al., 2016; Narr et al., 2017), sonication (Williamson et al., 2003; Trubl et al., 2016) or bead-beating (Williamson et al., 2013; Trubl et al., 2016) are used to detach viruses from soil particles. Soil material is removed from the suspension by centrifugation or filtration, and the virus particles are collected. Filtration through membranes of 0.22–0.45 µm pore size is commonly used for the separation of free viruses from cells (Kimura 2008). Selection of the filter material is important because viruses can adsorb to filter membranes (Tartera et al., 1993). Filtration can result in even two-third reduction of viruses in the extracts (Paul et al., 1991). Sequential re-extraction of the soil sample results in more complete elution (Williamson et al., 2005). As shown by Göller et al. (2020), the recovery of bacteriophages added to the sample increased from 46% to 67% by resuspending the soil pellet twice instead of only once. The number of isolated viruses can be determined by titration (e.g., plaque assay), epifluorescence microscopy, transmission electron microscopy (Pratama and van Elsas 2018; Trubl et al., 2020) or flow cytometry (Ballaud et al., 2015). For the cultivation of virus-host pairs and quantification of infectious virions in a given sample, the plaque assay method is commonly used. The host culture is infected with a serially diluted virus sample and incubated on an agar plate overlayed with a soft top agar. The number of plaques on a homogeneous bacterial lawn is counted after incubation and used for calculating the number of infectious viruses in the sample.

1.5 Metagenomics as a powerful tool to study microbial communities

Soil is one of the least understood habitats on the Earth (Handelsmann 1998; Howe et al., 2014; Kalam et al., 2020). The last 25 years of research have verified that culturing is a slow and laborious method to learn a lot about a very small number of the microorganisms present in the environment. Only a fraction of the microorganisms in soil are readily cultured using current techniques, and, at the same time, the soil microbiota is shown to contain unique metabolic potential and amazing genetic diversity (Fierer et al., 2007). Thus, microbial communities contain a vast number of species, most of which have never been cultured or identified (Howe et al., 2014; Trubl et al., 2018). Metagenomics is the study of the total genetic material in a defined environment, and the total pooled gene content of a microbial community forms the community metagenome (Handelsman 1998; Schloss and Handelsman 2003). As metagenomics involves the extraction of the collected DNA of all species in a sample, it differs greatly from the traditional DNA isolation analysis, which is usually performed using a pure culture of a microbe clone. The isolated metagenomic DNA is broken up into numerous small fragments and sequenced. This untargeted sequencing of all microbial genomes in a sample is called shotgun sequencing (Quince et al., 2017). The resulting sequences are analyzed, and the microbial genomes are reconstructed. Metagenomics as a method does not require isolation or culturing, and it has been of great value in extending our understanding of microbial communities in various environments (Howe et al., 2014; Quince et al., 2017; Emerson et al., 2018; Trubl et al., 2018, Roux 2019).

Metagenomic analysis of environmental DNA can give information about the abundance and distribution of specific microbial taxa in different environments (Howe et al., 2014; Hultman et al., 2015; Delmont et al., 2018). Metatranscriptomics, i.e., the analysis of total RNA, allows to identify active genes and changes in gene activity in changing environmental conditions (Hultman et al., 2015; Emerson et al., 2018). Thus, it can also help to understand the functioning of the communities and changes caused by environmental upheavals (Luo et al., 2014; Männistö et al., 2016; Emerson et al., 2018). These methods are so efficient that they can be performed even on single cells (Bankevich et al., 2012).

Many environmental samples are very complex to analyze, and getting complete genome assemblies might be challenging (Delmont et al., 2011; Howe et al., 2014; Delmont et al., 2018). The DNA extraction is a key step for successful metagenomic analysis (Fierer 2017, Trubl et al., 2019). In soil, the composition and physicochemical properties of soil particles and aggregates affect the adhesion of materials on their surface and how these materials can be extracted. As these properties vary greatly in different types of soils, different DNA extraction methods have been developed (Trubl et al., 2019,

2020; Göller et al., 2020). After the DNA extraction yield and purity is optimized, the sequencing, library construction and bioinformatics methods can all significantly affect the results of the analysis (Trubl et al., 2019, 2020). A variety of bioinformatics tools are currently available to meet the specific needs of metagenomics-based analyses, including the software for analyzing viral sequences (Roux et al., 2015a; Bolduc et al., 2017; Ren et al., 2017).

1.6 Metagenomics used to unravel soil virus mysteries

Two main approaches are used for studying viral DNA in soils: viral sequences can be obtained directly from virus particles extracted from soil, or they can be selected from metagenomes where all DNA was sequenced from a soil sample (Trubl et al., 2020). Virus particles can be extracted from soil with the same methods and limitations that apply to isolating viruses for *in vitro* culture (see section 1.4, Kimura et al., 2008; Göller et al., 2020; Trubl et al., 2020). Viromes, i.e., the total genetic material of (free) viruses in a given environment, are produced from the extracted virus particles separated from microbial cells. Due to the prior removal of cellular DNA, the use of isolated viruses gives increased coverage specifically for viral genomes in comparison to the use of total DNA metagenomes, but the sampling excludes proviruses (Trubl et al., 2020). Both methods have some common disadvantages: sample preparation and the methods used to extract and amplify DNA are biased, and the results are affected by the choice of bioinformatic tools (Trubl et al., 2020).

Soil viral communities are often expected to be dominated by bacteriophages, as the analysis methods are biased to their recovery (Emerson 2019). However, plant and animal viruses and especially mycoviruses can have a major role in some soil types, if the high abundance of fungi is taken as an indication of the abundance of mycoviruses (Sutela et al., 2019). The viral genomes can be either DNA or RNA, double-stranded or single-stranded. Almost all fungal and oomycete viruses have genomes composed of double-stranded or single-stranded RNA (Zheng et al., 2014; Sutela et al., 2018). Several studies have indicated RNA viruses may outnumber DNA viruses in some cases, indicating their importance in these ecosystems (Shi et al., 2016; Stough et al., 2018; Starr et al., 2019). At the moment, most viral analyses are performed using DNA metagenomes (Emerson 2019), and specific methods for viral metagenome analysis for single-stranded and double-stranded DNA viruses from different soil types are being developed (Trubl et al., 2019).

Metagenomics-based research on soil viruses has been more challenging than that for viruses in marine environments. As viral biomass in soil is relatively low, the yield of viral DNA is also typically low, which complicates the assembly of reads, and a low number of viral contigs is produced, as genome assembly requires a massive sequence library to start with (van der Walt et al., 2017; Trubl et al., 2019). The data processing requires computational resources and takes time (Peltola et al.,

1984; Thomas et al., 2012; Quince et al., 2017; van der Walt et al., 2017). The lack of a universal viral marker gene, such as the 16S rRNA gene used for bacterial phylogenetics, or any substituting method to make viral taxonomy surveys has hindered the direct studies of virus dynamics in ecosystems (Trubl et al., 2020). As metagenomics is based on comparisons to reference libraries, it does not need universal marker genes for the phylogeny analysis. However, identified viruses mostly belong to unidentified taxons, as only a small number of soil viruses have been characterized and are found in the reference databases needed for identification (Emerson et al., 2018; Trubl et al., 2018). Due to the mentioned challenges, there is a limited number of thorough metagenomics-based characterizations of the ecological significance of virus-host interactions in soil (Emerson et al., 2018; Trubl et al., 2018, 2021; Wu et al., 2021). The metagenomics-based projects to characterize the role of viruses in permafrost peatlands have provided new insights into the role of viral populations in the Arctic permafrost ecosystems (Emerson et al., 2018; Trubl et al., 2018). The results revealed habitat specificity of viral communities, a shift from soil-like to aquatic-like community identity along the thaw gradient, infection and lysing of dominant microbial hosts, and carriage of auxiliary metabolic genes (AMGs) (Emerson et al., 2018). The identified AMGs suggested virus-mediated adjustments in host carbon metabolism, soil organic matter degradation, binding of polysaccharide, and regulation of sporulation (Trubl et al., 2018). All these factors suggest a major impact of viral populations on the ecosystem carbon cycling.

Environmental factors have been demonstrated to cause significant changes in the composition and functions of soil viral populations *in vitro*. Wu et al. (2021) showed increased viral activity in wet prairie soil, and Trubl et al. (2021) revealed that viruses can continue to infect and replicate in Arctic peat soils below freezing temperatures. Wu et al. (2021) measured activity of DNA and RNA viruses in water-saturated and air-dried soil samples after the incubation for 15 days using the combination of metagenomics, metatranscriptomics and metaproteomics. Differences in soil moisture changed the activities of both DNA and RNA viruses. Most of the transcriptionally active DNA viral contigs were unique to either wet or dry treatments. The number of transcribed DNA viral contigs was higher in dry soils, but the levels of transcriptional activity were significantly higher for DNA viruses in wet soils. Of the putative DNA virus-host pairs, 44% were unique to the dry soil treatment, 28% were detected only in the wet soil, and 28% were found in both dry and wet treatments. The total RNA viral abundances were strongly correlated with the abundance of active eukaryotic species, especially in wet soils (Wu et al. 2021). Trubl et al. (2021) studied virus-host interactions in Alaskan peat soil in simulated winter conditions using stable isotope probing metagenomics. Peat samples were incubated at –1.5 °C in anoxic conditions for 184 and 370 days using heavy water (water containing

deuterium, i.e., heavier hydrogen isotope, instead of the common hydrogen) to label actively replicating microbes and viruses. Metagenome analysis of the samples revealed 46 bacterial populations and 243 viral populations that actively took up heavy water and produced CO_2. The most abundant active bacterial populations belonged to Acidobacteriota, Bacteroidota, and Firmicutes. Notably, active bacterial populations represented only a small portion of the microbial community detected in the peat soil, while active viral populations represented a large portion of the detected viral community. Lysogeny was common in the samples, being probably linked to low host abundances and harsh environmental conditions. According to Trubl et al. (2021), the large number of active virus-host interactions in sub-freezing anoxic conditions emphasizes the potential that viruses have in modulating soil microbial communities and the significant carbon losses in the Arctic during the long winters.

To conclude, both culture-dependent (isolation and culturing) and metagenomics-based methods are powerful tools, each with their own advantages and limitations. Their combination can provide the most comprehensive information about the diversity of viruses and their roles in regulating the functions of microbial communities in various environments.

2. Aims of the study

Viruses thrive in all microbial communities, affecting their functions and development (Emerson et al., 2018). Microbial communities largely consist of species that have never been cultured or identified. Soil contains high numbers of virus particles, but soil viruses remain understudied (Swanson et al., 2009). This research is aimed to characterize viruses residing in permafrost-affected soils (i) by isolating viruses infecting Acidobacteria from Kilpisjärvi Arctic tundra soil, and (ii) by analyzing metagenomes from the same sampling site. Research outcomes are expected to contribute to better understanding of virus-host interactions in Arctic soils.

3. Materials and methods

3.1 Isolation and characterization of viruses

3.1.1 Field site and collection of soil samples

The summer soil samples were collected in July 2018 and 2019 for metagenome analyzes (Table 2) (Viitamäki 2019; Pessi et al., 2020). The research site in Malla nature reserve is located in the oroarctic mountain tundra area in Kilpisjärvi in the northwestern Finland (69.04°N, 20.79°E). In this study area, main vegetation types are classified as barren soils, heathlands, meadows, and fens. Soil cores were collected using a soil corer sterilized with 70% ethanol. The cores were divided into organic and mineral subsamples when both soil types were available. Samples were spooned in Whirl-pak bags, immediately frozen in dry ice or liquid nitrogen and stored at –80 °C until further analyses.

Winter samples (Table 2) were collected for the virus isolation in March 2021. Snow was removed with a spade, and the frozen plant material was removed. Samples were chiseled from the soil surface and spooned in ziplock bags and stored at 4 °C until further analyses. All tools were sterilized with 70% ethanol.

Table 2. Soil samples used in this study.

Soil sample	Sampling date	Vegetation type	Sampling depth, cm
o12218	July 2018	Graminoid	5–10
o12212		Fen	5–10
o12205	July 2019	Graminoid	5–10
o12209		Fen	5–10
o12215		Fen	5–10
o12216		Fen	5–10
o12204		Fen	5–10
o25	April 2021	Deciduous shrub	1–2
o37		Evergreen shrub	1–2
o109		Deciduous shrub	1–2
o115		Deciduous shrub	1–2
o181		Graminoid	1–2
o193		Graminoid	1–2
o427		Evergreen shrub	1–2
o577		Evergreen shrub	1–2
o12217		Fen	1–2
o12222		Graminoid	1–2
o1075		Deciduous shrub	1–2
o733		Evergreen shrub	1–2

3.1.2 Acidobacteria host strains

The 16 Acidobacteria strains used as potential hosts in virus isolation were kindly provided by Minna Männistö, Natural Resources Institute Finland, Oulu (Table 3). The host strains were grown on DSMZ medium 1284, containing 0.5 g glucose, 0.1 g yeast extract (Neogen, Lansing, USA), 0.1 g casamino acids (MP Biomedicals,Solon, USA), 0.04 g $MgSO_4 \times 7\ H_2O$, and 0.02 g $CaCl_2 \times 2\ H_2O$ per 1 l of distilled water (www.dsmz.de/microorganisms/medium/pdf/ DSMZ_Medium1284.pdf). Fifteen g and 4 g of agar were added per 1 l for plates and top-layer agar, respectively, and pH was adjusted to 5.5. The strains were incubated with aeration at room temperature (RT).

Table 3. Acidobacteria host strains for virus isolation.

Host strain no.	Species	Strain code	Reference
1	*Granulicella sapmiensis*	MP7CTX5	Männistö et al., 2012a
2	*Granulicella arctica*	MP5ACTX2	Männistö et al., 2012a
3	*Granulicella* sp.	X4BP1	unpublished
4	*Edaphobacter* sp.	M8UP27	unpublished
5	*Granulicella mallensis*	MP5ACTX8	Männistö et al., 2012a
6	*Edaphobacter* sp.	MP8S11	unpublished
7	*Edaphobacter* sp.	X5P2	unpublished
8	*Edaphobacter* sp.	M8UP28	unpublished
9	*Granulicella* sp.	X5P3	unpublished
10	*Edaphobacter* sp.	M8US30	unpublished
11	*Granulicella* sp.	MP8S9	unpublished
12	*Edaphobacter* sp.	M8UP15	unpublished
13	*Acidicapsa* sp.	MP8S7	unpublished
14	*Granulicella tundricola*	MP5ACTX9	Rawat et al., 2014
15	*Granulicella* sp.	M8UP17	unpublished
16	*Edaphobacter* sp.	M8UP30	unpublished

3.1.3 Virus isolation

Only organic layer soil samples were used for virus isolation. Viruses were eluted from the soil samples using three buffers: DSMZ medium 1284, phosphate-buffered saline (PBS) and protein supplemented PBS (PPBS: 2% BSA, 10% PBS, 1% potassium citrate, 150 mM $MgSO_4$) (Göller et al., 2020). Five protocols for the isolation of viruses from the soil samples were tested (Table 4). The variables tested were elution buffer, volume, time and temperature, centrifugation speed, time and temperature, as well as the pretreatment of soil samples. After incubating soil with a buffer, soil particles were removed by centrifugation, and the supernatant was filtered using LLG Syringe Filters Spheros filters with 220 nm pore size. The filtered or both filtered and non-filtered supernatants were used for plaque assay with a selection of host strains (Table 5). 100–150 µl of supernatant and 300 µl of fresh liquid host culture (exponential or early stationary phase) were mixed with 3 ml of top-layer

agar (46 °C) and added on top of plates. The plates were incubated aerobically at RT and regularly checked for the presence of viral plaques. Single plaques were picked up, resuspended in broth and subjected to plaque assay, which was repeated sequentially three times to ensure the purity of virus isolates.

Table 4. Virus isolation protocols 1–5.

Protocol	**1**	**2**	**3**	**4**	**5**
Elution buffer	PBS	1284 DSMZ	PPBS and 1284 DSMZ	1284 DSMZ	1284 DSMZ host suspension
Elution volume	1:3	1:3	1:1	1:5	1:10
Shaking: time, temp.	30 min, RT	o.n.[a], 5 ˚C	manual 10 min and o.n. at 4 ˚C	7 days, RT	7 days, RT
Centrifugation: g, time, temp.	2,500 g, 30 min, RT	2,500 g, 30 min, RT	10,000 g, 2×30 min, 4 ˚C	10,000 g, 10 min, RT	10,000 g, 10 min, RT
Filtration, 220 nm	+, -	+, -	+	+	+

a. o.n., overnight

Table 5. Host strains used for the soil samples in the isolation protocols 1–5.

Protocol	**1**		**2**		**3**		**4 and 5**	
Year[a]	Sample	Host[b]	Sample	Host	Sample	Host	Sample	Host
2018			o12218	1–11				
			o12212	1–11				
2019	o12205	1–6	o12216	1–11	o12215	1–7	o12211	1–7
			o12204	1–11	o12217	1–7		
			o12215	1–7	o12218	1–7		
			o12209	1–7				
			o12205	1–7				
2021	o193	1–11						
	o181	1–16						
	o427	1–11						
	o577	1–11						
	o12222	1–11						
	o12217	1–11						
	o25	1–11						
	o109	1–11						
	o37	1–11						
	o115	1–11						
	o733	1–16						
	o1075	1–16						

a. Sample collection year.
b. Numbers refer to Table 3.

3.1.4 Preparation of virus agar stocks

Semiconfluent plates were used for stock preparation. The top agar layers were collected and suspended in broth (3 ml per plate), incubated for 1 h (150 rpm, RT) and centrifuged (10,000 g, 30 min, 4 °C). The supernatant was collected and filtered (220 nm). Virus stocks were titrated by plaque assay and stored at 4°C.

3.1.5 Virus DNA extraction

Genomic DNA was extracted from virus stocks using Thermo Scientific GeneJET Genomic DNA purification Kit (modified from Santos et al., 1991). Virus stock (7.2 ml) was divided into four 1.8 ml aliquots and mixed with 4 µl of DNase I (0.5 mg/ml) and 20 µl of RNase A (10 mg/ml) and incubated for 30 min at 37 °C. Freshly prepared 2 M $ZnCl_2$ was added to the mixture (40 µl per aliquot) and incubated for 5 min at 37 °C. After centrifugation (10,000 g, 1 min), the supernatant was discarded and the pellet was resuspended by pipetting gently once or twice in 1 ml TES buffer [0.1 M Tris-HCl, pH 8; 0.1 M EDTA; 0.3% (w/v) SDS] and incubated at 60 °C for 15 min, being mixed two times during the incubation by turning the tube gently. Proteinase K (20 mg/ml) was added (40 µl per aliquot) and incubated for 90 min at 37 °C. The kits's lysis solution and 70% EtOH were mixed 1:1 and added to the mixtures (1 ml per aliquot). The samples were loaded in 800 µl batches to the GeneJET Genomic DNA Purification Column. The column was centrifuged (1 min, 6,000 g), the flow-through solution was discarded, and a new batch was loaded in the column until all of the samples have been put through the same purification column. The column was washed with 500 µl of Wash Buffer I (centrifuged 1 min, 8,000 g) and 500 µl of Wash Buffer II (centrifuged 3 min, 14,000 g). DNA was eluted by adding 50 µl of Qiagen elution buffer (AllPrep DNA/RNA Mini Kit, Qiagen, Hilden, Germany) to the column, incubating for 2 min at RT and centrifugating (1 min, 8,000 g). The flow-through was loaded back to the column, incubated for 2 min at RT, and centrifuged (1 min, 8,000 g). The concentration and purity of DNA were determined using Nanodrop (ThermoFisher Scientific, Waltham, MA, USA). The eluted DNA was stored at –20 °C.

3.1.6 Sequencing and genome analyses

Nextera sequencing library was made from the extracted viral DNA and sequencing was conducted with Illumina MiSeq at the DNA Sequencing and Genomics Laboratory, Institute of Biotechnology, University of Helsinki. Reads were trimmed and adaptors removed with Cutadapt using Phred quality score of 30 and trimming length of 50 (Martin 2011). The assembly was done using Spades v. 3.15.0 (Bankevich et al., 2012). The sequences were handled and analyzed using Geneious Prime 2021.2.2 (https://www.geneious.com). ORFs were predicted using Glimmer3 v. 1.5 (Delcher et al., 2007) and annotated using Blastx searches against NCBI non-redundant protein sequences database with e-

value threshold of 0.001, searches performed in August 2021 (Altschul et al., 1990). Viral genomes were compared to each other using Emboss stretcher (Rice et al., 2000) and Circoletto (Darzentas 2010). Putative classification of viruses was performed using Virfam (Lopez et al., 2014).

3.2 Metagenomics

3.2.1 DNA extraction and sequencing

The total DNA isolated from the soil sample o12217 (Kilpisjärvi, July 2018) was used for the metagenomic analysis. The DNA was extracted with a modified bead beating protocol (DeAngelis et al., 2009; Griffiths et al., 2000) as previously described by Viitamäki (2019), and the library for Illumina metagenome sequencing was prepared (Nextera XT DNA Library Preparation Kit, Illumina, San Diego, CA, USA). Metagenomes were obtained across two paired-end NextSeq (132–170 bp) and one NovaSeq (2 x 151 bp) runs. Sequence reads were trimmed by removing primers, short reads and low-quality reads with Cutadapt (Martin 2011).

3.2.2 Assembly of metagenome

The assembly of trimmed reads was performed either by MEGAHIT (Li et al., 2015 and Li et al., 2016) within the Lazypipe pipeline (Plyusnin et al., 2020) or by metaSPAdes v. 3.15.0 (Nurk et al., 2017). The quality of assembled metagenome was assessed using MetaQUAST (Mikheenko et al., 2015). The volume of sequencing data that was used for the assemblies, i.e., the percentage of reads that were mapped to the contigs (alignment rate) was analyzed using Bowtie2 (Langmead & Salzberg 2012). Completeness of obtained contigs was assessed using Samtools (Danecek et al., 2021).

3.2.3 Recovering and annotating viral contigs and taxonomic assignments

Within Lazypipe, i.e., using MEGAHIT assembly, gene-like regions in the assembled contigs were scanned for using MetaGeneAnnotator (Noguchi et al., 2006, 2008), homology search was conducted using Centrifuge (Kim et al., 2016), and viral contigs were annotated with Blastn (Altschul et al., 1990). The contigs obtained with metaSPAdes were subjected to the What-the-Phage pipeline (Marquet et al., 2020), which used several virus prediction programs that run in parallel: VirFinder v. 1.1 (Ren et al., 2017), PPR-Meta v. 1.1 (Fang et al., 2019), VirSorter v. 1.0.6 (Roux et al., 2015a), Metaphinder (Jurtz et al., 2016), Sourmash v. 2.0.1 (Brown and Irber 2016), Vibrant v. 1.2.1 (with and without virome mode) (Kieft, Zhou and Anantharaman 2020), Phigaro v. 2.2.6 (Starikova et al., 2020), Virsorter 2 v. 2.0.beta (Guo et al., 2021) and Seeker (Marquet et al., 2020). The pipeline also included annotation and taxonomic assignment using Prodigal v. 2.6.3 (Hyatt et al., 2020) and hmmer v. 3.3.2 (Eddy 1995). The quality and completeness for the “phage-positive” contigs was assessed by

CheckV (Nayfach et al., 2021) within the pipeline. The subset of phage-positive contigs containing at least one viral gene was retained to avoid false positive results.

3.2.4 Analyzes of taxonomic relations

From What-the-Phage output, the subset of viral contigs that were at least 10 kbp long or at least 50% complete were selected for the taxonomic analysis using vConTACT2 (Jang et al., 2019) at the Cyverse platform (https://de.cyverse.org/). Genes were predicted using Prodigal (Hyatt et al., 2020), the resulting protein sequences file was used in vConTACT2-Gene2Genome (Jang et al., 2019) for the creating of gene-to-genome mapping file, which was then used in vConTACT2 (Jang et al., 2019) with NCBI Bacterial and Archaeal Viral Refseq V201 database within International Committee on Taxonomy of Viruses+NCBI taxonomy. The resulting network file was visualized in Cytoscape v. 3.8.2. (Shannon et al., 2003).

4. Results

4.1 Five virus isolates infecting soil Acidobacteria were obtained from Arctic soil samples

Sixteen acidobacterial strains belonging to the genera *Granulicella*, *Edaphobacter* and *Acidicapsa* (Table 3) were used in the attempts to isolate viruses from the Kilpisjärvi Arctic tundra soil samples (Table 2). In total, six phage isolates were obtained from soil samples collected at Kilpisjärvi in April 2021 (Table 5). Four of these isolates were obtained on *Edaphobacter* sp. X5P2 (named here EV1, EV2, EV3 and EV4), one on *Edaphobacter* sp. M8UP27 (EV5) and one on *Granulicella* sp. X4BP1 (GV1). The viruses formed plaques after five to eight days of incubation. All plaques were clear and round, but their sizes differed. Plaques formed by EV1 and EV5 were about 5 mm in diameter, while plaques formed by GV1, EV2, EV3 and EV4 were smaller, about 1-2 mm in diameter (Fig. 1). Titers of the virus stocks ranged from 8.6×10^8 to 6.6×10^9 PFU/ml. The genome analysis showed that five of the isolates were unique, whereas EV3 and EV4 genomes were identical (chapter 4.2.). Thus, five different isolates were obtained.

Table 6. Virus isolates obtained from the Kilpisjärvi winter 2021 soil samples.

Soil sample	Host[a]	Virus isolate	Stock titer, PFU/ml	Plaque morphology
o181	Ed. X5P2	EV1	2.0×10^9	clear
o12217	Ed. X5P2	EV2	3.7×10^9	clear
o12217	Ed. X5P2	EV3[b]	6.6×10^9	clear
o12217	Ed. X5P2	EV4[b]	2.4×10^9	clear
o12222	Ed. M8UP27	EV5	4.0×10^9	clear
o12222	G. X4BP1	GV1	8.6×10^8	clear

a. Ed., *Edaphobacter* sp., G., *Granulicella* sp.,
b. Same virus (based on the genome analysis, see below).

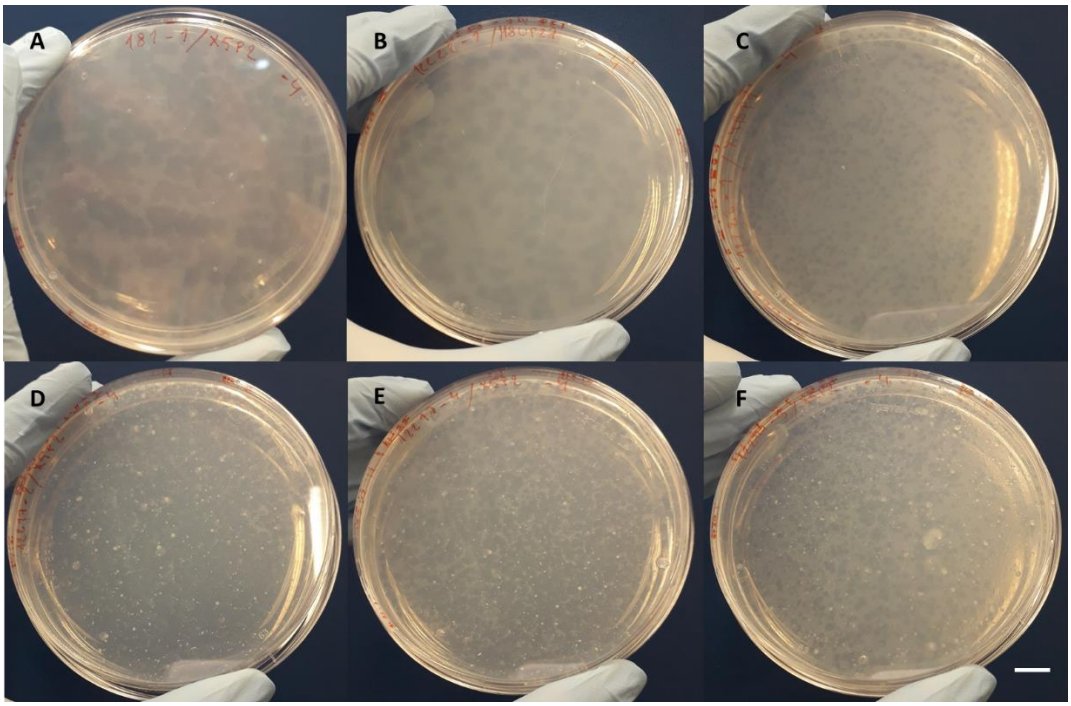


Figure 1. Plaque morphology of the six virus isolates from the Kilpisjärvi 2021 soil samples: (A) EV1, (B) EV5, (C) GV1, (D) EV2, (E) EV3, (F) EV4. Scale bar 1 cm for all sections.

4.2 The five acidobacterial virus genomes are dsDNA molecules

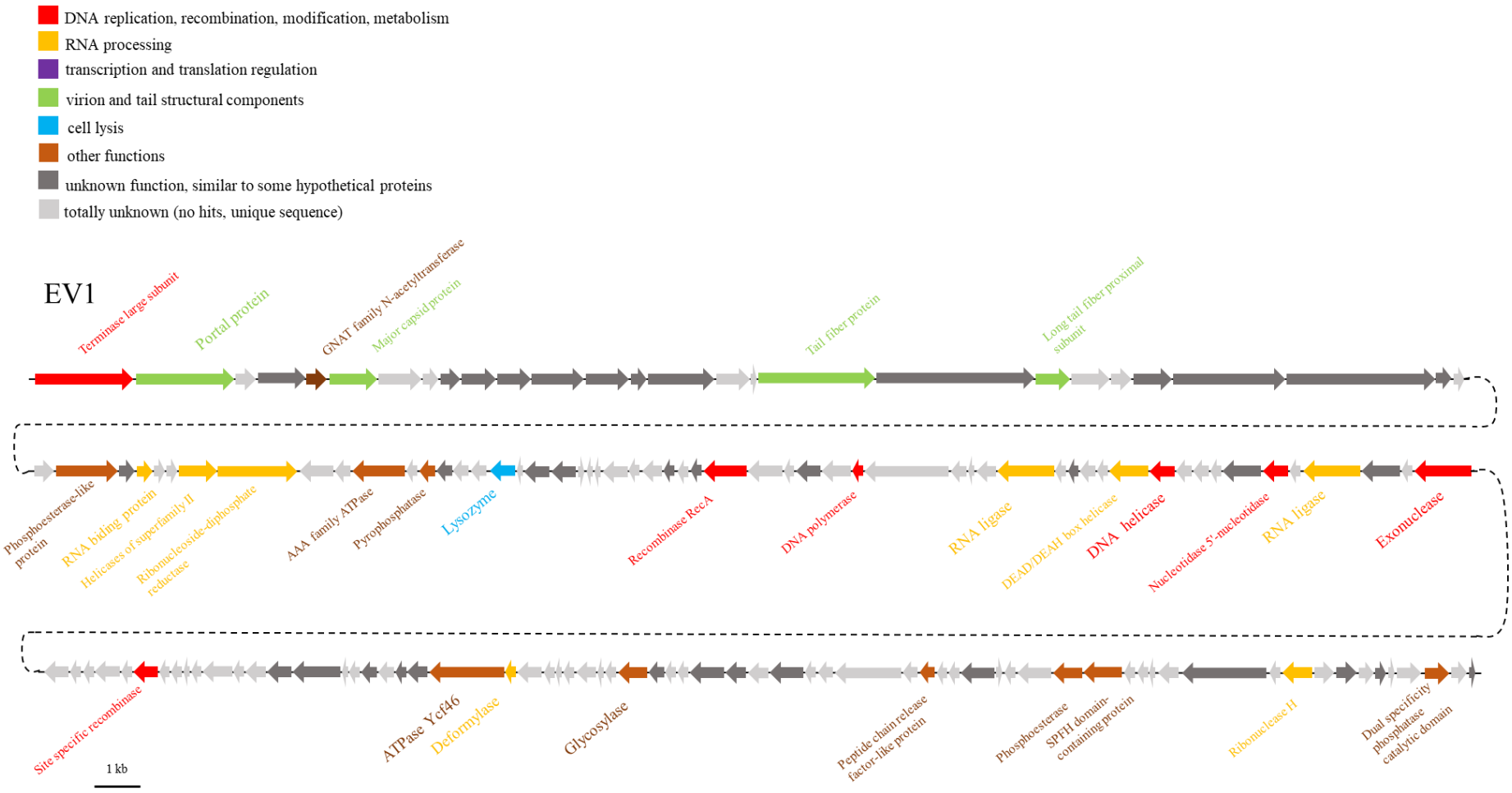
Sequencing revealed that the genome sizes ranged from 63,196 to 308,711 bp and the genomic GC contents (molar content of guanine plus cytosine) varied from 51.3 to 58.4% (Table 7). All genomes had direct 127 bp end repeats, suggesting circular sequences, and one of the two repeats in each pair was removed before further analyses. The virus genomes were predicted to contain from 108 to 348 open reading frames (ORFs, Table 7, Suppl. Tables S2–S6). The ORFs were numbered starting from the ORF encoding terminase large subunit (see below). The viruses had ORFs tightly packed in the genome, 1.1-1.7 ORFs/kbp (Fig. 2, Table 7). An ORF encoding the highly conserved replication initiation protein of the ssDNA viruses (Malathi and Renuka Devi 2019), was not detected in any of genome sequences. The assembly of the sequences and the lack of Rep gene indicate that the genomes are most likely circular dsDNA molecules.

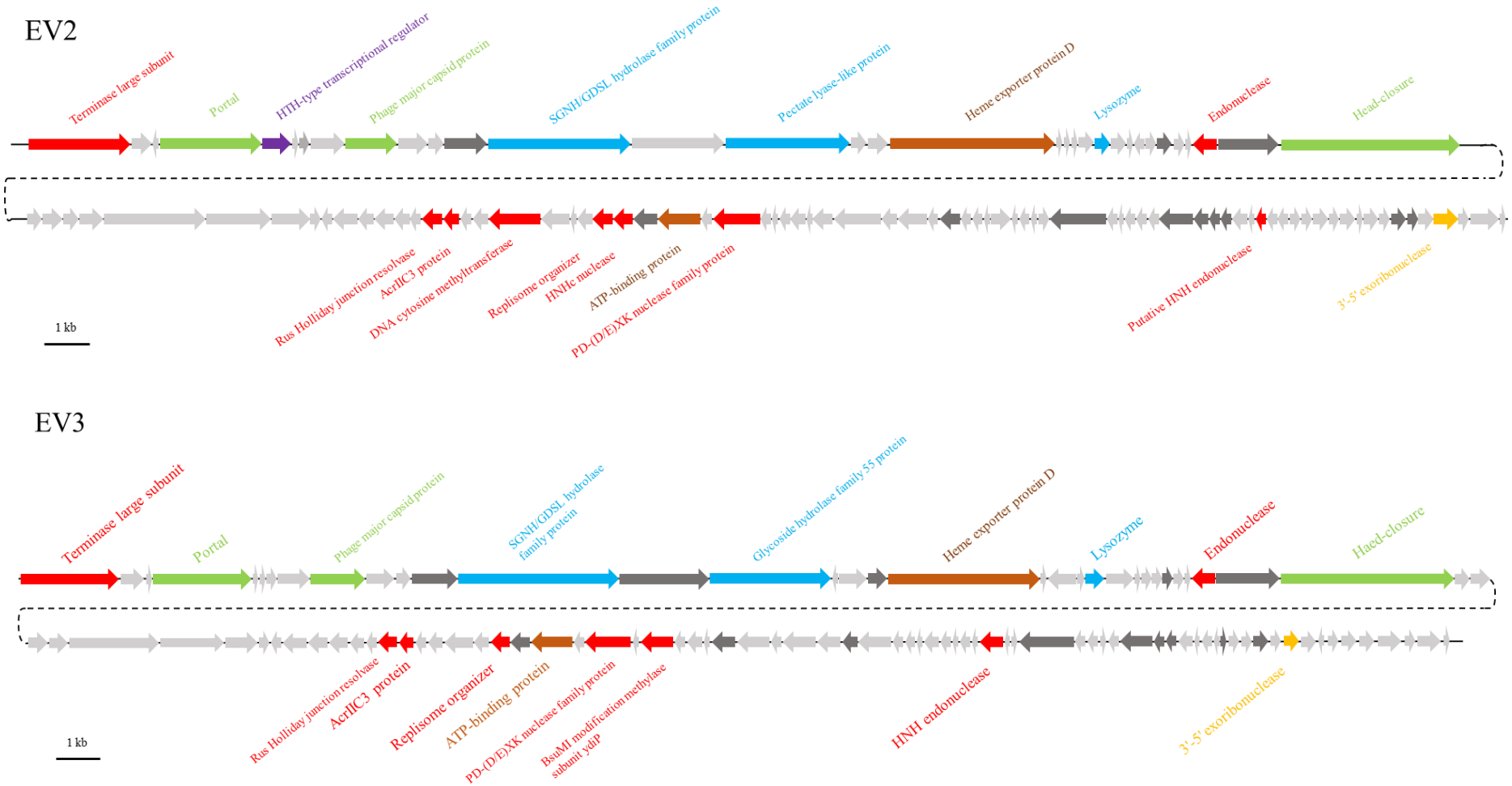
Table 7. Genomic features of the virus isolates.

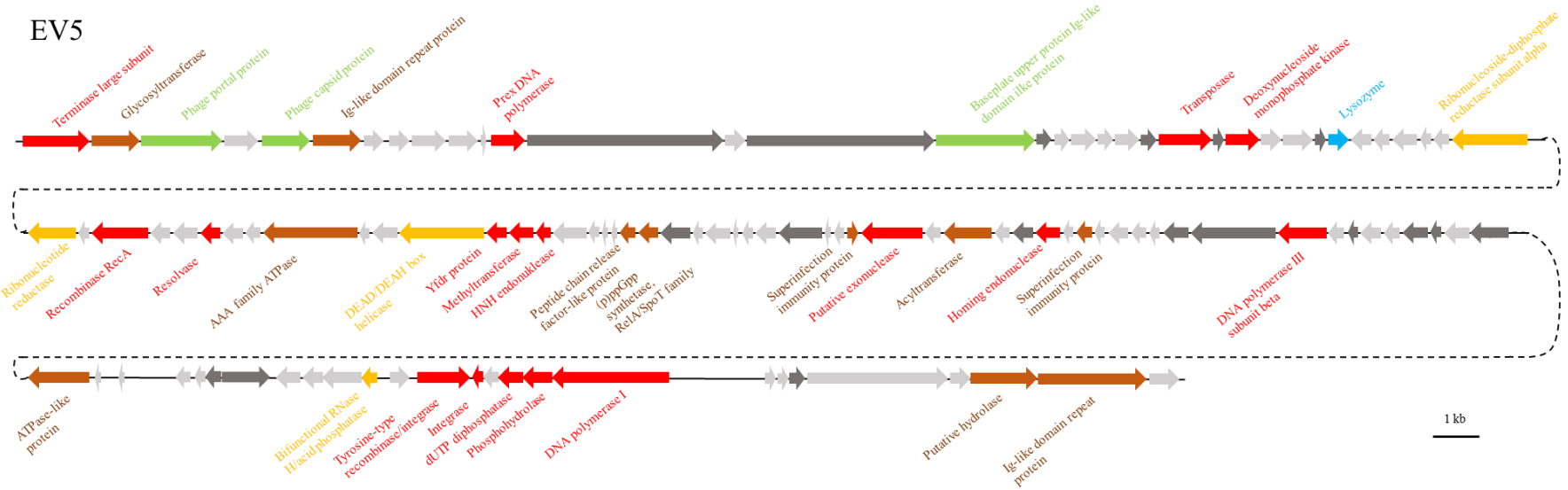
Virus isolate	Genome size, bp	GC content, %	Total no. of ORFs	ORFs/kbp
EV1	97,608	51.3	152	1.6
EV2	63,169	55.3	109	1.7
EV3	63,277	55.1	108	1.7
EV5	88,042	55.4	115	1.3
GV1	308,711	58.4	348	1.1

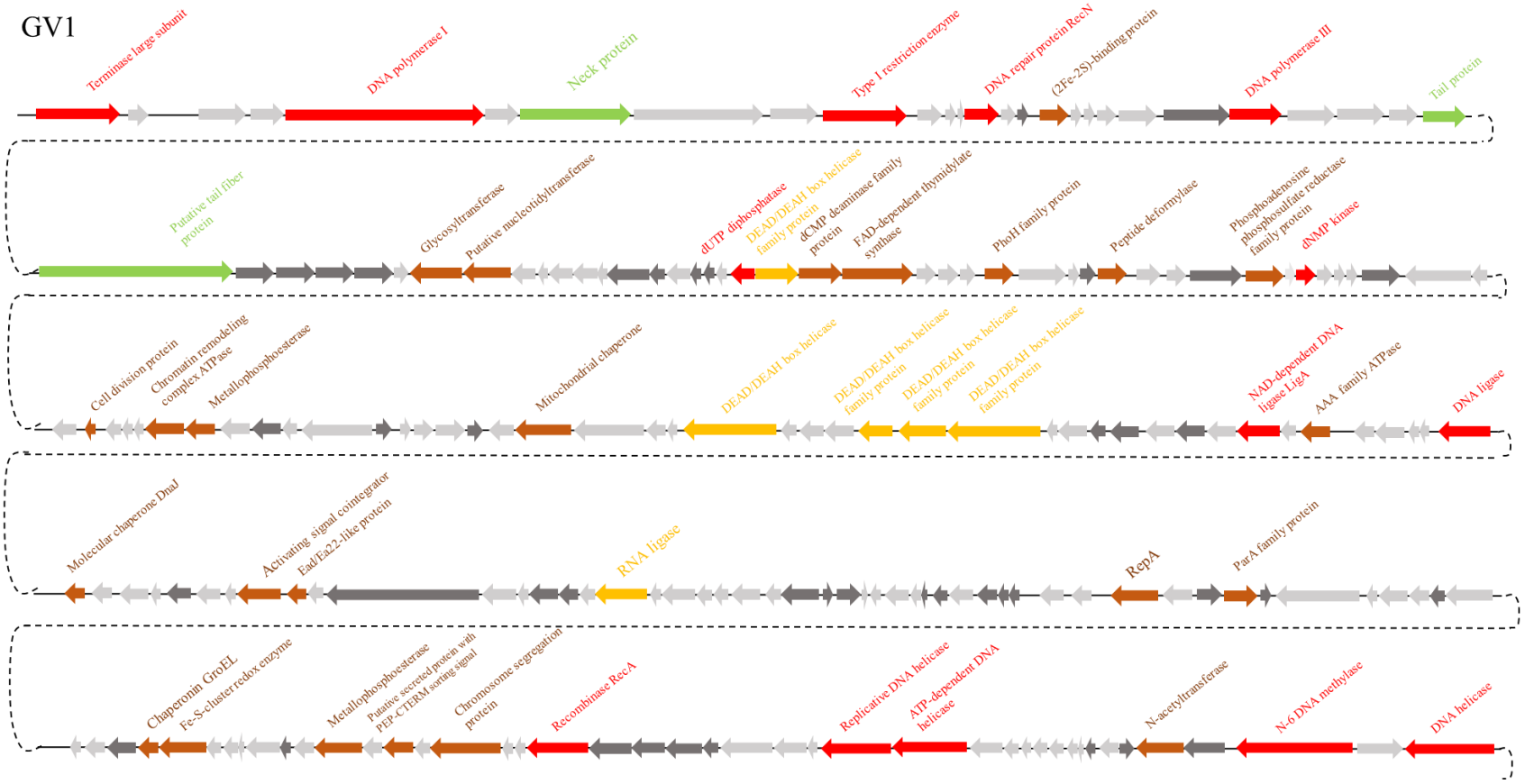
Among the five virus isolates, only EV2 and EV3 genomes were similar, having overall nucleotide identity of 87.8% (Fig. S1). The ORFs were annotated based on the blastx (Stephen et al., 1997) similarity searches (Fig. 2). The ORF encoding terminase large subunit was found in all five genomes and was named ORF1. Of all the ORFs, 54-72% were unique to each of the five viruses, i.e., not having homologs in the NCBI protein sequences database (Fig. 3, light grey segment). From 10 to 24% of the ORFs had similarities to microbial sequences with unknown functions (Fig. 3, dark grey segment). In the genome of EV5, 30% of the 115 ORFs could be annotated (Fig. 3). In the genomes of EV1, EV2, EV3 and GV1, only 16-19% of the number of ORFs could be annotated. Two largest groups of the ORF products that could be functionally assigned were enzymes needed for DNA replication, recombination, modification and metabolism (Fig. 3, red segment) and the mixed group of enzymes and other proteins involved e.g., in energy, sugar and protein metabolism (Fig. 3, brown segment). Viral structural proteins were the third largest group of ORF products that could be annotated (Fig. 3, green segment). ORFs encoding putative tail structural proteins could be found in the EV1, EV5 and GV1 genomes, suggesting that these are tailed phages.

All four *Edaphobacter* virus genomes contained an ORF encoding a putative protein with a lysozyme activity. EV2 and EV3 genomes also contained other ORFs whose products are involved in cell wall hydrolyses: a lipolytic SGNH/GDSL hydrolase family protein, gene product 13 (gp13), a pectate lyase-like protein in EV2 (gp15), and a glycoside hydrolase family 55 protein in EV3 (gp15). EV1, EV5 and GV1 genomes had an ORF encoding the recombinase RecA (ORF57, ORF38 and ORF181 respectively). RecA has been shown to facilitate homologous recombination between viruses and is also used in horizontal gene transfer (Lee et al., 2018). EV5 proteins gp66 and gp74 were assigned with the putative function of superinfection immunity, which is associated with lysogeny and serves to prevent bacteria from being infected by two or more related viruses, or to protect the host cell from being lysed (Berngruber et al., 2010). Two putative host derived AMGs were found in the genomes. EV1 gp107 was annotated to have a putative function of Ycf46 protein. In cyanobacteria, Ycf46 protein was shown to have a role in inorganic carbon utilization by regulating photosynthesis (Jiang et al. 2015). The gene was active under CO_2 starvation: the production of the Ycf46 protein was increased under a low concentration of dissolved inorganic carbon. The other AMG was a PhoH family protein in GV1 (gp54). PhoH is one of phosphate (Pho) regulon proteins, which regulates phosphate uptake and metabolism under low-phosphate conditions (Goldsmith et al., 2011).


DNA replication, recombination, modification, metabolism
RNA processing
transcription and translation regulation
virion and tail structural components
cell lysis
other functions
unknown function, similar to some hypothetical proteins
totally unknown (no hits, unique sequence)
EV1
Terminase large subunit
Portal protein
GNAT family N-acetyltransferase
Major capsid protein
Tail fiber protein
Long tail fiber proximal subunit
Phosphoesterase-like protein
RNA biding protein
Helicases of superfamily II
Ribonucleoside-diphosphate reductase
AAA family ATPase
Pyrophosphatase
Lysozyme
Recombinase RecA
DNA polymerase
RNA ligase
DEAD/DEAH box helicase
DNA helicase
Nucleotidase 5'-nucleotidase
RNA ligase
Exonuclease
Site specific recombinase
ATPase Ycf46
Deformylase
Glycosylase
Peptide chain release factor-like protein
Phosphoesterase SPFH domain-containing protein
Ribonuclease H
Dual specificity phosphatase catalytic domain
1 kb


EV2
Terminase large subunit
Portal
HTH-type transcriptional regulator
Phage major capsid protein
SGNH/GDSL hydrolase family protein
Pectate lyase-like protein
Heme exporter protein D
Lysozyme
Endonuclease
Head-closure
Rus Holliday junction resolvase
AcrIIC3 protein
DNA cytosine methyltransferase
Replisome organizer
HNHe nuclease
ATP-binding protein
PD-(D/E)XK nuclease family protein
Putative HNH endonuclease
3'-5' exoribonuclease
1 kb
EV3
Terminase large subunit
Portal
Phage major capsid protein
SGNH/GDSL hydrolase family protein
Glycoside hydrolase family 55 protein
Heme exporter protein D
Lysozyme
Endonuclease
Haed-closure
Rus Holliday junction resolvase
AcrIIC3 protein
Replisome organizer
ATP-binding protein
PD-(D/E)XK nuclease family protein
BsuMI modification methylase subunit ydiP
HNH endonuclease
3'-5' exoribonuclease
1 kb


EV5
Terminase large subunit
Glycosyltransferase
Phage portal protein
Phage capsid protein
Ig-like domain repeat protein
Prex DNA polymerase
Baseplate upper protein Ig-like domain like protein
Transposase
Deoxynucleoside monophosphate kinase
Lysozyme
Ribonucleoside-diphosphate reductase subunit alpha
Ribonucleotide reductase
Recombinase RecA
Resolvase
AAA family ATPase
DEAD/DEAH box helicase
Ydr protein
Methyltransferase
HNH endonuklease
Peptide chain release factor-like protein
(p)ppGpp synthetase, RelA/SpoT family
Superinfection immunity protein
Putative exonuclease
Acyltransferase
Homing endonuclease
Superinfection immunity protein
DNA polymerase III subunit beta
ATPase-like protein
Bifunctional RNase H/acid phosphatase
Tyrosine-type recombinase/integrase
Integrase
dUTP diphosphatase
Phospholydrolase
DNA polymerase I
Putative hydrolase
Ig-like domain repeat protein
1 kb


GV1
Terminase large subunit
DNA polymerase I
Neck protein
Type I restriction enzyme
DNA repair protein RecN
(2Fe-2S)-binding protein
DNA polymerase III
Tail protein
Putative tail fiber protein
Glycosyltransferase
Putative nucleotidyltransferase
dUTP diphosphatase
DEAD/DEAH box helicase family protein
dCMP deaminase family protein
FAD-dependent thymidylate synthase
PhoH family protein
Peptide deformylase
Phosphoadenosine phosphosulfate reductase family protein
dNMP kinase
Cell division protein
Chromatin remodeling complex ATPase
Metallophosphoesterase
Mitochondrial chaperone
DEAD/DEAH box helicase
DEAD/DEAH box helicase family protein
DEAD/DEAH box helicase family protein
DEAD/DEAH box helicase family protein
NAD-dependent DNA ligase LigA
AAA family ATPase
DNA ligase
Molecular chaperone DnaJ
Activating signal cointegrator Ead/Ea22-like protein
RNA ligase
RepA
ParA family protein
Chaperonin GroEL
Fe-S-cluster redox enzyme
Metallophosphoesterase
Putative secreted protein with PEP-CTERM sorting signal
Chromosome segregation protein
Recombinase RecA
Replicative DNA helicase
ATP-dependent DNA helicase
N-acetyltransferase
N-6 DNA methylase
DNA helicase

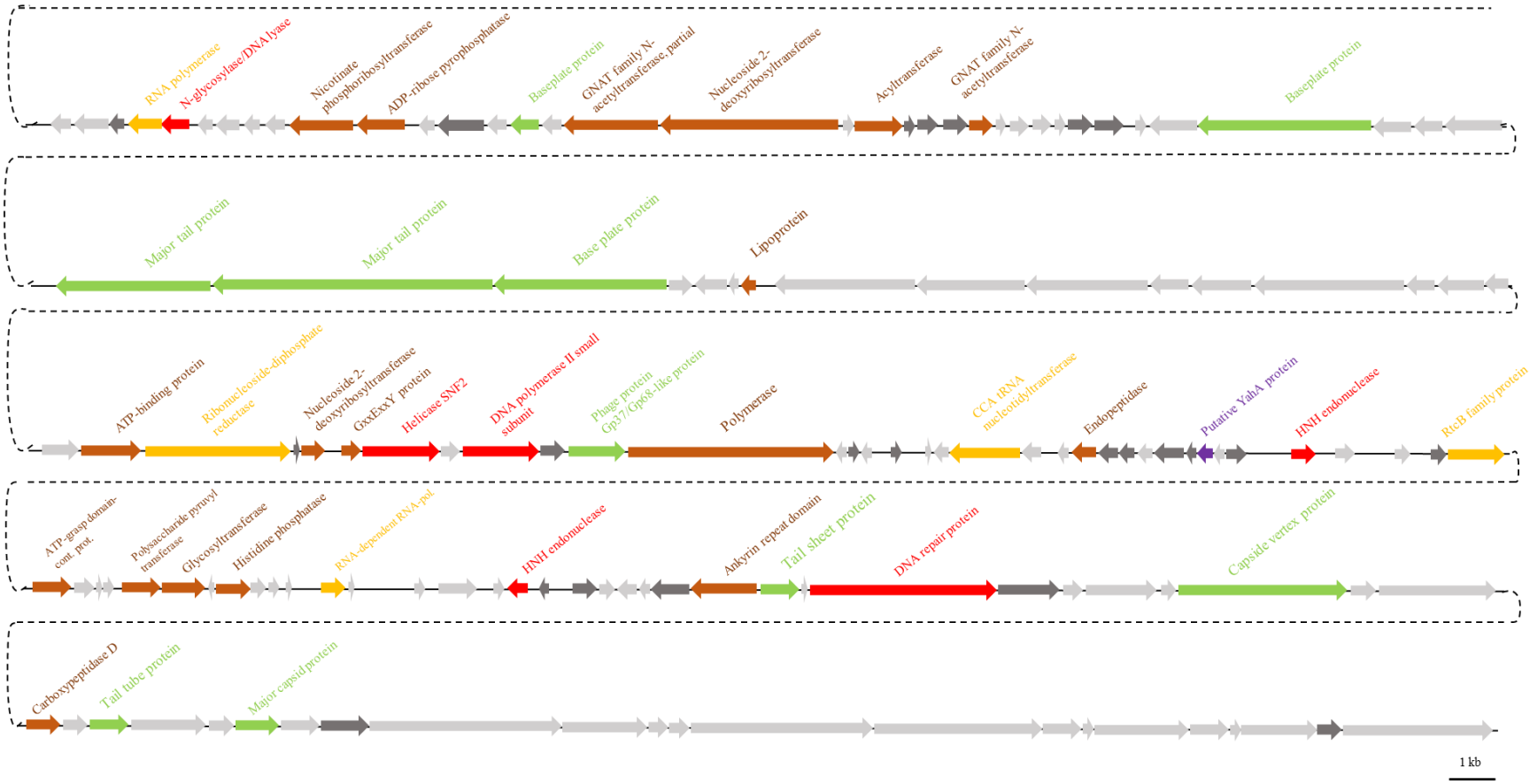


Figure 2. Linear representations of the EV1 (97.6 kb), EV2 (63.2 kb), EV3 (63.3 kb), EV5 (88.0 kb) and GV1 (308.7 kb) genomes. ORFs are shown as arrows that indicate the reading direction. Scale bar is 1 kbp.

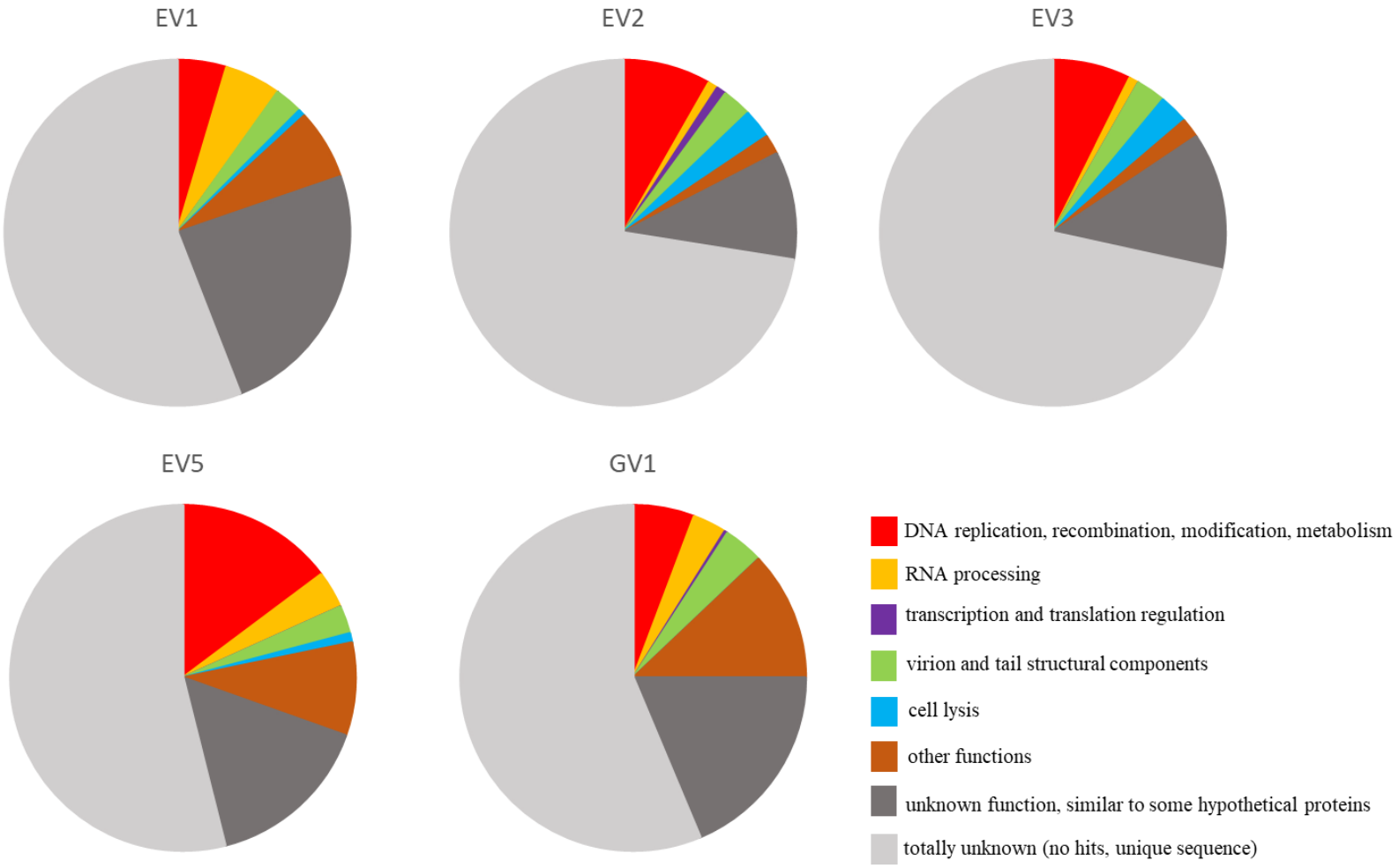


Figure 3. Percentage of different functional groups of the ORF products, colour shows the putative function.

For putative virus classification, we have used the Virfam programme (Lopez et al., 2014), which analyzes the ORFs corresponding to eight virus structural and packaging proteins (Table 8.). According to the Virfam analysis, EV2 and EV3 most likely belong to the *Podoviridae* family, i.e., have a short non-contractile tail, while EV5 belongs to the *Siphoviridae* family of Type 1 Cluster 5, i.e., has a long flexible tail. According to Lopez et al. (2014), the category of *Siphoviridae* Type 1 Cluster 5 includes phages that adopt the structural organization of bacteriophage SPP1 neck and infect Proteobacteria or Streptomyces. The Virfam analyzes could not classify EV1 or GV1. EV1 could not be confidently assigned to any of the four types of head-and-neck organization of tailed phages described in Lopez et al. (2014). However, the EV1 genome sequence contains ORFs encoding a tail fiber protein (gp18) and a long tail fiber proximal subunit (gp20), indicating that the virus belongs to the order *Caudovirales*. GV1 sequence was apparently too long (308.7 kb) for the Virfam analysis, as no results could be obtained, but the sequence contains several ORFs for structural proteins, including neck protein (gp7), tail protein (gp27), putative tail fiber protein (gp28), three base plate proteins (gp220, gp338 and gp344), two major tail proteins (gp342 and gp343), showing that the virus belongs to the order *Caudovirales*.

Table 8. structural and packaging proteins Virfam programme uses for morphological classification of tailed viruses (Lopez et al., 2014).

Structural part	**Specification**
Major Capsid Protein (MCP),	Self-assembling protein forming the procapsid
Portal	Protein which forms a ring through which the DNA is packaged into a procapsid
Terminase	Complex with ATPase and endonuclease activities and a DNA-recognition component, which cuts and translocates viral dsDNA into a procapsid
Adaptor	Head-completion protein bound to the portal
Head-closure	Head-completion protein belonging to the connector, which provides the docking point for tail attachment
Tail completion	Protein belonging to long tails
Major Tail Protein	Protein forming the phage's tail tube
Sheath	Component of the contractile tail that surrounds the central tail tube protein, only in Myophages

4.3 Assembly of Kilpisjärvi metagenome

Sequencing of total DNA from the sample o12217 resulted in 77,704,310 reads after trimming. Trimmed reads were assembled using MEGAHIT as part of the Lazypipe pipeline (Plyusnin et al., 2020) and metaSPAdes (van der Walt et al., 2017), and their performance was compared (Table 9). The number of contigs assembled by MEGAHIT was almost double to the number of contigs assembled by metaSPAdes. The assembly span of MEGAHIT (1,660 Mbp) was also much larger than the assembly span of metaSPAdes (1,345 Mbp). MEGAHIT also assembled more of the long contigs (≥1 kbp) with larger total length than metaSPAdes, whereas metaSPAdes assembled slightly more of the very long contigs (≥25 kbp) than MEGAHIT (Fig. 4). MetaSPAdes also produced the longest contig (374,190 bp). In general, the performance of both assemblers was very good and comparable, as both produced high N50 values, the length of the shortest contig at 50% of the total genome sequence.

Table 9. Assembly statistics of o12217 metagenome.

Parameters	metaSPAdes	MEGAHIT
Number of contigs	633,529	912,843
Total length	1,345,076,145	1 659,853,037
Number of long contigs (≥1 kbp)	235,032	311,133
Total length of long contigs (≥1 kbp)	641,309,222	766,603,570
Largest contig, bp	374,190	218,556
N50	1,841	1,484
L50	95,015	163,251

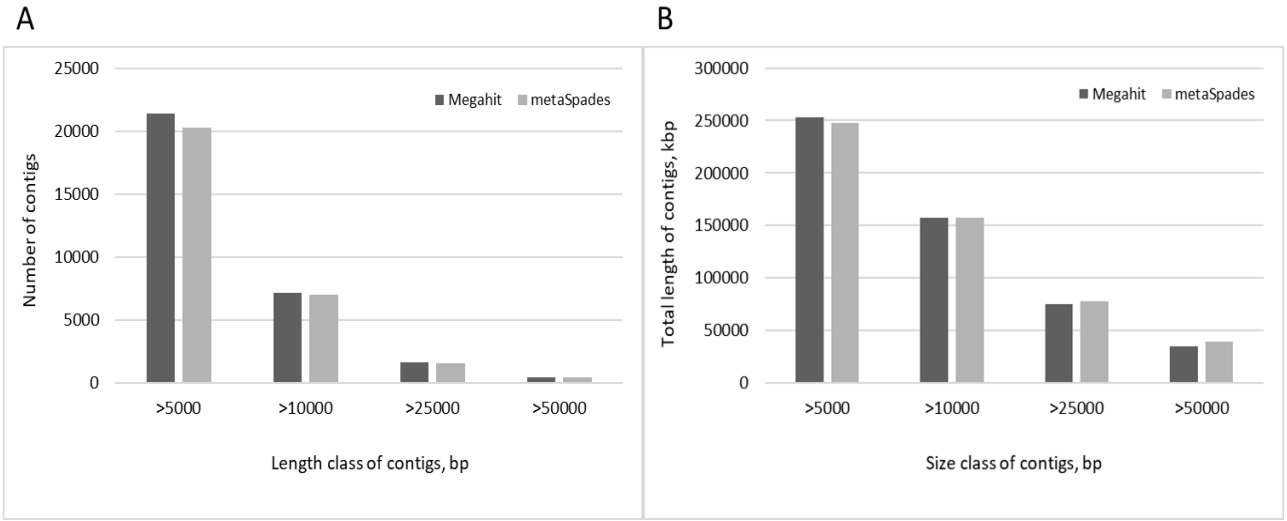


Figure 4. Comparison of contigs assembled by metaSPAdes and Megahit. (A) Number of large contigs in different size classes, (B) total length of contigs in different size classes.

When reads were mapped to contigs, the overall alignment rates were 68.74% for metaSPAdes and 70.49% for MEGAHIT assemblies, respectively. Coverage and mean depth values were higher for the metaSPAdes assembly (Table 10).

Table 10. Completeness of the contigs (%).

Assembler	Horizontal coverage	Vertical coverage
MEGAHIT	69.8	3.5
metaSPAdes	98.7	4.7

4.4 Recovery and annotation of viral contigs

4.4.1 Annotation and taxonomic profiling within the Lazypipe

Taxonomic abundancies of bacteria, viruses and archaea were mapped within the Lazypipe pipeline. Over 99% of the contigs obtained with MEGAHIT were annotated as Bacteria, 0.5% as Archaea, and no contigs were annotated as viruses. The most prominent taxons were Actinobacteria (32%) and Acidobacteria (29%), the others were Chloroflexi (11%), Proteobacteria (8%), Planctomycetes (7%) and Verrucomicrobia (6%).

The Acidobacteria contigs included ones putatively represented the following species: "*Candidatus* Sulfotelmatobacter" (72%), *"Candidatus* Sulfotelmatomonas" (9%), *Granulicella* (4%), "*Candidatus* Koribacter" (4%), *Edaphobacter* (3%) and *Acidiphila* (2%) (Fig. 5A). The *Granulicella* species included *Granulicella* sp. GAS466 (57%), *Granulicella sibirica* (10%), *Granulicella mallensis* (9%), *Granulicella rosea* (7%), *Granulicella pectinivorans* (6%), *Granulicella* sp. 5B5 (5%) and *Granulicella tundricola* (5%) (Fig. 5B). The Edaphobacter species were *Edaphobacter aggregans* (46%), *Edaphobacter modestus* (41%), *Edaphobacter dinghuensis* (9%) and *Edaphobacter* sp. 12200R-103 (Fig. 5C).

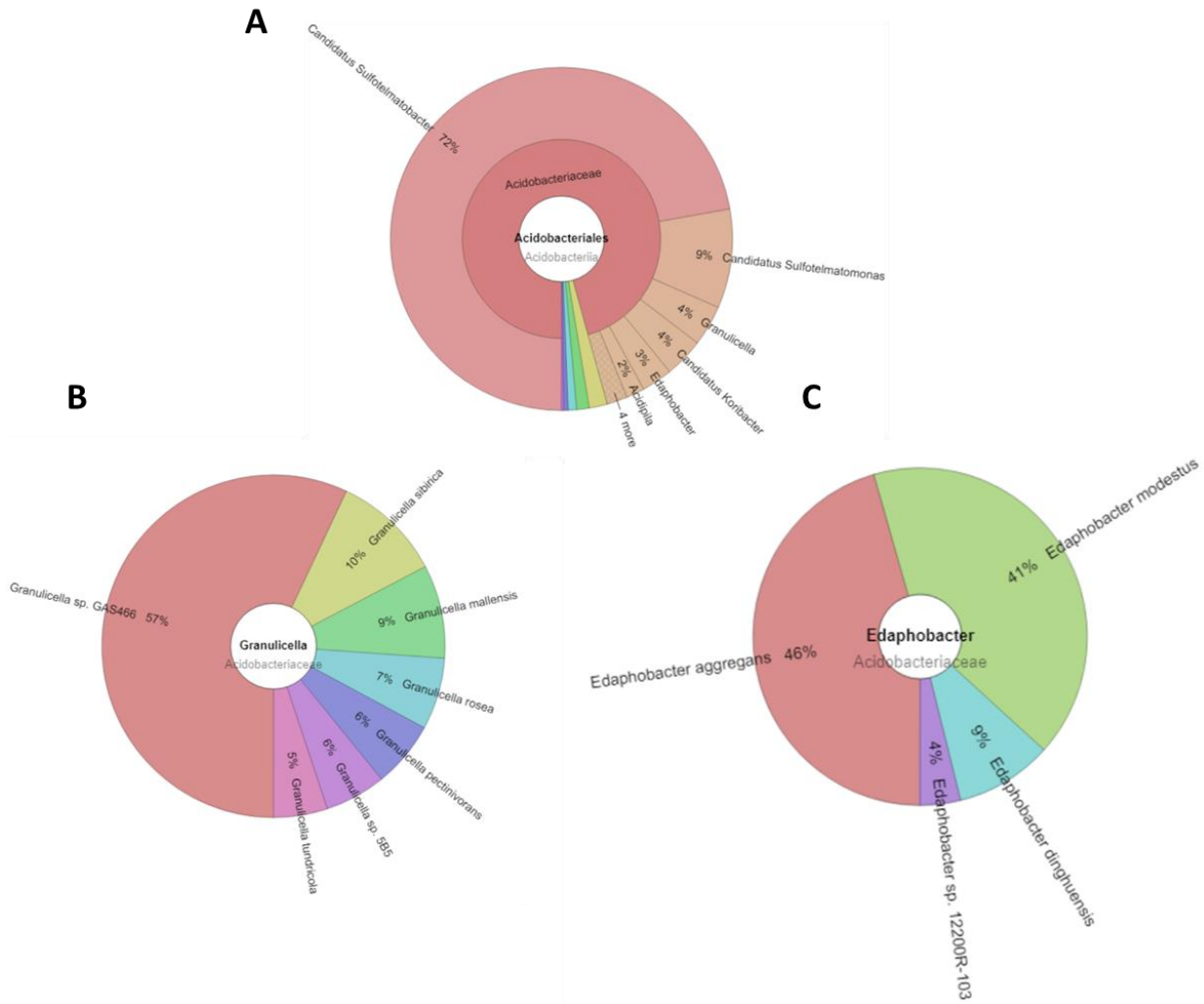


Figure 5. Taxonomy profile of metagenome from the soil sample o12217 Kilpisjärvi (collected in July 2018). (A) The order Acidobacteriales, (B) the genus *Granulicella* of the family Acidobacteriaceae, (C) the genus *Edaphobacter* of the family Acidobacteriaceae.

4.4.2 Annotation and taxonomic profiling for the assembly obtained with metaSPAdes

The contigs obtained with metaSPAdes were subjected to the What-the-Phage pipeline (Marquet et al. 2020), which uses several virus prediction programs. Figure 6 summarizes the identification performance of the ten programs What-the-Phage pipeline used. In total, 9,412 contigs (nodes) were identified as "phage-positive" by all the programmes together. However, only 627 of them contained at least one viral gene as confirmed by CheckV and were thus selected for the further analyzes. From these, only two contigs were of high quality, two of medium quality, 138 were low-quality, and 485 had no determined quality. The medium- and high-quality contigs were 66-100% complete, while others were 0.8-38% complete. The length of the contigs varied from 1,502 bp to 125,568 bp. The number of genes in the contigs varied between 1 and 118. The contigs contained 1–15 viral genes: 452 contigs had one viral gene, 123 contigs had two viral genes, 34 contigs had three viral genes and

18 contigs had four or more viral genes. In addition to the viral genes, the contigs contained 0–73 host genes: 522 contigs had no host genes, 62 contigs had one host gene, 18 contigs had two host genes and 25 contigs had three or more host genes. Eighteen contigs were annotated as proviruses. DNA homology search for taxonomic assignments (nhmmer, Eddy 1995) within the What-the-Phage pipeline could not produce any classifications.

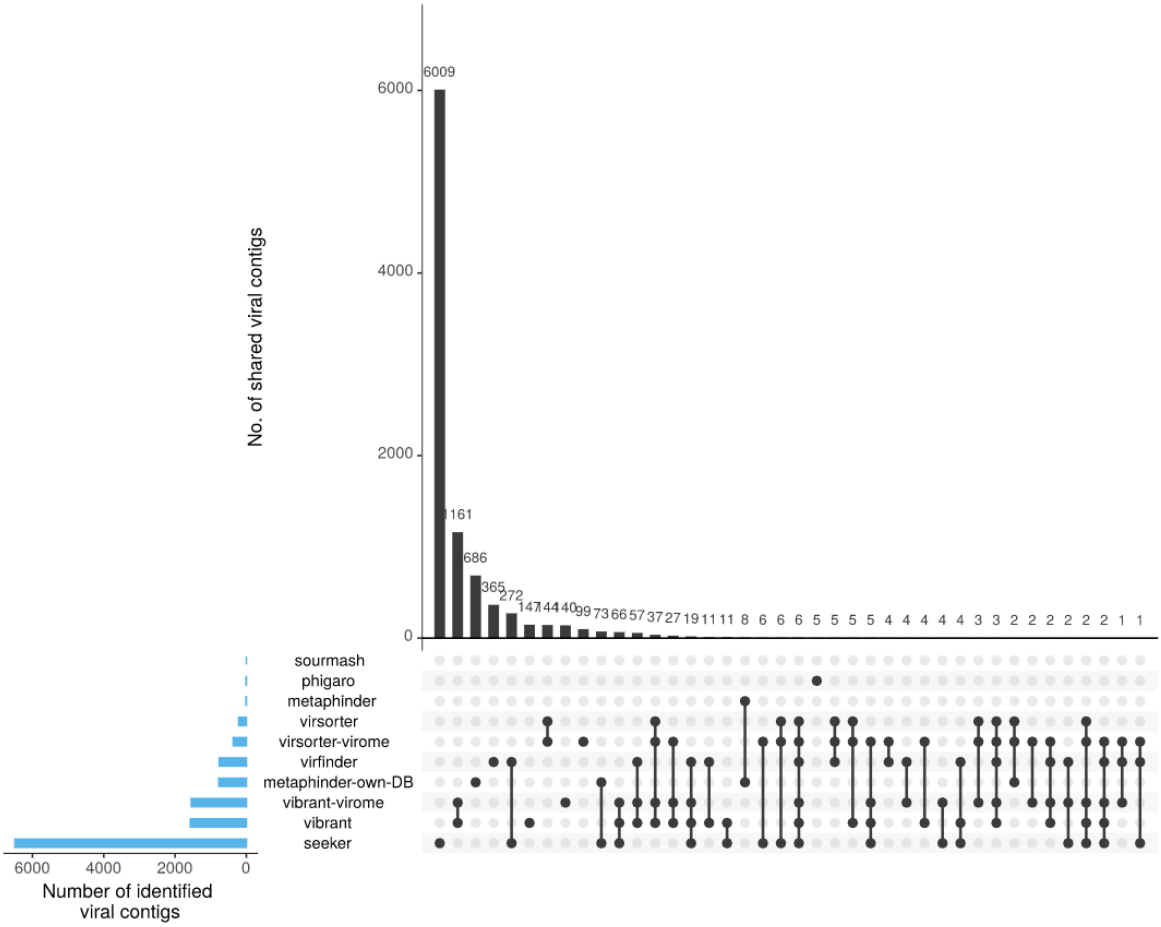


Figure 5. Viral contig analysis of the o12217 metagenome by the What-the-phage pipeline. UpSetr plot summarizing the identification performance of each program. Black bars: number of contigs uniquely identified by each program or program combination (dot matrix below each column). Blue bars: total amount of identified phage-contigs of each program.

4.2.1 Relating Kilpisjärvi metagenome-derived viral contigs to known viral sequences

The metagenome-derived viral sequences were analyzed by clustering using the gene-content based classification method (Jang et al., 2019). For this, the subset of viral contigs that are at least 10 kbp long or at least 50% complete was selected (74 contigs total) due to the software's better performance on longer sequences. This dataset also included the five virus isolates complete genomes described in 4.1. In total, 134 clusters (= approximately genus-level groupings) were obtained. From the five

virus isolates described here, GV1 clustered with two metagenomic viral contigs, EV2 and EV3 clustered with each other and three metagenome-derived viral contigs, and EV1 and EV5 clustered with each other and four metagenomic viral contigs. Only one metagenome node clustered with several previously identified viruses from the database, all of them were tailed phages: Pseudomonas phage vB_PaeP_Tr60_Ab31, Escherichia phage TL-2011b, Escherichia phage phiV10, Bordetella virus BPP1, Salmonella phage epsilon15, Salmonella phage ST64B, Xanthomonas citri phage CP2 (Fig. 6). The rest of the metagenome nodes clustered only with other nodes of the metagenome (Fig. 6).

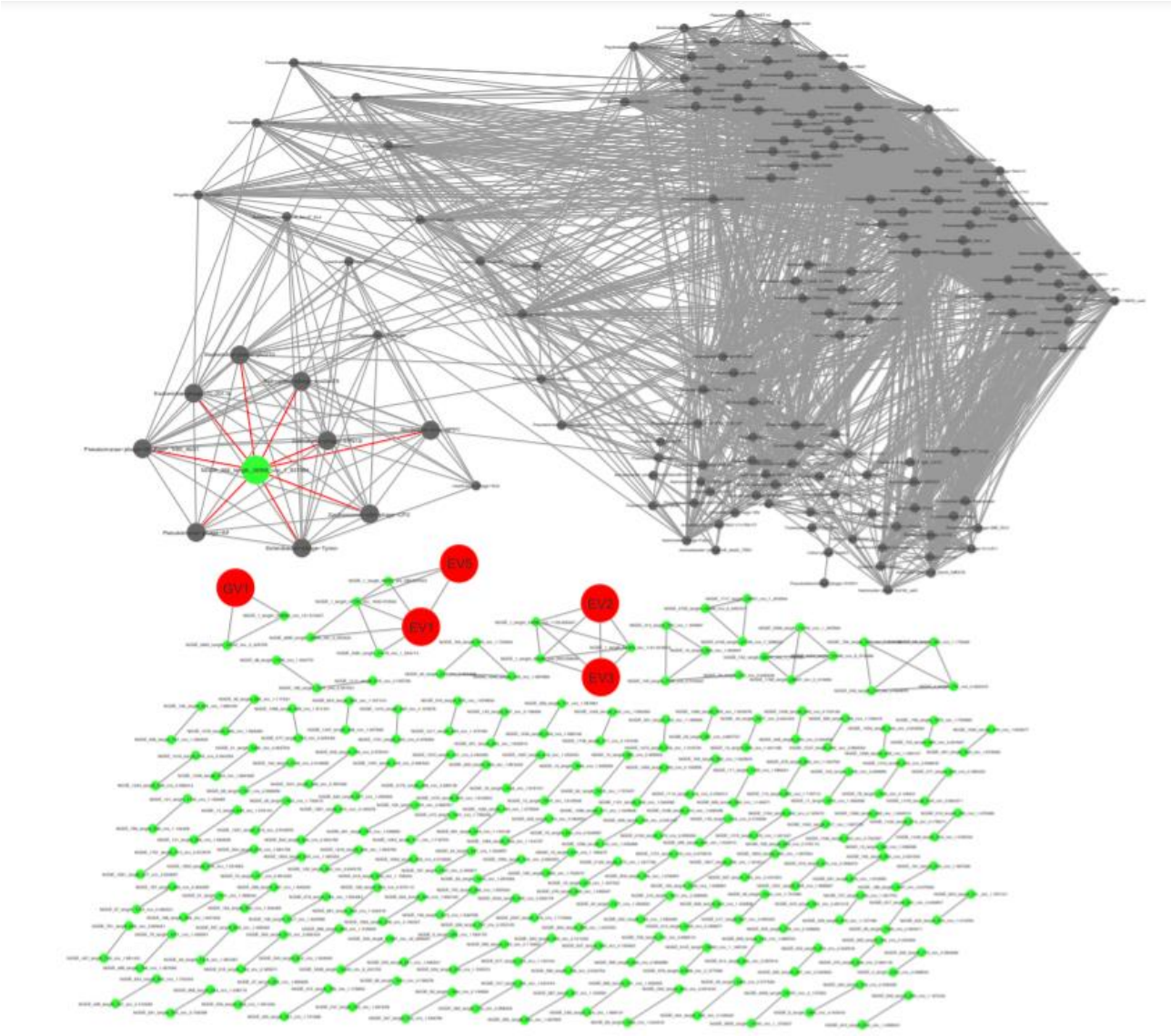


Figure 6. Clustering of the isolated Kilpisjärvi viruses (red nodes) with the Kilpisjärvi viral metagenome-derived viral contigs (green nodes) and all RefSeq (v201) viral genomes (gray nodes). Edges (lines) between nodes indicate pairwise similarity.

5. Discussion

5.1 The first five acidobacterial virus isolates and their genetic diversity

Although Acidobacteria is one of the most prominent bacterial phyla in different soil types around the world, to our knowledge, the five viruses isolated in the present study are the first Acidobacteria-infecting virus isolates ever reported. The viruses were isolated by the plaque assay technique from Arctic soil samples using three different acidobacterial strains: three viruses on *Edaphobacter* sp. X5P2, one on *Edaphobacter* sp. M8UP27, and one on *Granulicella* sp. X4BP1. The two *Edaphobacter* strains are the most similar to *Edaphobacter lichenicola* SBC68, and *Granulicella* sp. X4BP1 is the most similar to *Granulicella aggregans* (Belova et al., 2018). The used host strains had been previously isolated from the same Kilpisjärvi mountain area, where Acidobacteria have been shown to dominate in the acidic tundra soils (Männistö et al., 2007, 2013; Viitamäki 2019).

The two soil sample sets collected from Kilpisjärvi in July 2018 and 2019 failed to yield virus isolates with any of the five different isolation protocols consecutively tested, including protocols 3 and 4 that were based on enrichment cultures. These samples had been frozen immediately after being collected and have been stored at –80 °C for two to three years before the isolation attempts. Noticeably, the isolation succeeded only from the "fresh" soil samples collected from Kilpisjärvi in April 2021, stored at 4 °C and eluted within a month. According to Gould (1999), both DNA and RNA viruses are typically stable when kept for months at refrigerator temperatures and stored for years at very low temperatures without any special preservatives or carefully regulated freezing techniques: the tiny size, simple structure and the absence of free water are the reasons for such stability of viruses. Preserving the viruses at low or ultra-low temperatures significantly increases the length of time that the virus can be stored as infectious material (Gould 1999). In the study on viruses in wastewater, Olson et al. (2004) observed less bacteriophage MS2 inactivated after short storage (8 days) at 4 °C (20%) compared to storage at –80 °C (58%), while during the extended storage (300 days), less MS2 was inactivated at –80 °C (75%) compared to 4 °C (93%). Differing storage temperatures of the Arctic soil samples (–80 °C versus 4°C, processed within a month) did not affect viral abundances evaluated by direct count (epifluorescence microscopy, Trubl et al., 2016). Trubl et al. (2018) extracted and sequenced DNA from viral particles purified from soil samples and evaluated the effect of the sample storage on vOTU recovery (–80 °C versus 4 °C, processed within a month). Their results showed a broader recovery of vOTUs from the chilled bog sample in comparison to the frozen sample, but no effect of storage temperature was observed in the palsa and fen habitat samples. They speculated that the higher diversity in the chilled bog virome in comparison to the frozen one could have several potential causes: freezing could damage viral particles due to the rapidity of freezing with liquid

nitrogen, while the chilled sample could have contained active microbes with ongoing viral infections differing from those in the field environment, or there could have been a specific induction of temperate viruses in the chilled bog sample. The bog habitat is very acidic (pH ~4) and very wet with changing water levels, both factors that are connected to the increased selection for temperate viruses (Evans et al., 2012; Payet and Suttle 2013). Both examples of negative effects of freezing (storage at –80 °C) on virus activity deal with very wet samples, Olson et al. (2004) with wastewater and Trubl et al. (2018) with bog, which consists of soggy *Sphagnum* moss. These are both chemically and physically different environment for microbes than our samples of humus-rich topsoil, even though both the summer and winter samples had high moisture content of 44–95% and acidic pH (pH 5.1-5.7 for summer samples, pH 4.7-6.0 for winter samples). The tundra soils at Kilpisjärvi are exposed to the wide annual temperature fluctuation including very low (down to –15 °C) winter temperatures and repeated freeze-thaw cycles during autumn and in spring (Männistö et al., 2009). Microbial activity has been reported in soil even at temperatures down to –20 °C (Rivkina et al., 2000). As reported by Männistö et al. (2009), the freeze-thaw cycles down to –10 °C had only a marginal effect on the soil bacterial composition in the samples from Kilpisjärvi tundra soils, and the authors hypothesized that these conditions have selected a stable frost-tolerant bacterial community that is only little affected by the temperature fluctuations and freeze-thaw cycles (Männistö et al., 2009). It would be reasonable to expect that also the viral community in the area has adjusted to the prevailing environmental conditions. However, in case of our samples collected in 2018 and 2019, the exposure to –80 °C for a few years could have affected the integrity of viral particles, especially the structures responsible for host recognition, which might explain why no viruses could have been isolated from these samples.

The obtained virus isolates had dsDNA genomes of sizes falling in a range reported for the Arctic soil viruses, 10,000–440,000 bp (Trubl et al., 2021), and GC content values (56–60%) similar to those reported for Acidobacteria (Männistö et al., 2012; Thrash and Coates 2015; Eichorst et al., 2020). One of the several critical parameters that determine the selectivity with which viruses infect their hosts is the molecular specificity of recognition needed for the virus entry into the host cell (Bahir et al., 2009). All bacteriophages are tuned to match their bacterial hosts, as also seen in their genomic GC contents: most phages show slightly lower GC values than their hosts (Bahir et al., 2009, Simón et al., 2021). The genomic GC percentages are 58.4% for the *Granulicella* virus GV1 and vary from 51.3 to 55.4% for the four *Edaphtobacter* viruses, which corresponds well with the GC values of 56.0–59.9% reported for the *Granulicella* species isolated from Kilpisjärvi (Männistö et al., 2012) and 55.8–56.9% reported for *Edaphtobacter* (Thrash and Coates 2015).

Based on the similarity searches against the publicly available viral sequences, the isolated acidobacterial viruses are not closely related to any known viruses. They are also not similar one to another, except EV2 and EV3, which were isolated from the same sample using the same host. The viral genome sequences are largely unique, as only a small fraction of ORF products could be assigned with putative functions (Fig. 2), highlighting yet unknown viral diversity present in soil. The annotated ORF products relate to virion structures or the major steps of the phage life cycle, such as e.g., cell lysis. The putative proteins involved in the regulation of virus replication and superinfection immunity, which is associated with lysogeny, were also found. The average number of lysogenic phages in bacterial genomes is 2.6, and some genomes can contain up to 17 different lysogenic phages (Casjens 2003). Superinfection immunity prevents the host bacterium from being infected by other viruses and helps to protect the host cell from being lysed (Berngruber et al. 2010). High occurrence of lysogeny is linked with low nutrient availability limiting cell growth rate and inducing low host cell density (Maurice et al., 2010). Ycf46-like protein in EV1 (gp107) and PhoH family protein in GV1 (gp54) may be linked to low nutrient conditions: CO_2 starvation (Jiang et al., 2014), and low-phosphate conditions (Goldsmith et al., 2011), respectively. It remains to be studied which role these putative AMGs may play in virus-host interactions between EV1 and GV1 and their respective hosts.

Based on the presence of ORFs putatively encoding virion structural proteins and the Virfam analysis, all the isolates are tailed phages belonging to the order *Caudovirales*. EV2 and EV3 could be further tentatively classified to the *Podoviridae* family (short, noncontractile tails), and EV5 to the *Siphoviridae* family (long, noncontractile tails). Williamson et al. (2005) compared viral particle morphologies from six different biomes (agricultural, forested, and wetland soils) and detected different types of viral communities in each soil. Five soils were dominated by tailed phages (~80%) and one of the agricultural soils was dominated by spherical, non-tailed particles (56%). The results are in accordance with previous studies in aquatic systems, suggesting that most viruses found in environmental samples are tailed phages (Wommac & Colwell 2000).

Up to 72% of the ORFs were unique to the isolated phages, emphasizing how little acidobacterial viruses have been studied. To our knowledge, the only available information about acidobacterial virus sequences is from metagenomic viral contigs *in silico* linked to acidobacterial hosts (Emerson et al. 2018, Trubl et al., 2018) and putative proviruses found in acidobacterial genomes (Eichhorst *et al.* 2020). The high portion of unknown sequences observed in the obtained isolates is also in accordance with the notion that soils contain a substantial viral genetic diversity for us to discover:

some cosmopolitan groups have been identified, but the majority of viruses show high habitat-type specificity across diverse ecosystems (Paez-Espino et al., 2016).

5.2 Viral sequences detected in the metagenomes from Kilpisjärvi soil

The genome analysis of the isolated viruses was complemented with the metagenome analysis to estimate the presence of the viruses related to the obtained isolates in the studied soil, as well as to assess the overall viral diversity. The soil sample from the same collection site representing *Empetrum*-dominated fen was selected for the metagenomics-based analyses.

The small amount of virus DNA in soil extracts can lead to poor or no assembly of virus sequences in the data and downstream analyses easily missing viruses (Pratama & van Elsas 2018). Thus, an important part of the metagenome analysis is the assembly of the contigs: a good assembler utilizes most of the raw sequence data to produce the biggest possibly assembly span, it also should produce a high number of long contigs (>1,000 bp) needed for the accurate interpretation of full genes and reconstruction of single genomes (van der Walt et al., 2017). Two different assemblers that have shown good performance with soil metagenomes (van der Walt et al., 2017), metaSPAdes and Megahit, were compared to get the best possible assembly of virus sequences. MEGAHIT is an open-source well performing and memory-efficient next-generation sequencing (NGS) assembler optimized for metagenomes (van der Walt et al., 2017). SPAdes is a commonly used open-source metagenome assembler that can be used for both single-cell and multicell data assembly (Bankevich et al., 2012). Here, we used it with the --meta option, i.e., metaSPAdes, an assembler based on SPAdes, but specifically developed for the assembly of metagenomes (Nurk et al., 2017). QUAST (QUality ASsessment Tool) evaluates genome assemblies based on the alignment of contigs to a reference. Among other parameters, it calculates basic assembly statistics, e.g., the number of contigs and assembly span in different length groups, N50 and L50 lengths. *N50* is the length of the shortest contig at 50% of the total genome sequence, and *L50* is the least number of contigs that sum up to half of the genome length. In the assembler comparison of van der Walt et al., (2017) metaSPAdes performed best with high-coverage metagenomes, being less suitable for low-coverage ones, whereas MEGAHIT performed well with both high- and low-coverage samples. On our sample, both assemblers showed good results in terms of the assembly quality, but MEGAHIT was more memory efficient and faster than metaSPAdes, similarly to the previous reports (van der Walt et al., 2017).

In the studied metagenome, no viral contigs could be detected by the Lazypipe pipeline, which performs assembling and taxonomic profiling of bacterial and viral sequences from different environments (Plyusnin et al., 2020). This pipeline has, however, shown that Acidobacteria is one of the largest bacterial groups in the studied samples, which is in accord with other reports (Männistö et

al., 2009, 2013). The analysis using the What-the-Phage pipeline that utilizes several different virus-detection tools yielded 627 phage-positive contigs, which is about 0.1% of all metagenomic contigs initially assembled. The metagenome approach is intrinsically not very efficient at capturing virus signals, because in soil samples, typically, less than 2% of reads are identified as viral (Emerson et al., 2018). Also, viruses in bulk-soil metagenomes represent probably only a fraction of the virome, since most free viruses adsorb to soil. This means that the metagenomics-based analysis generally favours actively reproducing and temperate viruses, but at the moment, it is impossible to determine to which extent viral abundances in metagenomes correlate with the activity and infectivity (Emerson et al., 2017). In the work of Carini et al. (2016), extracellular DNA caused significant misestimation of the relative abundances of different taxa. They analyzed DNA from a wide range of soils and found that, on average, 40% of both prokaryotic and fungal DNA was extracellular or from cells that were no longer intact. After cell death, extracellular DNA can stay amplifiable in soils for weeks to years (Levy-Booth et al., 2007).

The exact sequences of the five described acidobacterial virus isolates were not detected in the analyzed metavirome, but some clustering of approximately genus level was observed (Figure 6), suggesting that the studied soil contains also other viruses similar to the isolated ones. The viral contigs found using the What-the-Phage pipeline could not be taxonomically classified and are apparently specific for the studied soil, as the clustering was observed mainly within the metagenome. Similar results were obtained in metagenome analyses of Swedish Stordalen peatland soils in the study of Emerson et al. (2018). They used the shared protein network analysis to compare the similarity of 1,907 Stordalen viral populations to three published data sets: (i) 2,010 procaryotic viral genomes from RefSeq v75; (ii) 2,040 soil viral contigs from Roux et al. (2015b); and (iii) 3,112 soil-associated viral contigs from Paez-Espino et al. (2016). Stordalen viral populations formed 738 clusters, 451 of which contained only Stordalen viral populations. Only 17% of the Stordalen viral populations could be taxonomically classified (Emerson et al., 2018). The study of Paez-Espino et al. (2016) shows that the habitat specificity of viruses is not limited to soils, as in their analysis of 3,042 assembled metagenomes from 10 ecotypes most viruses appeared to be habitat specific.

In conclusion, we anticipate that further studies of the isolated viruses as well as the isolation and characterization of new viruses infecting tundra soil bacteria will make it possible to better understand the physiology of virus-host interactions and their effects on the structure and functions of soil microbial communities, especially those residing in the permafrost-affected soils. Such information will ultimately help in estimating the microbial input into global nutrient cycling processes and the global climate change feedback loop.

6. Acknowledgements

First, I would like to thank my supervisors Ph.D. Tatiana Demina and Docent Jenni Hultman. Thank you, Tatiana, for your endless enthusiasm, patience in guiding me through the maze of bioinformatics, and your insightful advice on my thesis. Jenni, thank you for always being there to answer my questions and especially for your advice for metagenome analyzes, and commenting on my writing. I would also like to thank Ph.D. Minna Männistö for providing host bacterial stains for this study.

7. References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Anderson, J.P.E., and Domsch, K.H. (2006) Quantities of plant nutrients in the microbial biomass of selected soils. *Soil Sci* **171:** 106–111.

Bahir, I., Fromer, M., Prat, Y., & Linial, M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Molecular systems biology*, *5*, 311.

Ballaud, F., Dufresne, A., Francez, A., Colombet, J., Sime-Ngando, T., and Quaiser, A. (2015) Dynamics of Viral Abundance and Diversity in a Sphagnum-Dominated Peatland: Temporal Fluctuations Prevail Over Habitat. *Front Microbiol* **6:** 1494.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., *et al.* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* **19:** 455–477.

Barns, S.M., Takala, S.L., and Kuske, C.R. (1999) Wide Distribution and Diversity of Members of the Bacterial Kingdom Acidobacterium in the Environment. *Appl Environ Microbiol* **65:** 1731–1737.

Barns, S.M., Cain, E.C., Sommerville, L., and Kuske, C.R. (2007) Acidobacteria Phylum Sequences in Uranium-Contaminated Subsurface Sediments Greatly Expand the Known Diversity within the Phylum. *Appl Environ Microbiol* **73:** 3113–3116.

Beaulaurier, J., Luo, E., Eppley, J.M., Uyl, P.D., Dai, X., Burger, A., *et al.* (2020) Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res* **30:** 437–446.

Beijerinck MW (1898). Über ein Contagium vivum fluidum als Ursache der Fleckenkrankheit der Tabaksblätter. *Verhandelingen der Koninklijke Akademie van Wetenschappen te Amsterdam* **65:** 1–22.

Belova, S.E., Suzina, N.E., Rijpstra, W.I., Sinninghe Damsté, J.S., and Dedysh, S.N. (2018) Edaphobacter lichenicola sp. nov., a member of the family Acidobacteriaceae from lichen-dominated forested tundra. *Int J Syst Evol Microbiol* **68:** 1265–1270.

Berngruber, T.W., Weissing, F.J., and Gandon, S. (2010) Inhibition of Superinfection and the Evolution of Viral Latency. *J Virol* **84:** 10200–10208.

Bolduc, B., Youens-Clark, K., Roux, S., Hurwitz, B.L., and Sullivan, M.B. (2017) iVirus: facilitating new insights in viral ecology with software and community data sets imbedded in a cyberinfrastructure. *ISME J* **11:** 7–14.

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120.

Bonetti, G., Trevathan-Tackett, S., Carnell, P.E., and Macreadie, P.I. (2019) Implication of Viral Infections for Greenhouse Gas Dynamics in Freshwater Wetlands: Challenges and Perspectives. *Front Microbiol* **10:** 1962.

Bonetti, G., Trevathan-Tackett, S., Carnell, P.E., and Macreadie, P.I. (2021) The potential of viruses to influence the magnitude of greenhouse gas emissions in an inland wetland. *Water Res* **193:** 116875.

Brown, T.C., and Irber, L. (2016) sourmash: a library for MinHash sketching of DNA. *Journal of Open Source Software* **1:** 27.

Campos, R.K., Boratto, P.V., Assis, F.L., Aguiar, Eric R. G. R., Silva, L.C.F., Albarnaz, J.D., *et al.* (2014) Samba virus: a novel mimivirus from a giant rain forest, the Brazilian Amazon. *Virol J* **11:** 95.

Carini, P., Marsden, P.J., Leff, J.W., Morgan, E.E., Strickland, M.S., and Fierer, N. (2016) Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. *Nature microbiology* **2:** 16242.

Casjens, S. (2003) Prophages and bacterial genomics: what have we learned so far? *Mol Microbiol* **49:** 277–300.

Challacombe, J., and Kuske, C. (2012) Mobile genetic elements in the bacterial phylum Acidobacteria. *Mob Genet Elements* **2:** 179–183.

Chen, F., Wang, K., Stewart, J., and Belas, R. (2006) Induction of Multiple Prophages from a Marine Bacterium: a Genomic Approach. *Appl Environ Microbiol* **72:** 4995–5001.

Chu H.Y., Fierer N., Lauber C.L., Caporaso J.G., Knight R., Grogan, P. (2010) Soil bacterial diversity in the Arctic is not fundamentally different from that found in other biomes. *Environ Microbiol* **12:** 2998–3006.

Cresawn, S.G., Pope, W.H., Jacobs-Sera, D., Bowman, C.A., Russell, D.A., Dedrick, R.M., *et al.* (2015) Comparative genomics of Cluster O mycobacteriophages. *PLOS ONE* **10:** e0118725.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience* **10:** 1–4

Darzentas, N. (2010) Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* **26:** 2620–2621.

DeAngelis, K.M., Brodie, E.L., DeSantis, T.Z., Andersen, G.L., Lindow, S.E., and Firestone, M.K. (2009) Selective progressive response of soil microbial community to wild oat roots. *The ISME Journal* **3:** 168–178.

Dedysh, S.N., Pankratov, T.A., Belova, S.E., Kulichevskaya, I.S., and Liesack, W. (2006) Phylogenetic Analysis and In Situ Identification of Bacteria Community Composition in an Acidic Sphagnum Peat Bog. *Appl Environ Microbiol* **72:** 2110–2117.

Dedysh, S.N., and Yilmaz, P. (2018) Refining the taxonomic structure of the phylum Acidobacteria. *Int J Syst Evol Microbiol* **68:** 3796–3806.

Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23:** 673–679.

Delmont, T.O., Quince, C., Shaiber, A., Esen, O.C., Lee, S.T.M., Rappe, M.S., *et al.* (2018) Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* **3:** 804–816.

Delmont, T.O., Robe, P., Clark, I., Simonet, P., and Vogel, T.M. (2011) Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods* **86:** 397–400.

Eddy, S.R. (1995) Multiple alignment using hidden Markov models. *Proc Int Conf Intell Syst Mol Biol* **3:** 114–120.

Eichorst, S.A., Kuske, C.R., and Schmidt, T.M. (2011) Influence of Plant Polymers on the Distribution and Cultivation of Bacteria in the Phylum Acidobacteria. *Appl Environ Microbiol* **77:** 586–596.

Eichorst, S.A., Trojan, D., Huntemann, M., Clum, A., Pillay, M., Palaniappan, K., *et al.* (2020) One Complete and Seven Draft Genome Sequences of Subdivision 1 and 3 Acidobacteria Isolated from Soil. *Microbiology Resource Announcements* **9:** e01087-19.

Eichorst, S.A., Trojan, D., Roux, S., Herbold, C., Rattei, T., and Woebken, D. (2018) Genomic insights into the Acidobacteria reveal strategies for their success in terrestrial environments: Genomic insights into acidobacteria. *Environ Microbiol* **20:** 1041–1063.

Emerson, J.B. (2019) Soil Viruses: A New Hope. *Appl Environ Sci* **4:** 120.

Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B.J., Jang, H.B., *et al.* (2018) Host-linked soil viral ecology along a permafrost thaw gradient **3:** 870–880.

Enav, H., Kirzner, S., Lindell, D., Mandel-Gutfreund, Y., and Béjà, O. (2018) Adaptation to sub-optimal hosts is a driver of viral diversification in the ocean. *Nat Commun* **9:** 4698.

Evans, C., and Brussaard, C.P.D. (2012) Regional Variation in Lytic and Lysogenic Viral Infection in the Southern Ocean and Its Contribution to Biogeochemical Cycling. *Appl Environ Microbiol* **78:** 6741–6748.

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., and Zhu, H. (2019) PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience* **8:** giz066.

Faoro, H., Alves, A.C., Souza, E.M., Rigo, L.U., Cruz, L.M., Al-Janabi, S., *et al.* (2010) Influence of Soil Characteristics on the Diversity of Bacteria in the Southern Brazilian Atlantic Forest. *Appl Environ Microbiol* **76:** 4744–4749.

Fierer, N. (2017) Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev* **15:** 579–590.

Fierer, N., Breitbart, M., Knight, R., Rohwer, F., Jackson, R.B., Nulton, J., *et al.* (2007) Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil. *Appl Environ Microbiol* **73:** 7059–7066.

Ganzert, L., Lipski, A., Hubberten, H., and Wagner, D. (2011) The impact of different soil parameters on the community structure of dominant bacteria from nine different soils located on Livingston Island, South Shetland Archipelago, Antarctica. *FEMS Microbiol Ecol* **76:** 476–491.

Ghosh, D., Roy, K., Williamson, K.E., White, D.C., Wommack, K.E., Sublette, K.L., and Radosevich, M. (2008) Prevalence of Lysogeny among Soil Bacteria and Presence of 16S rRNA and trzN Genes in Viral-Community DNA. *Appl Environ Microbiol* **74:** 495–502.

Goldsmith, D.B., Crosti, G., Dwivedi, B., McDaniel, L.D., Varsani, A., Suttle, C.A., *et al.* (2011) Development of phoH as a Novel Signature Gene for Assessing Marine Phage Diversity. *Appl Environ Microbiol* **77:** 7730–7739.

Gould, E.A. (1999) Methods for long-term virus preservation. *Mol Biotechnol* **13:** 57–66.

Goulden, M.L., Wofsy, S.C., Harden, J.W., Trumbore, S.E., Crill, P.M., Gower, S.T., *et al.* (1998) Sensitivity of boreal forest carbon balance to soil thaw. *Science* **279:** 214–217.

Griffiths, R.I., Whiteley, A.S., O'Donnell, A.G., and Bailey, M.J. (2000) Rapid Method for Coextraction of DNA and RNA from Natural Environments for Analysis of Ribosomal DNA- and rRNA-Based Microbial Community Composition. *Appl Environ Microbiol* **66:** 5488–5491.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., *et al.* (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532:** 465–470.

Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., *et al.* (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9:** 37.

Göller, P.C., Haro-Moreno, J., Rodriguez-Valera, F., Loessner, M.J., and Gómez-Sanz, E. (2020) Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil. *Microbiome* **8:** 1–17.

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., and Goodman, R.M. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5:** 245–249.

Hodgkins, S.B., Tfaily, M.M., McCalley, C.K., Logan, T.A., Crill, P.M., Saleska, S.R., *et al.* (2014) Changes in peat chemistry associated with permafrost thaw increase greenhouse gas production. *Proc Natl Acad Sci U S A* **111:** 5819–5824.

Howard-Varona, C., Roux, S., Dore, H., Solonenko, N.E., Holmfeldt, K., Markillie, L.M., *et al.* (2017) Regulation of infection efficiency in a globally abundant marine Bacteriodetes virus. *The ISME* J **11:** 284–295

Howe, A.C., Jansson, J.K., Malfatti, S.A., Tringe, S.G., Tiedje, J.M., and Brown, C.T. (2014) Tackling soil diversity with the assembly of large, complex metagenomes. *Proc Natl Acad Sci U S A* **111:** 4904–4909.

Hultman, J., Waldrop, M.P., Mackelprang, R., David, M.M., McFarland, J., Blazewicz, S.J., *et al.* (2015) Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature* **521:** 208–212.

Hurst, C.J., Gerba, C.P., and Cech, I. (1980) Effects of environmental variables and soil characteristics on virus survival in soil. *Appl Environ Microbiol* **40:** 1067–1079.

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

Hyman, P. (2019) Phages for Phage Therapy: Isolation, Characterization, and Host Range Breadth. *Pharmaceuticals (Basel).* **12:** 35.

Jagdale, S.S., and Joshi, R.S. (2018) Enemies with benefits: mutualistic interactions of viruses with lower eukaryotes. *Arch Virol* **163:** 821–830.

Jang, H.B., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., *et al.* (2019) Gene sharing networks to automate genome-based prokaryotic viral taxonomy. *Nat Biotechnol* **37:** 632–639.

Janssen, P.H. (2006) Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* **72:** 1719–1728.

Jiang, H., Song, W., Cheng, H., and Qiu, B. (2015) The hypothetical protein Ycf46 is involved in regulation of CO_2 utilization in the cyanobacterium Synechocystis sp. PCC 6803. *Planta* **241:** 145–155.

Jones, R.T., Robeson, M.S., Lauber, C.L., Hamady, M., Knight, R., and Fierer, N. (2009) A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *ISME J* **3:** 442–453.

Jurtz, V.I., Villarroel, J., Lund, O., Voldby Larsen, M., and Nielsen, M. (2016) MetaPhinder-Identifying Bacteriophage Sequences in Metagenomic Data Sets. *PLOS One* **11:** e0163111.

Kalam, S., Basu, A., Ahmad, I., Sayyed, R.Z., Hesham Ali El-Enshasy, Hesham Ali El-Enshasy, *et al.* (2020) Recent Understanding of Soil Acidobacteria and Their Ecological Significance: A Critical Review. *Front Microbiol* **11:** 580024.

Kieft, K., Zhou, Z., and Anantharaman, K. (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8:** 90.

Kim, D., Song, L., Breitwieser, F.P., and Salzberg, S.L. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **26:** 1721–1729.

Kimura, M., Jia, Z., Nakayama, N., and Asakawa, S. (2008) Ecology of viruses in soils: Past, present and future perspectives. *Soil Science and Plant Nutrition* **54:** 1–32.

Kishimoto, N., Kosako, Y., and Tano, T. (1991) Acidobacterium capsulatum gen. nov., sp. nov.: an acidophilic chemoorganotrophic bacterium containing menaquinone from acidic mineral environment. *Curr Microbiol* **22:** 1–7.

Kuzyakov, Y., and Mason-Jones, K. (2018) Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biology and Biochem* **127:** 305–317.

Langmead, B., and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9:** 357–359.

Lee, S.H., and Cho, J.C. (2009) Distribution Patterns of the Members of Phylum Acidobacteria in Global Soil Samples. *Journal of microbiol and biotechnol* **19:** 1281–7.

Lee, A.J., Endo, M., Hobbs, J.K., and Wälti, C. (2018) Direct Single-Molecule Observation of Mode and Geometry of RecA-Mediated Homology Search. *ACS nano* **12:** 272–278.

Lehmann, J., and Kleber, M. (2015) The contentious nature of soil organic matter. *Nature* **528:** 60–68.

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

Levy-Booth, D., Campbell, R.G., Gulden, R.H., Hart, M.M., Powell, J.R., Klironomos, J.N., *et al.* (2007) Cycling of extracellular DNA in the soil environment. *Soil biology & biochemistry* **39:** 2977–2991.

Li, D., Liu, C., Luo, R., Sadakane, K., and Lam, T. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31:** 1674–1676.

Li, D., Luo, R., Liu, C., Leung, C., Ting, H., Sadakane, K., *et al.* (2016) MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102:** 3–11.

Liang, X., Wagner, R.E., Zhuang, J., DeBruyn, J.M., Wilhelm, S.W., Liu, F., *et al.* (2019) Viral abundance and diversity vary with depth in a southeastern United States agricultural ultisol. *Soil Biology & Biochem* **137:** 107546.

Liang, X., Zhang, Y., Wommack, K.E., Wilhelm, S.W., DeBruyn, J.M., Sherfy, A.C., *et al.* (2020) Lysogenic reproductive strategies of viral communities vary with soil depth and are correlated with bacterial diversity. *Soil biology & biochem* **144:** 107767.

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008) Reticulate Representation of Evolutionary and Functional Relationships between Phage Genomes. *Mol Biol Evol* **25:** 762–777.

Lloyd, K.G., Steen, A.D., Ladau, J., Yin, J., and Crosby, L. (2018) Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *mSystems* **3:** 5.

Lopes, A., Tavares, P., Petit, M., Guérois, R., and Zinn-Justin, S. (2014) Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics* **15:** 1027.

Luo, C., Rodriguez-R, L.M., Johnston, E.R., Wu, L., Cheng, L., Xue, K., *et al.* (2014) Soil Microbial Community Responses to a Decade of Warming as Revealed by Comparative Metagenomics. *Appl Environ Microbiol* **80:** 1777–1786.

Mackelprang, R., Waldrop, M.P., Deangelis, K.M., David, M.M., Chavarria, K.L., BLAZEWICZ, S.J., *et al.* (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* **480:** 368–371.

Malathi, V.G., and Renuka Devi, P. (2019) ssDNA viruses: key players in global virome. *VirusDis* **30:** 3–12.

Marquet, M., Hölzer, M., Pletz, M.W., Viehweger, A., Makarewicz, O., Ehricht, R., and Brandt, C. (2020) What the Phage: A scalable workflow for the identification and analysis of phage sequences. Preprint

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17:** 10–12.

Maurice, C.F., Bouvier, T., Comte, J., Guillemette, F., and Del Giorgio, P.A. (2010) Seasonal variations of phage life strategies and bacterial physiological states in three northern temperate lakes. *Environ Microbiol* **12:** 628–641.

Michen, B., and Graule, T. (2010) Isoelectric points of viruses. *J Appl Microbiol* **109:** 388–397.

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016) MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* **32:** 1088–1090.

Männistö, M.K., Ganzert, L., Tiirola, M., Häggblom, M.M., and Stark, S. (2016) Do shifts in life strategies explain microbial community responses to increasing nitrogen in tundra soil?. *Soil Biology and Biochemistry* **96:** 216–228.

Männistö, M.K., and Häggblom, M.M. (2006) Characterization of psychrotolerant heterotrophic bacteria from Finnish Lapland. *Syst and Appl Microbiol* **29**: 229–243.

Männistö, M.K., Rawat, S., Starovoytov, V., and Häggblom, M.M. (2011) Terriglobus saanensis sp. nov., an acidobacterium isolated from tundra soil. *Int J Syst Evol Microbiol* **61:** 1823–1828.

Männistö, M.K., Rawat, S., Starovoytov, V., and Häggblom, M.M. (2012) Granulicella arctica sp. nov., Granulicella mallensis sp. nov., Granulicella tundricola sp. nov. and Granulicella sapmiensis sp. nov., novel acidobacteria from tundra soil. *Int J Syst Evol Microbiol* **62:** 2097–2106.

Männistö, M.K., Tiirola, M., and Haggblom, M.M. (2009) Effect of Freeze-Thaw Cycles on Bacterial Communities of Arctic Tundra Soil. *Microb Ecol* **58:** 621–631.

Männistö, M.K., Tiirola, M., and Häggblom, M.M. (2007) Bacterial communities in Arctic fjelds of Finnish Lapland are stable but highly pH-dependent. *FEMS Microbiol Ecol* **59:** 452–465.

Nakayama, N., Okumura, M., Inoue, K., Asakawa, S., and Kimura, M. (2007) Seasonal variations in the abundance of virus-like particles and bacteria in the floodwater of a Japanese paddy field. *Soil science & plant nutrition* **53:** 420–429.

Narr, A., Nawaz, A., Wick, L.Y., Harms, H., and Chatzinotas, A. (2017) Soil Viral Communities Vary Temporally and along a Land Use Transect as Revealed by Virus-Like Particle Counting and a Modified Community Fingerprinting Approach (fRAPD). *Front Microbiol* **8:** 1975.

Nauta, A.L., Heijmans, M. M. P. D, Blok, D., Limpens, J., Elberling, B., Gallagher, A., *et al.* (2015) Permafrost collapse after shrub removal shifts tundra ecosystem to a methane source. *Nature Climate Change* **5:** 67–70.

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39:** 578–585.

Neufeld, J.D., and Mohn, W.W. (2005) Unexpectedly High Bacterial Diversity in Arctic Tundra Relative to Boreal Forest Soils, Revealed by Serial Analysis of Ribosomal Sequence Tags. *Appl Environ Microbiol* **71:** 5710–5718.

Noguchi, H., Park, J., and Takagi, T. (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* **34:** 5623–5630.

Noguchi, H., Taniguchi, T., and Itoh, T. (2008) MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes. *DNA Res* **15:** 387–396.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27:** 824–834.

Olson, M.R., Axler, R.P., and Hicks, R.E. (2004) Effects of freezing and storage temperature on MS2 viability. *J Virol Methods* **122:** 147–152.

Paez-Espino, D., Eloe-Fadrosh, E., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., *et al.* (2016) Uncovering Earth's virome. *Nature* **536:** 425–430.

Pankratov, T.A., Serkebaeva, Y.M., Kulichevskaya, I.S., Liesack, W., and Dedysh, S.N. (2008) Substrate-induced growth and isolation of Acidobacteria from acidic Sphagnum peat. *The ISME J* **2:** 551–560.

Pankratov, T.A., Ivanova, A.O., Dedysh, S.N., and Liesack, W. (2011) Bacterial populations and environmental factors controlling cellulose degradation in an acidic Sphagnum peat. *Environ Microbiol* **13:** 1800–1814.

Paul, J.H., Jiang, S.C., and Rose, J.B. (1991) Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration. *Appl Environ Microbiol* **57:** 2197–2204.

Payet, J.P., and Suttle, C.A. (2013) To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. *Limnol Oceanogr* **58:** 465–474.

Peltola, H., Söderlund, H., and Ukkonen, E. (1984) SEQAID: a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Res* **12:** 307–321.

Pessi, I., Viitamäki, S., Virkkala, A.M., Eronen-Rasimus, E., Delmont, T.O., Marushchak, M.E., Luoto, M. and Hultman, J. (2021) Truncated denitrifiers dominate the denitrification pathway in tundra soil metagenomes. *bioRxiv.* Preprint

Plyusnin, I., Kant, R., Jääskeläinen, A.J., Sironen, T., Holm, L., Vapalahti, O., Smura, T. (2020) Novel NGS pipeline for virus discovery from a wide spectrum of hosts and sample types. *Virus Evolution* **6:** veaa091.

Post, E., Alley, R.B., Christensen, T.R., Macias-Fauria, M., Forbes, B.C., Gooseff, M.N., *et al.* (2019) The polar regions in a 2°C warmer world. *Sci Adv* **5:** eaaw9883.

Pradeu, T. (2016) Mutualistic viruses and the heteronomy of life. *Studies in History and Philosophy of Biological and Biomedical Sciences* **59:** 80–88.

Pratama, A.A., and van Elsas, J.D. (2018) The 'Neglected' Soil Virome – Potential Role and Impact. *Trends Microbiol* **26:** 649–662.

Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J., and Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* **35:** 833–844.

Quirós, P., and Muniesa, M. (2017) Contribution of cropland to the spread of Shiga toxin phages and the emergence of new Shiga toxin-producing strains. *Sci Rep* **7:** 7796.

Rawat, S.R., Männistö, M.K., Bromberg, Y., and Häggblom, M.M. (2012) Comparative genomic and physiological analysis provides insights into the role of Acidobacteria in organic carbon utilization in Arctic tundra soils. *FEMS Microbiol Ecol* **82:** 341–355.

Rawat, S.R., Männistö, M.K., Starovoytov, V., Goodwin, L., Nolan, M., Hauser, L., *et al.* (2014) Complete genome sequence of Granulicella tundricola type strain MP5ACTX9(T), an Acidobacteria from tundra soil. *Stand Genomic Sci.* **9:** 449–461.

Ren, J., Ahlgren, N.A., Lu, Y.Y., Fuhrman, J.A., and Sun, F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5:** 69.

Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16:** 276–277.

Rivkina, E.M., Friedmann, E.I., McKay, C.P., and Gilichinsky, D.A. (2000) Metabolic Activity of Permafrost Bacteria below the Freezing Point. *Appl Environ Microbiol* **66:** 3230–3233.

Roux, S. (2019) A Viral Ecogenomics Framework To Uncover the Secrets of Nature's "Microbe Whisperers". *mSystems* **4:** 111.

Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015a) VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3:** e985.

Roux, S., Hallam, S.J., Woyke, T., and Sullivan, M.B. (2015b) Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife* **4:** e08490.

Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2019) GenBank. *Nucleic acids research* **47:** D94–D99.

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford)* 2020: baaa062.

Schloss, P.D., and Handelsman, J. (2003) Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* **14:** 303–310.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13:** 2498–2504.

Shi, M., Lin, X., Tian, J., Chen, L., Chen, X., Li, C., *et al.* (2016) Redefining the invertebrate RNA virosphere. *Nature* **540:** 539–543.

Simón, D., Cristina, J., and Musto, H. (2021) Nucleotide Composition and Codon Usage Across Viruses and Their Respective Hosts. *Front Microbiol* **12:** 646300.

Starikova, E.V., Tikhonova, P.O., Prianichnikov, N.A., Rands, C.M., Zdobnov, E.M., Ilina, E.N., and Govorun, V.M. (2020) Phigaro: high-throughput prophage sequence annotation. *Bioinformatics* **36:** 3882–3884.

Starr, E.P., Nuccio, E.E., Pett-Ridge, J., Banfield, J.F., and Firestone, M.K. (2019) Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil. *Proc Natl Acad Sci U S A* **116:** 25900–25908.

Stough, J.M.A., Kolton, M., Kostka, J.E., Weston, D.J., Pelletier, D.A., and Wilhelm, S.W. (2018) Diversity of Active Viral Infections within the Sphagnum Microbiome. *Appl Environ Microbiol* **84:** e01124-18.

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

Sutela, S., Poimala, A., and Vainio, E.J. (2019) Viruses of fungi and oomycetes in the soil environment. *FEMS Microbiol Ecol* **95:** fiz119.

Suttle, C.A. (2007) Marine viruses - major players in the global ecosystem. *Nat Rev Microbiol* **5:** 801–812.

Tarnocai, C., Canadell, J.G., Schuur, E.A.G., Kuhry, P., Mazhitova, G., and Zimov, S. (2009) Soil organic carbon pools in the northern circumpolar permafrost region. *Global Biogeochem Cycles* **23:** GB2023-n/a.

Tartera, C., Araujo, R., Michel, T. and Jofre J. (1993) Culture and Decontamination Methods Affecting Enumeration of Phages Infecting Bacteroides fragilis in Sewage. *Appl Environ Microbiol* **59:** 351.

Thomas, T., Gilbert, J., and Meyer, F. (2012) Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* **2:** 3.

Thrash, J.C., and Coates, J.D. (2015) Edaphobacter. Bergey's Manual of Systematics of Archaea and Bacteria. Chichester (7st ed.). UK: John Wiley & Sons, Ltd. 1–3 p.

Torsvik, V., Goksøyr, J., and Daae, F.L. (1990) High diversity in DNA of soil bacteria. *Appl Environ Microbiol* **56:** 782–787.

Trubl, G., Hyman, P., Roux, S., and Abedon, S.T. (2020) Coming-of-Age Characterization of Soil Viruses: A User's Guide to Virus Isolation, Detection within Metagenomes, and Viromics. *Soil Syst.* **4:** 1–23.

Trubl, G., Jang, H.B., Roux, S., Emerson, J.B., Solonenko, N., Vik, D.R., *et al.* (2018) Soil Viruses Are Underexplored Players in Ecosystem Carbon Processing. *mSystems* **3:**.

Trubl, G., Kimbrel, J. A., Liquet-Gonzalez, J., Nuccio, E. E., Weber, P. K., Pett-Ridge, J., Jansson, J. K., Waldrop, M. P., & Blazewicz, S. J. (2021). Active virus-host interactions at sub-freezing temperatures in Arctic peat soil. *Microbiome*, **9:** 208.

Trubl, G., Roux, S., Solonenko, N., Li, Y., Bolduc, B., Rodríguez-Ramos, J., *et al.* (2019) Towards optimized viral metagenomes for double-stranded and single-stranded DNA viruses from challenging soils. *PeerJ* **7:** e7265.

Trubl, G., Solonenko, N., Chittick, L., Solonenko, S.A., Rich, V.I., and Sullivan, M.B. (2016) Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. *PeerJ* **4:** e1999.

Vainio, E.J., Pennanen, T., Rajala, T., and Hantula, J. (2017) Occurrence of similar mycoviruses in pathogenic, saprotrophic and mycorrhizal fungi inhabiting the same forest stand. *FEMS Microbiol Ecol* **93:** fix003.

van der Walt, A.J, van Goethem, M.W., Ramond, J., Makhalanyane, T.P., Reva, O., and Cowan, D.A. (2017) Assembling metagenomes, one community at a time. *BMC Genomics* **18:** 521.

Viitamäki, S. (2019) The activity and functions of soil microbial communities across a climate gradient in Finnish subarctic. University of Helsinki

Ward, N.L., Challacombe, J.F., Janssen, P.H., Henrissat, B., Coutinho, P.M., Wu, M., *et al.* (2009) Three Genomes from the Phylum Acidobacteria Provide Insight into the Lifestyles of These Microorganisms in Soils. *Appl Environ Microbiol* **75:** 2046–2056.

Williamson, K.E., Corzo, K.A., Drissi, C.L., Buckingham, J.M., Thompson, C.P., and Helton, R.R. (2013) Estimates of viral abundance in soils are strongly influenced by extraction and enumeration methods. *Biol Fertil Soils* **49:** 857–869.

Williamson, K.E., Helton, R.R., and Wommack, K.E. (2012) Bias in bacteriophage morphological classification by transmission electron microscopy due to breakage or loss of tail structures. *Microsc Res Tech* **75:** 452–457.

Williamson, K.E., Radosevich, M., Smith, D.W., and Wommack, K.E. (2007) Incidence of lysogeny within temperate and extreme soil environments. *Environ Microbiol* **9:** 2563–2574.

Williamson, K.E., Radosevich, M., and Wommack, K.E. (2005) Abundance and Diversity of Viruses in Six Delaware Soils. *Appl Environ Microbiol* **71:** 3119–3125.

Williamson, K.E., Schnitker, J.B., Radosevich, M., Smith, D.W., and Wommack, K.E. (2008) Cultivation-Based Assessment of Lysogeny among Soil Bacteria. *Microb Ecol* **56:** 437–447.

Williamson, K.E., Wommack, K.E., and Radosevich, M. (2003) Sampling Natural Viral Communities from Soil for Culture-Independent Analyses. *Appl Environ Microbiol* **69:** 6628–6633.

Wommack, K.E., and Colwell, R.R. (2000) Virioplankton: Viruses in Aquatic Ecosystems. *Microbiol Mol Biol Rev* **64:** 69–114.

Woodcroft, B.J., Singleton, C.M., Boyd, J.A., Evans, P.N., Emerson, J.B., Zayed, A.A.F., *et al.* (2018) Genome-centric view of carbon processing in thawing permafrost. *Nature* **560:** 49–54.

World Meteorological Organization (2021) State of the Global Climate in 2020. WMO-No. 1264. Geneva. 54 p.

Wu, R., Davison, M.R., Gao, Y., Nicora, C.D., McDermott, J.E., Burnum-Johnson, K., *et al.* (2021) Moisture modulates soil reservoirs of active DNA and RNA viruses. *Commun Biol* **4:** 992.

Zablocki, O., van Zyl, L., Adriaenssens, E.M., Rubagotti, E., Tuffin, M., Cary, S.C., and Cowan, D. (2014) High-level diversity of tailed phages, eukaryote-associated viruses, and virophage-like elements in the metaviromes of antarctic soils. *Appl Environ Microbiol* **80:** 6888–6897.

Zhang, J., Gao, Q., Zhang, Q., Wang, T., Yue, H., Wu, L., *et al.* (2017) Bacteriophage-prokaryote dynamics and interaction within anaerobic digestion processes across time and space. *Microbiome* **5:** 57.

Zheng, L., Zhang, M., Chen, Q., Zhu, M., and Zhou, E. (2014) A novel mycovirus closely related to viruses in the genus Alphapartitivirus confers hypovirulence in the phytopathogenic fungus Rhizoctonia solani. *Virology* **456:** 220–226.

Supplements

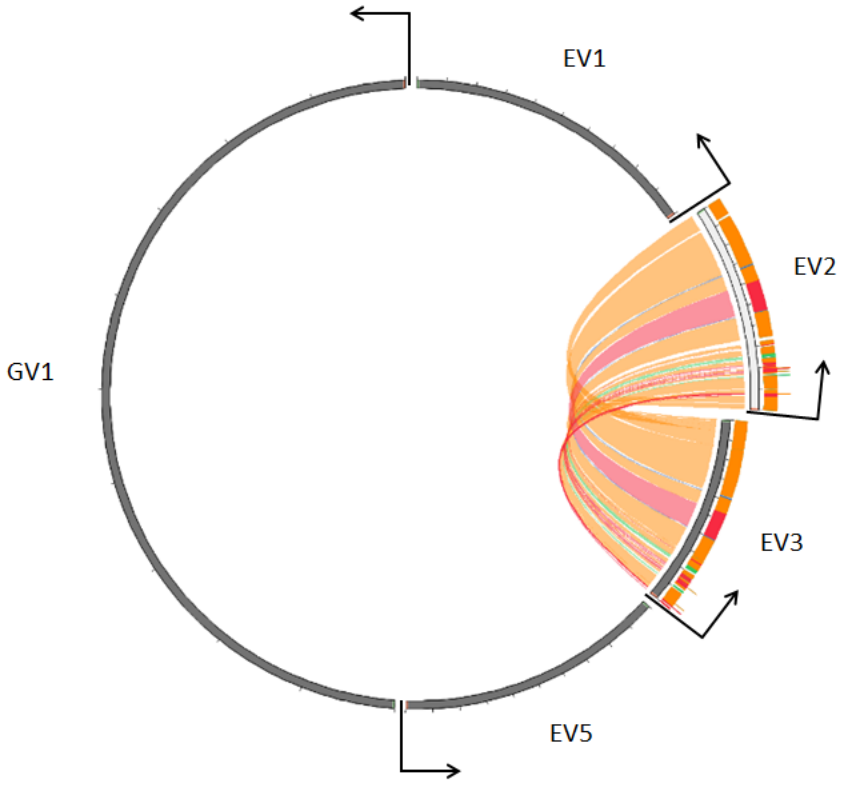


Figure S1. Visualising the five virus sequence similarities using Circoletto (Darzentas et al. 2010). Ribbons are coloured by % identity with absolute colouring: blue ≤90%, green ≤95%, orange ≤99%, red >99% identity. The orientation of sequences is shown by arrows.

Table S1. EV1 ORFs and their putative functions.

EV1							
Name	Start	End	Length	Direction	ORF product	Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 29.8.2021)	Putative function:
ORF1	1	2061	2061	forward	gp1	QGH73314.1, terminase large subunit [Siphoviridae sp. cttb18], 88.6%, 23.81%, 8.38e-16	Terminase large subunit
ORF2	2073	4205	2133	forward	gp2	CAB4142966.1, hypothetical protein UFOVP434_56 [uncultured Caudovirales phage], 95.5%, 25.33%, 1.08e-17	Portal protein
ORF3	4422	4796	375	forward	gp3		
ORF4	4786	5817	1032	forward	gp4	PWT75549, hypothetical protein C5B59_08705 [Bacteroidetes bacterium], 53.49 %, 33.7 %, 1.92e-18	
ORF5	5821	6246	426	forward	gp5	WP_048812770, GNAT family N-acetyltransferase [Acidilobus saccharovorans], 99.30 %, 27.9 %, 3.00e-8	GNAT family N-acetyltransferase
ORF6	6469	7500	1032	forward	gp6	CAB4142941.1, hypothetical protein UFOVP434_53 [uncultured Caudovirales phage], -, 26.54%, 7.00e-23	Major capsid protein
ORF7	7573	8511	939	forward	gp7		
ORF8	8550	8834	285	forward	gp8		
ORF9	9051	9527	477	forward	gp9	PWT75543, hypothetical protein C5B59_08675 [Bacteroidetes bacterium], 74.84 %, 31.9 %, 3.02e-13	
ORF10	9543	10208	666	forward	gp10	PWT75544, hypothetical protein C5B59_08680 [Bacteroidetes bacterium], 92.34 %, 53.7 %, 5.54e-72	
ORF11	10225	10893	669	forward	gp11	PWT75543, hypothetical protein C5B59_08675 [Bacteroidetes bacterium], 63.23 %, 34.8 %, 7.18e-17	
ORF12	10964	12079	1116	forward	gp12	PWT75542, hypothetical protein C5B59_08670 [Bacteroidetes bacterium], 99.19 %, 51.5 %, 5.08e-123	
ORF13	12089	13003	915	forward	gp13	PWT75539, hypothetical protein C5B59_08655 [Bacteroidetes bacterium], 98.69%, 36.1%, 1.49e-51	
ORF14	13003	13290	288	forward	gp14	PWT75538, hypothetical protein C5B59_08650 [Bacteroidetes bacterium], 77.08%, 35.1%, 5.94e-4	
ORF15	13298	14779	1482	forward	gp15	PWT75537, hypothetical protein C5B59_08645 [Bacteroidetes bacterium], 19.03 %, 55.2 %, 4.98e-19	
ORF16	14798	15538	741	forward	gp16		
ORF17	15549	15683	135	forward	gp17		
ORF18	15693	18173	2481	forward	gp18	QOC54124.1, tail fiber protein [Teseptimavirus S2B], 66.6%, 40.5%, 3.59e-09	Tail fiber protein
ORF19	18182	21592	3411	forward	gp19	PWT75536, hypothetical protein C5B59_08640 [Bacteroidetes bacterium], 62.27 %, 46.7 %, 2.24e-176	
ORF20	21600	22304	705	forward	gp20	BAR32946.1, long tail fiber proximal subunit [uncultured Mediterranean phage uvMED], 50.8%, 31.42%, 9.84e-05	Long tail fiber proximal subunit
ORF21	22304	23095	792	forward	gp21		
ORF22	23103	23501	399	forward	gp22		
ORF23	23498	24262	765	forward	gp23	MBS1799559, hypothetical protein [Acidobacteria bacterium], 65.88%, 54.2%, 4.12e-31	
ORF24	24270	26663	2394	forward	gp24	MBW4039200, hypothetical protein [Acidobacteria bacterium], 35.71 %, 28.6 %, 1.39e-6	
ORF25	26674	29886	3213	forward	gp25	WP_089838633, hypothetical protein [Granulicella pectinivorans], 55.18%, 38.0%, 2.33e-94	
ORF26	30004	30315	312	forward	gp26	PWT75524, hypothetical protein C5B59_08580 [Bacteroidetes bacterium], 94.23%, 38.1%, 2.66e-12	
ORF27	30456	30671	216	forward	gp27		
ORF28	30655	31068	414	forward	gp28		
ORF29	31139	32497	1359	forward	gp29	APU89236, phosphoesterase-like protein [Virus Rctr41k], 88.3%, 32.0%, 2.23e-54	Phosphoesterase-like protein
ORF30	32512	32775	264	forward	gp30	MBW4039197, hypothetical protein [Acidobacteria bacterium], 93.18%, 52.4%, 6.44e-21	
ORF31	33005	33313	309	forward	gp31	QHJ75341.1, hypothetical protein SnaR1_gp27 [Sphaerotilus phage vB_SnaP-R1], -, 67.86%, 2.67e-35	RNA binding protein
ORF32	33377	33559	183	forward	gp32		
ORF33	33562	33765	204	forward	gp33		
ORF34	33819	34586	768	forward	gp34	VVB52887, Uncharacterised protein [uncultured archaeon], 89.84%, 26.8%, 1.08e-4	
ORF35	34600	36279	1680	forward	gp35	YP_009831807.1, ribonucleotide-diphosphate reductase [Streptomyces phage BRock], -, 54.03%, 0	Ribonucleoside-diphosphate reductase
ORF36	36384	37052	669	reverse	gp36		
ORF37	37040	37309	270	reverse	gp37		
ORF38	37309	38436	1128	reverse	gp38	WP_193333366, AAA family ATPase [Duganella sp. FT27W], 95.48%, 39.3%, 6.55e-74	AAA family ATPase
ORF39	38483	38677	195	reverse	gp39		
ORF40	38674	39021	348	reverse	gp40	BAR35259.1, pyrophosphatase [uncultured Mediterranean phage uvMED], -, 49.57%, 2.40e-30	Pyrophosphatase
ORF41	39008	39340	333	reverse	gp41	QEM41504, hypothetical protein SEA_BOOPY_35 [Gordonia phage Boopy], 61.26 %, 32.9 %, 8.09e-4	
ORF42	39340	39639	300	reverse	gp42		
ORF43	39636	39797	162	reverse	gp43		
ORF44	39798	40262	465	reverse	gp44	WP_121470746, lysozyme [Edaphobacter dinghuensis], 97.42 %, 44.2 %, 1.28e-35	Lysozyme
ORF45	40259	40405	147	reverse	gp45		
ORF46	40437	40994	558	reverse	gp46	ABO60549, hypothetical protein Bcep1808_7679 [Burkholderia vietnamiensis G4], 64.52 %, 30.6 %, 4.67e-6	
ORF47	40978	41625	648	reverse	gp47	VIP05982, Uncharacterized protein OS=Sphingobium yanoikuyae [Gemmataceae bacterium], 67.59 %, 40.1 %, 2.61e-21	
ORF48	41674	41793	120	forward	gp48		
ORF49	41774	41926	153	reverse	gp49		
ORF50	41923	42039	117	reverse	gp50		

EV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 29.8.2021)**	**Putative function:**
ORF51	42137	42649	513	reverse	gp51		
ORF52	42646	42888	243	reverse	gp52		
ORF53	43093	43518	426	reverse	gp53		
ORF54	43570	43818	249	reverse	gp54	WP_073631414, hypothetical protein [Pseudoxanthobacter soli], 90.36 %, 41.3 %, 2.57e-5	
ORF55	43818	44015	198	reverse	gp55		
ORF56	44015	44188	174	reverse	gp56	WP_151617148, hypothetical protein [Bacillus cereus], 93.10 %, 50 %, 6.39e-7	
ORF57	44188	45075	888	reverse	gp57	DAV08528.1, TPA: MAG TPA: Protein recA [Myoviridae sp.], -, 58.63%, 1.65e-109	Recombinase RecA
ORF58	45203	45946	744	reverse	gp58		
ORF59	45946	46134	189	reverse	gp59		
ORF60	46097	46570	474	reverse	gp60	PWT76409, hypothetical protein C5B59_06705 [Bacteroidetes bacterium], 59.49 %, 30.5 %, 1.05e-4	
ORF61	46940	47518	579	reverse	gp61		
ORF62	49341	49517	177	reverse	gp62	DAP97983.1, TPA: MAG TPA: DNA directed DNA polymerase [Siphoviridae sp.], -, 25.23%, 5.44e-30	DNA polymerase
ORF63	47518	49311	1794	reverse	gp63		
ORF64	49734	49991	258	reverse	gp64		
ORF65	50161	50322	162	reverse	gp65		
ORF66	50319	50696	378	reverse	gp66		
ORF67	50823	52049	1227	reverse	gp67	DAJ42433.1, TPA: MAG TPA: RNA ligase [Myoviridae sp.], 92.4%, 23.83%, 5.76e-18	RNA ligase
ORF68	52107	52277	171	reverse	gp68		
ORF69	52277	52438	162	reverse	gp69	RTL06189, hypothetical protein EKK58_06115 [Candidatus Dependentiae bacterium], 98.15 %, 43.4 %, 1.48e-4	
ORF70	52456	52791	336	reverse	gp70		
ORF71	52792	53013	222	reverse	gp71		
ORF72	53645	54406	762	reverse	gp72	PWT76404, hypothetical protein C5B59_06675 [Bacteroidetes bacterium], 91.94 %, 36.1 %, 2.00e-26	DEAD/DEAH box helicase
ORF73	53010	53567	558	reverse	gp73	CAB4142576.1, SSL2 DNA or RNA helicases of superfamily II [uncultured Caudovirales phage], -, 43.15%, 5.01e-55	
ORF74	54613	54873	261	reverse	gp74		
ORF75	55248	55556	309	reverse	gp75		
ORF76	55526	55696	171	reverse	gp76		
ORF77	55680	56444	765	reverse	gp77	PWT76397, hypothetical protein C5B59_06640 [Bacteroidetes bacterium], 84.31 %, 37.2 %, 6.72e-34	
ORF78	56643	57122	480	reverse	gp78	DAK53723.1, TPA: MAG TPA: nucleotidase 5'-nucleotidase [Siphoviridae sp.], 63.9%, 35.08%, 3.35e-11	Nucleotidase 5'-nucleotidase
ORF79	57112	57315	204	reverse	gp79		
ORF80	57272	58462	1191	reverse	gp80	QGH79931, RNA ligase [Streptomyces phage Bordeaux], 97.73%, 29.3%, 4.00e-26	RNA ligase
ORF81	58474	59316	843	reverse	gp81	PWT76392, hypothetical protein C5B59_06615 [Bacteroidetes bacterium], 76.51%, 32.3%, 2.05e-28	
ORF82	59395	59550	156	reverse	gp82		
ORF83	59537	60715	1179	reverse	gp83	PWT76391, hypothetical protein C5B59_06610 [Bacteroidetes bacterium], 81.68%, 42%, 5.67e-43	Exonuclease
ORF84	60748	61287	540	reverse	gp84		
ORF85	61551	61736	186	reverse	gp85		
ORF86	61774	62004	231	reverse	gp86		
ORF87	62017	62487	471	reverse	gp87		
ORF88	62584	62811	228	reverse	gp88		
ORF89	62811	63359	549	reverse	gp89	DAU63467.1, TPA: MAG TPA: SITE SPECIFIC RECOMBINASE XERD [Myoviridae sp.]63.9%, 37.04%, 8.83e-10	Site specific recombinase
ORF90	63501	63719	219	reverse	gp90		
ORF91	63716	63868	153	reverse	gp91		
ORF92	64044	64178	135	reverse	gp92		
ORF93	64369	64584	216	reverse	gp93		
ORF94	64650	65207	558	reverse	gp94		
ORF95	65209	65394	186	reverse	gp95		
ORF96	65426	65806	381	reverse	gp96		
ORF97	65983	66156	174	reverse	gp97		
ORF98	66252	66680	429	reverse	gp98		
ORF99	66754	67287	534	reverse	gp99	WP_135908499, hypothetical protein [Mesorhizobium sp. M4B.F.Ca.ET.143.01.1.1], 84.83 %, 33.8 %, 3.78e-9	

EV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 29.8.2021)**	**Putative function:**
ORF100	67395	68354	960	reverse	gp100	RMH38352, hypothetical protein D6690_00090 [Nitrospirae bacterium], 96.56 %, 32.3 %, 3.55e-37	
ORF101	68383	68523	141	reverse	gp101		
ORF102	68870	69079	210	reverse	gp102		
ORF103	69063	69392	330	reverse	gp103	PWT76232, hypothetical protein C5B59_07110 [Bacteroidetes bacterium], 96.36 %, 38.7 %, 7.63e-16	
ORF104	69379	69657	279	reverse	gp104		
ORF105	69751	69933	183	reverse	gp105	TAE87551, DUF2997 domain-containing protein [Verrucomicrobia bacterium], 83.61 %, 38.5 %, 2.86e-4	
ORF106	69973	70377	405	reverse	gp106	MBT3307776, DUF1257 domain-containing protein [Alphaproteobacteria bacterium], 90.37 %, 33.1 %, 2.19e-8	
ORF107	70454	72049	1596	reverse	gp107	DAW88856.1, TPA: MAG TPA: Ycf46 [Podoviridae sp.], 83.2%, 23.48%, 2.94e-14	ATPase Ycf46
ORF108	72661	72870	210	reverse	gp108	DAY87167, TPA: MAG TPA: Deformylase, HYDROLASE [Siphoviridae sp.], 70.91 %, 32 %, 1.75e-4	Deformylase
ORF109	72119	72613	495	reverse	gp109		
ORF110	72892	73128	237	reverse	gp110		
ORF111	73197	73349	153	reverse	gp111		
ORF112	73425	73601	177	reverse	gp112		
ORF113	73674	74042	369	reverse	gp113		
ORF114	74045	74203	159	reverse	gp114		
ORF115	74206	74400	195	reverse	gp115		
ORF116	74402	74998	597	reverse	gp116	YP_010062387, hypothetical protein KIW74_gp04 [Mycobacterium phage Kimona], 89.45 %, 36.7 %, 3.34e-24	Glycosylase
ORF117	74998	75276	279	reverse	gp117	MBW1931439, hypothetical protein [Deltaproteobacteria bacterium], 93.55 %, 33.3 %, 4.49e-10	
ORF118	75345	75557	213	reverse	gp118		
ORF119	75634	75828	195	reverse	gp119		
ORF120	75815	76471	657	reverse	gp120	CAB5219167, hypothetical protein UFOVP229_35 [uncultured Caudovirales phage], 48.86 %, 32.5 %, 5.76e-4	
ORF121	76608	76994	387	reverse	gp121	CAB4156328, hypothetical protein UFOVP663_57 [uncultured Caudovirales phage], 94.57 %, 35.9 %, 3.90e-8	
ORF122	77299	77661	363	reverse	gp122		
ORF123	77738	78442	705	reverse	gp123	WP_065815571, hypothetical protein [Nitratireductor aquibiodomus], 94.89 %, 32.9 %, 1.13e-31	
ORF124	78524	78676	153	reverse	gp124		
ORF125	80337	80603	267	reverse	gp125		
ORF126	80600	82018	1419	reverse	gp126		
ORF127	82028	82303	276	reverse	gp127		
ORF128	82469	82807	339	reverse	gp128	MBU2249439, peptide chain release factor-like protein [Gammaproteobacteria bacterium], 97.35 %, 45.5 %, 1.86e-24	Peptide chain release factor-like protein
ORF129	82794	83033	240	reverse	gp129		
ORF130	83047	83277	231	reverse	gp130		
ORF131	83380	84048	669	reverse	gp131	CAB4214602, hypothetical protein UFOVP1454_53 [uncultured Caudovirales phage], 16.59 %, 81.1 %, 1.11e-7	
ORF132	84148	84297	150	forward	gp132		
ORF133	84443	84655	213	reverse	gp133		
ORF134	84677	85495	819	reverse	gp134		
ORF135	85495	86142	648	reverse	gp135	NIQ80602, phosphoesterase [Anaerolineae bacterium], 90.74 %, 34.2 %, 1.06e-25	Phosphoesterase
ORF136	86898	87707	810	reverse	gp136	WP_056763053, SPFH domain-containing protein [Rhodanobacter sp. Root561], 98.15 %, 52.6 %, 1.19e.89	
ORF137	88419	88634	216	reverse	gp137		
ORF138	88624	88827	204	reverse	gp138		
ORF139	88864	88995	132	reverse	gp139		
ORF140	89234	89641	408	reverse	gp140		
ORF141	89641	92415	2775	reverse	gp141	PWT75565, hypothetical protein C5B59_08785 [Bacteroidetes bacterium], 99.89 %, 38 %, 0	
ORF142	92504	92686	183	reverse	gp142		
ORF143	92673	93131	459	reverse	gp143	QBK93949.1, ribonuclease H [Pithovirus LCPAC406], 53.9%, 28.97%, 2.25e-07	Ribonuclease H
ORF144	93785	94165	381	forward	gp144		
ORF145	94774	95220	447	forward	gp145	TFG23847, hypothetical protein EU532_13090 [Candidatus Lokiarchaeota archaeon], 65.1 %, 27.1 %, 7.07e-4	
ORF146	95228	95539	312	forward	gp146		
ORF147	95541	95759	219	forward	gp147	PWT75557, hypothetical protein C5B59_08745 [Bacteroidetes bacterium], 58.9 %, 56.8 %, 1.52e-5	
ORF148	95856	95984	129	forward	gp148		
ORF149	95994	96482	489	forward	gp149		
ORF150	96572	97081	510	forward	gp150	AUR85196.1, dual specificity phosphatase catalytic domain [Vibrio phage 1.070.O._10N.261.45.B2], 70.9%, 31.86%, 2.22e-13	Dual specificity phosphatase catalytic domain
ORF151	97074	97409	336	forward	gp151		
ORF152	97474	97596	123	forward	gp152	VVB52063, Uncharacterised protein [uncultured archaeon], 57.56 %, 62.5 %, 1.22e-9	

Table S2. EV2 ORFs and their putative functions.

EV2							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF1	1	2172	2172	forward	gp1	WP_195564997, hypothetical protein [Parabacteroides merdae], 31.49 %, 30.3 %, 1.97e-15	Terminase large subunit
ORF2	2272	2727	456	forward	gp2		
ORF3	2146	2277	132	reverse	gp3		
ORF4	2731	4920	2190	forward	gp4	PWT73821, hypothetical protein C5B60_07670 [Chloroflexi bacterium], 87.95 %, 28.4 %, 6.24e-75	
ORF5	5009	5605	597	forward	gp5	MBN2293007, helix-turn-helix domain-containing protein [Pirellulales bacterium], 84.42 %, 34.6 %, 9.61e-26	HTH-type transcriptional regulator
ORF6	5821	5952	132	forward	gp6		
ORF7	5959	6210	252	forward	gp7		
ORF8	6292	7047	756	forward	gp8		
ORF9	7072	8265	1194	forward	gp9	PWT93081, hypothetical protein C5B54_02450 [Acidobacteria bacterium], 98.74 %, 28.7 %, 8.34e-33	Phage major capsid protein
ORF10	8322	8960	639	forward	gp10		
ORF11	8975	9301	327	forward	gp11		
ORF12	9463	10362	900	forward	gp12	MBL7983922, hypothetical protein [Flavobacteriales bacterium], 76.67 %, 30.4 %, 8.16e-13	
ORF13	10371	13943	3573	forward	gp13	WP_142988179, SGNH/GDSL hydrolase family protein [Granulicella rosea], 55 %, 25.3 %, 2.83e-5	SGNH/GDSL hydrolase family protein
ORF14	13940	15967	2028	forward	gp14		
ORF15	15931	18690	2760	forward	gp15	WP_179584031, hypothetical protein [Edaphobacter lichenicola], 86.09 %, 36.8 %, 9.85e-126	Pectate lyase-like protein
ORF16	18687	19028	342	forward	gp16		
ORF17	19025	19474	450	forward	gp17		
ORF18	19459	23028	3570	forward	gp18	MBB5316895, hypothetical protein [Edaphobacter lichenicola], 49.66 %, 63.2 %, 0	Heme exporter protein D
ORF19	23055	23189	135	forward	gp19		
ORF20	23471	23635	165	forward	gp20		
ORF21	23186	23326	141	reverse	gp21		
ORF22	23892	24203	312	forward	gp22		
ORF23	24200	24619	420	forward	gp23	EEK6741156, lysozyme [Salmonella enterica subsp. enterica serovar Enteritidis], 97.14 %, 56.7 %, 1.61e-42	Lysozyme
ORF24	24623	24979	357	forward	gp24		
ORF25	24979	25122	144	forward	gp25		
ORF26	25094	25273	180	reverse	gp26		
ORF27	25291	25539	249	forward	gp27		
ORF28	25539	25742	204	forward	gp28	WP_176125780, hypothetical protein [Paraburkholderia youngii], 64.71 %, 70.5 %, 6.39e-13	
ORF29	25768	25947	180	forward	gp29		
ORF30	25944	26099	156	forward	gp30		
ORF31	26096	26617	522	reverse	gp31	MBN1507285, HNH endonuclease [Sedimentisphaerales bacterium], 90.23 %, 41.8 %, 4.98e-23	Endonuclease
ORF32	26680	28077	1398	forward	gp32	MBO0758621, hypothetical protein [Bradyrhizobiaceae bacterium], 79.18 %, 26.2 %, 2.75e-24	
ORF33	28093	32016	3924	forward	gp33	PYX86513, hypothetical protein DMG70_00535, partial [Acidobacteria bacterium], 31.04 %, 33.3 %, 3.82e-45	
ORF34	32017	32313	297	forward	gp34		
ORF35	32462	32836	375	forward	gp35		
ORF36	32875	33216	342	forward	gp36		
ORF37	33217	33717	501	forward	gp37		
ORF38	33723	35930	2208	forward	gp38		
ORF39	35930	37339	1410	forward	gp39		
ORF40	37336	38187	852	forward	gp40		

EV2							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF41	38165	38371	207	forward	gp41		
ORF42	38354	38542	189	reverse	gp42		
ORF43	38652	39149	498	reverse	gp43		
ORF44	39173	39529	357	reverse	gp44		
ORF45	39529	39912	384	reverse	gp45		
ORF46	39968	40231	264	reverse	gp46		
ORF47	40242	40439	198	reverse	gp47		
ORF48	40494	40907	414	reverse	gp48	CAB4199560, Rus Holliday junction resolvase [uncultured Caudovirales phage], 95.65 %, 48.9 %, 2.57e-23	Rus Holliday junction resolvase
ORF49	40894	41178	285	reverse	gp49	DAY74410, TPA: MAG TPA: AcrIIC3 protein [Myoviridae sp.], 81.05 %, 41.6 %, 3.85e-07	AcrIIC3 protein
ORF50	41175	41372	198	reverse	gp50		
ORF51	41353	41670	318	reverse	gp51		
ORF52	41667	42755	1089	reverse	gp52	MBU6488275, DNA cytosine methyltransferase [Burkholderiales bacterium], 99.17 %, 56 %, 4.53e-111	DNA cytosine methyltransferase
ORF53	42752	43381	630	reverse	gp53		
ORF54	43431	43574	144	reverse	gp54		
ORF55	43571	43897	327	reverse	gp55		
ORF56	43929	44324	396	reverse	gp56	DAJ06814, TPA: MAG TPA: replisome organizer [Siphoviridae sp.], 90.15 %, 32.8 %, 1.55e-4	Replisome organizer
ORF57	44328	44771	444	reverse	gp57	YP_009986333, hypothetical protein JR324_gp213 [Escherichia phage nieznany], 97.3 %, 36.4 %, 9.99e-25	HNHc nuclease
ORF58	44936	45457	522	reverse	gp58	MBL8793140, hypothetical protein [Planctomycetia bacterium], 99.43 %, 34.6 %, 1.57e-27	
ORF59	45458	46354	897	reverse	gp59	NLZ00572, ATP-binding protein [Pirellulaceae bacterium], 95.65 %, 45.3 %, 1.14e-83	ATP-binding protein
ORF60	46365	46565	201	reverse	gp60		
ORF61	46555	47589	1035	reverse	gp61	ANS03326, hypothetical protein [uncultured Mediterranean phage uvDeep-CGR2-KM19-C37], 93.91 %, 49.7 %, 1.16e-94	PD-(D/E)XK nuclease family protein
ORF62	47586	47792	207	reverse	gp62		
ORF63	47785	47925	141	reverse	gp63		
ORF64	48066	48308	243	reverse	gp64		
ORF65	48501	48845	345	reverse	gp65		
ORF66	48848	48982	135	reverse	gp66		
ORF67	48983	49390	408	reverse	gp67		
ORF68	49377	50459	1083	reverse	gp68		
ORF69	50456	50806	351	reverse	gp69		
ORF70	50784	51350	567	reverse	gp70		
ORF71	51337	51534	198	reverse	gp71		
ORF72	51531	51935	405	reverse	gp72	WP_159332858, hypothetical protein [Sphingobacterium sp. 8BC], 37.78 %, 49 %, 6.59e-04	
ORF73	52011	52178	168	reverse	gp73		
ORF74	52212	52358	147	reverse	gp74		
ORF75	52421	52798	378	forward	gp75		
ORF76	53086	53205	120	reverse	gp76		
ORF77	52892	53089	198	reverse	gp77		
ORF78	53250	53369	120	reverse	gp78		
ORF79	53389	53538	150	forward	gp79		

EV2							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF80	53744	54946	1203	reverse	gp80	WP_183980275, hypothetical protein [Edaphobacter lichenicola], 70.07 %, 54.9 %, 5.76e-84	
ORF81	54936	55121	186	reverse	gp81		
ORF82	55199	55336	138	reverse	gp82		
ORF83	55300	55479	180	reverse	gp83		
ORF84	55476	55646	171	reverse	gp84		
ORF85	55691	55912	222	reverse	gp85		
ORF86	55983	56675	693	reverse	gp86	RUP38893, hypothetical protein EKK63_10995 [Acinetobacter sp.], 97.84 %, 45.1 %, 9.85e-54	
ORF87	56762	57094	333	reverse	gp87	WP_214829273, hypothetical protein [Chryseobacterium sp. ISL-6], 76.58 %, 43.2 %, 4.12e-12	
ORF88	57122	57319	198	reverse	gp88	MAH50751, hypothetical protein [Candidatus Pacearchaeota archaeon], 96.97 %, 71.9 %, 1.55e-26	
ORF89	57316	57507	192	reverse	gp89	QXV73543, hypothetical protein [Rhizobium phage RHph_X2_30], 93.75 %, 46.9 %, 1.07e-08	
ORF90	57504	57785	282	reverse	gp90		
ORF91	57782	57895	114	reverse	gp91		
ORF92	57886	58089	204	reverse	gp92	EGE5776580, hypothetical protein [Escherichia coli], 80.88 %, 69.1 %, 1.68e-20	Putative HNH endonuclease
ORF93	58123	58323	201	reverse	gp93		
ORF94	58354	58545	192	reverse	gp94		
ORF95	59040	59264	225	forward	gp95		
ORF96	59252	59422	171	forward	gp96		
ORF97	59419	59709	291	forward	gp97		
ORF98	59706	59921	216	forward	gp98		
ORF99	59906	60217	312	forward	gp99		
ORF100	60201	60335	135	forward	gp100		
ORF101	60359	60685	327	forward	gp101		
ORF102	60685	60891	207	forward	gp102		
ORF103	60896	61234	339	forward	gp103	WP_153529011, hypotentical protein [Sinorizobium meliloti], 71.68 %, 39 %, 2.41e-5	
ORF104	61234	61449	216	forward	gp104	MBT4124567, hypothetical protein [Candidatus Pacebacteria bacterium], 93.06 %, 64.2 %, 9.74e-22	
ORF105	61449	61772	324	forward	gp105		
ORF106	61769	62239	471	forward	gp106	WP_074830503, 3'-5' exoribonuclease [Bradyrhizobium lablabi], 99.36 %, 56.7 %, 4.71e-56	3'-5' exoribonuclease
ORF107	62262	62456	195	forward	gp107		
ORF108	62446	63027	582	forward	gp108		
ORF109	63027	1	144	forward	gp109		

Table S3. EV3 ORFs and their putative functions.

EV3							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 28.8.2021)**	**Putative function:**
ORF1	1	2172	2172	forward	gp1	PWT73825, hypothetical protein C5B60_07690 [Chloroflexi bacterium], 84.94 %, 31.2 %, 2.34e-82	Terminase large subunit
ORF2	2266	2727	462	forward	gp2		
ORF3	2146	2277	132	reverse	gp3		
ORF4	2731	4920	2190	forward	gp4	PWT73821, hypothetical protein C5B60_07670 [Chloroflexi bacterium], 87.95 %, 28.4 %, 1.63e-74	
ORF5	4889	5005	117	forward	gp5		
ORF6	5252	5383	132	forward	gp6		
ORF7	5390	5641	252	forward	gp7		
ORF8	5723	6478	756	forward	gp8		
ORF9	6503	7696	1194	forward	gp9	PWT93081, hypothetical protein C5B54_02450 [Acidobacteria bacterium], 98.74 %, 28.7 %, 8.48e-33	Phage major capsid protein
ORF10	7753	8391	639	forward	gp10		
ORF11	8406	8732	327	forward	gp11		
ORF12	8801	9793	993	forward	gp12	MBL7983922, hypothetical protein [Flavobacteriales bacterium], 69.49 %, 30.4 %, 1.58e-12	
ORF13	9802	13374	3573	forward	gp13	WP_142988179, SGNH/GDSL hydrolase family protein [Granulicella rosea], 55 %, 25.3 %, 2.87e-5	SGNH/GDSL hydrolase family protein
ORF14	13371	15398	2028	forward	gp14		
ORF15	15380	18118	2739	forward	gp15	WP_183980467, hypothetical protein [Edaphobacter lichenicola], 86.64 %, 36.8 %, 1.66e-128	Glycoside hydrolase family 55 protein
ORF16	18983	19111	129	reverse	gp16		
ORF17	18115	18456	342	forward	gp17		
ORF18	18453	18902	450	forward	gp18	WP_183978929, right-handed parallel beta-helix repeat-containing protein [Edaphobacter lichenicola], 100 %, 81.4 %, 4.59e-13	
ORF19	19160	22744	3585	forward	gp19	WP_183978929, right-handed parallel beta-helix repeat-containing protein [Edaphobacter lichenicola], 25.52 %, 86.6 % %, 7.5e-131	heme exporter protein D
ORF20	22771	22905	135	forward	gp20		
ORF21	22902	23486	585	reverse	gp21		
ORF22	23692	23919	228	forward	gp22		
ORF23	23916	24335	420	forward	gp23	EEK6741156, lysozyme [Salmonella enterica subsp. enterica serovar Enteritidis], 97.14 %, 56.7 %, 1.61e-42	Lysozyme
ORF24	24339	24695	357	forward	gp24		
ORF25	24695	24838	144	forward	gp25		
ORF26	24840	25010	171	forward	gp26		
ORF27	25007	25255	249	forward	gp27		
ORF28	25255	25458	204	forward	gp28	WP_020949629, hypothetical protein [Paracoccus aminophilus], 60.29 %, 73.2 %, 8.63e-12	
ORF29	25484	25663	180	forward	gp29		
ORF30	25660	25815	156	forward	gp30		
ORF31	25812	26333	522	reverse	gp31	MBN1507285, HNH endonuclease [Sedimentisphaerales bacterium], 90.23 %, 41.8 %, 4.62e-23	Endonuclease
ORF32	26396	27793	1398	forward	gp32	MBO0758621, hypothetical protein [Bradyrhizobiaceae bacterium], 79.18 %, 26.2 %, 2.80e-24	
ORF33	27809	31732	3924	forward	gp33	PYX86513, hypothetical protein DMG70_00535, partial [Acidobacteria bacterium], 31.04 %, 33.3 %, 3.83e-45	
ORF34	31733	32029	297	forward	gp34		
ORF35	32178	32552	375	forward	gp35		
ORF36	32573	32932	360	forward	gp36		
ORF37	32933	33433	501	forward	gp37		
ORF38	33469	35622	2154	forward	gp38		
ORF39	35622	37031	1410	forward	gp39		
ORF40	37028	37879	852	forward	gp40		

EV3							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 28.8.2021)**	**Putative function:**
ORF41	37857	38063	207	forward	gp41		
ORF42	38046	38282	237	reverse	gp42		
ORF43	38345	38842	498	reverse	gp43		
ORF44	38866	39222	357	reverse	gp44		
ORF45	39222	39605	384	reverse	gp45		
ORF46	39661	39924	264	reverse	gp46		
ORF47	39935	40132	198	reverse	gp47		
ORF48	40187	40600	414	reverse	gp48	CAB4199560, Rus Holliday junction resolvase [uncultured Caudovirales phage], 95.65 %, 48.9 %, 2.57e-23	Rus Holliday junction resolvase
ORF49	40587	40871	285	reverse	gp49	DAY74410, TPA: MAG TPA: AcrIIC3 protein [Myoviridae sp.], 81.05 %, 41.6 %, 9.32e-8	AcrIIC3 protein
ORF50	40868	41065	198	reverse	gp50		
ORF51	41046	41363	318	reverse	gp51		
ORF52	41360	41995	636	reverse	gp52		
ORF53	42185	42511	327	reverse	gp53		
ORF54	42543	42938	396	reverse	gp54	DAJ06814, TPA: MAG TPA: replisome organizer [Siphoviridae sp.], 84.09 %, 35.7 %, 6.44e-5	Replisome organizer
ORF55	43103	43624	522	reverse	gp55	MBL8793140, hypothetical protein [Planctomycetia bacterium], 99.43 %, 35.1 %, 1.23e-27	
ORF56	43625	44521	897	reverse	gp56	TXH49472, ATP-binding protein [Desulfurellales bacterium], 97.66 %, 43.9 %, 1.65e-69	ATP-binding protein
ORF57	44533	44733	201	reverse	gp57		
ORF58	44723	45760	1038	reverse	gp58	MBN95370, hypothetical protein [Deltaproteobacteria bacterium], 96.53 %, 36.7 %, 5.23e-55	PD-(D/E)XK nuclease family protein
ORF59	45753	45893	141	reverse	gp59		
ORF60	45912	46655	744	reverse	gp60	WP_212005773, hypothetical protein [Chitinophaga sp. KRA15-503], 98.79 %, 50.4 %, 2.51e-56	BsuMI modification methylase subunit ydiP
ORF61	46652	46894	243	reverse	gp61		
ORF62	47087	47431	345	reverse	gp62		
ORF63	47434	47568	135	reverse	gp63		
ORF64	47569	48078	510	reverse	gp64	PZR36560, hypothetical protein DI526_03145 [Caulobacter segnis], 40 %, 44.1 %, 1.99e-5	
ORF65	48075	48755	681	reverse	gp65		
ORF66	48752	48922	171	reverse	gp66		
ORF67	48919	49692	774	reverse	gp67		
ORF68	49670	50203	534	reverse	gp68		
ORF69	50193	50498	306	reverse	gp69	WP_135918881, MULTISPECIES: hypothetical protein [unclassified Mesorhizobium], 85.29 %, 41.6 %, 5.04e-10	
ORF70	50485	51177	693	reverse	gp70		
ORF71	51227	51430	204	reverse	gp71		
ORF72	51486	51653	168	reverse	gp72		
ORF73	51687	51821	135	reverse	gp73		
ORF74	51949	52242	294	reverse	gp74		
ORF75	52253	52507	255	reverse	gp75		
ORF76	52504	52650	147	reverse	gp76		
ORF77	52647	52844	198	reverse	gp77		

EV3							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 28.8.2021)**	**Putative function:**
ORF78	52841	52960	120	reverse	gp78		
ORF79	52977	53465	489	reverse	gp79	MBN1363140, HNH endonuclease [Sedimentisphaerales bacterium], 98.77 %, 34.8 %, 1.51e-16	HNH endonuclease
ORF80	53515	53634	120	reverse	gp80		
ORF81	53654	53803	150	forward	gp81		
ORF82	54009	55208	1200	reverse	gp82	WP_183980275, hypothetical protein [Edaphobacter lichenicola], 70.25 %, 54.9 %, 7.24e-84	
ORF83	55198	55383	186	reverse	gp83		
ORF84	55562	55741	180	reverse	gp84		
ORF85	55762	55896	135	forward	gp85		
ORF86	55953	56174	222	reverse	gp86		
ORF87	56245	56937	693	reverse	gp87	RUP38893, hypothetical protein EKK63_10995 [Acinetobacter sp.], 97.84 %, 45.1 %, 2.84e-58	
ORF88	57024	57221	198	reverse	gp88	VVC05176, Uncharacterised protein [uncultured archaeon], 98.48 %, 72.3 %, 7.84e-25	
ORF89	57218	57409	192	reverse	gp89	QXV73543, hypothetical protein [Rhizobium phage RHph_X2_30], 93.57 %, 45.3 %, 6.44e-8	
ORF90	57406	57687	282	reverse	gp90		
ORF91	57684	57797	114	reverse	gp91		
ORF92	58025	58225	201	reverse	gp92		
ORF93	58258	58395	138	reverse	gp93		
ORF94	58738	58851	114	forward	gp94	WP_135090488, DUF559 domain-containing protein [Sphingomonas parva], 92.11 %, 51.4 %, 2.72e-4	
ORF95	59306	59530	225	forward	gp95		
ORF96	59518	59688	171	forward	gp96		
ORF97	59685	60026	342	forward	gp97	WP_153529011, hypothetical protein [Sinorhizobium meliloti], 75.44 %, 36.8 %, 2.16e-5	
ORF98	60023	60241	219	forward	gp98		
ORF99	60226	60537	312	forward	gp99	WP_074830503, 3'-5' exoribonuclease [Bradyrhizobium lablabi], 99.36 %, 56.1 %, 1.05e-55	3'-5' exoribonuclease
ORF100	60679	61005	327	forward	gp100		
ORF101	60550	60684	135	reverse	gp101		
ORF102	61005	61211	207	forward	gp102		
ORF103	61216	61557	342	forward	gp103		
ORF104	61557	61880	324	forward	gp104		
ORF105	61877	62347	471	forward	gp105		
ORF106	62334	62564	231	forward	gp106		
ORF108	62554	63135	582	forward	gp108		
ORF109	63135	1	144	forward	gp109		

Table S4. EV5 ORFs and their putative functions.

EV5							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 28.8.2021)**	**Putative function:**
ORF1	1	1458	1458	forward	gp1	CAB4136849.1, XtmB Phage terminase large subunit [uncultured Caudovirales phage], -, 29.16%, 1.93e-54	Phage terminase large subunit
ORF2	1585	2583	999	forward	gp2	WP_185650222, glycosyltransferase [Clostridium sp. DJ247], 76.26 %, 22.6 %, 1.64e-7	Glycosyltransferase
ORF3	2592	4325	1734	forward	gp3	DAL25760, TPA_asm: MAG TPA_asm: portal protein [Siphoviridae sp.], 82.87 %, 25.3 %, 4.36e-26	Phage portal protein
ORF4	4405	5151	747	forward	gp4		
ORF5	5226	6206	981	forward	gp5	WP_188580350, hypothetical protein [Tistrella bauzanensis], 99.08 %, 52.6 %, 6.15e-108	Phage capsid protein
ORF6	6291	7169	879	forward	gp6	WP_154673779, Ig-like domain repeat protein [Singulisphaera acidiphila], 32.76 %, 44 %, 8.60e-7	Ig-like domain repeat protein
ORF7	7172	7594	423	forward	gp7		
ORF8	7591	8022	432	forward	gp8		
ORF9	8015	8740	726	forward	gp9		
ORF10	8856	9416	561	forward	gp10		
ORF11	9551	9697	147	forward	gp11		
ORF12	9697	10374	678	forward	gp12	DAN74926.1, TPA: MAG TPA: Prex DNA polymerase [Podoviridae sp.], 100%, 26.06%, 4.29e-19	Prex DNA polymerase
ORF13	10562	14785	4224	forward	gp13	HET76244, hypothetical protein [Acidobacteria bacterium], 20.03 %, 30.4 %, 3.21e-12	
ORF14	14742	15149	408	forward	gp14		
ORF15	15177	19136	3960	forward	gp15	HEU44481, hypothetical protein [Acidobacteria bacterium], 38.49 %, 32.6 %, 5.76e-60	
ORF16	19150	21303	2154	forward	gp16	DAT97964.1, TPA: MAG TPA: Baseplate upper protein immunoglobulin like domain [Siphoviridae sp.], 52.0%, 29.69%, 1.86e-04	Baseplate upper protein Ig-like domain ilke protein
ORF17	21306	21662	357	forward	gp17	WP_213805765, hypothetical protein [Granulicella sp. dw_53], 99.16 %, 35.6 %, 1.11e-17	
ORF18	21667	22023	357	forward	gp18		
ORF19	22025	22498	474	forward	gp19	WP_130417546, hypothetical protein [Edaphobacter modestus], 99.37 %, 44.4 %, 4.88e-28	
ORF20	22499	22789	291	forward	gp20	WP_213805768, hypothetical protein [Granulicella sp. dw_53], 92.78 %, 45.6 %, 9.46e-18	
ORF21	22786	23322	537	forward	gp21	RZU39317, hypothetical protein BDD14_0686 [Edaphobacter modestus], 80.45 %, 48.3 %, 4.74e-45	
ORF22	23437	23793	357	forward	gp22		
ORF23	23904	24968	1065	forward	gp23	RWZ86779, hypothetical protein EO766_13310 [Hydrotalea sp. AMD], 96.62 %, 36 %, 5.51e-71	Transposase
ORF24	25041	25295	255	forward	gp24		
ORF25	25373	26113	741	forward	gp25	MBS1722828, hypothetical protein [Armatimonadetes bacterium], 94.33 %, 28.8 %, 3.18e-11	Deoxynucleoside monophosphate kinase
ORF26	26171	26560	390	forward	gp26		
ORF27	26580	27224	645	forward	gp27		
ORF28	27239	27475	237	forward	gp28	WP_066765656, hypothetical protein [Sphingobium sp. CCH11-B1], 92.41 %, 38.4 %, 4.28e-8	
ORF29	27539	27946	408	forward	gp29	RJQ54006, lysozyme [Desulfobacteraceae bacterium], 99.26 %, 45.2 %, 4.52e-30	Lysozyme
ORF30	28002	28451	450	reverse	gp30		
ORF31	28802	29104	303	reverse	gp31		
ORF32	29101	29583	483	reverse	gp32		
ORF33	29609	29842	234	reverse	gp33		
ORF34	29839	30171	333	reverse	gp34		
ORF35	30286	31929	1644	reverse	gp35	MBS3934059, ribonucleoside-diphosphate reductase subunit alpha [Truepera sp.], 99.27 %, 55.2 %, 0	Ribonucleoside-diphosphate reductase subunit alpha
ORF36	31943	32917	975	reverse	gp36	NIQ16161, ribonucleotide reductase [Candidatus Dadabacteria bacterium], 99.69 %, 63.9 %, 3.97e-154	Ribonucleotide reductase
ORF37	32914	33108	195	reverse	gp37		
ORF38	33105	34298	1194	reverse	gp38	MBI2448315, recombinase RecA [Candidatus Microgenomates bacterium], 86.93 %, 52.6 %, 9.43e-97	Recombinase RecA
ORF39	34295	34735	441	reverse	gp39		
ORF40	34766	35251	486	reverse	gp40		
ORF41	35241	35633	393	reverse	gp41	QDP60839, putative protein D14 [Prokaryotic dsDNA virus sp.], 95.42 %, 38.5 %, 1.00e-15	Resolvase

EV5							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 28.8.2021)**	**Putative function:**
ORF42	35620	35991	372	reverse	gp42		
ORF43	35981	36247	267	reverse	gp43		
ORF44	36397	38349	1953	reverse	gp44	MBU1209011, AAA family ATPase [Proteobacteria bacterium], 96.77 %, 26.4 %, 5.16e-41	AAA family ATPase
ORF45	38494	38700	207	reverse	gp45		
ORF46	38663	39151	489	reverse	gp46		
ORF47	39152	40960	1809	reverse	gp47	MBI5478857, DEAD/DEAH box helicase [Deltaproteobacteria bacterium], 99 %, 31.6 %, 3.14e-86	DEAD/DEAH box helicase
ORF48	40999	41367	369	reverse	gp48	MBK8467938, hypothetical protein [Chloracidobacterium sp.], 73.17 %, 53.8 %, 1.04e-22	Yfdr
ORF49	41414	41956	543	reverse	gp49	QBQ74851.1, methyltransferase [Caudovirales GX15bay], 74.7%, 31.36%, 1.49e-14	Methyltransferase
ORF50	41966	42286	321	reverse	gp50	MBP7211077, HNH endonuclease [Paludibacteraceae bacterium], 84.11 %, 35.8 %, 2.58e-9	HNH endonuclease
ORF51	42289	43008	720	reverse	gp51		
ORF52	43042	43314	273	reverse	gp52		
ORF53	43327	43464	138	reverse	gp53		
ORF54	43550	43696	147	reverse	gp54		
ORF55	43964	44311	348	reverse	gp55	QSM04704.1, RF-1 peptide chain release factor [Mycobacterium phage prophiGD54-2], 72.8%, 42.11%, 2.78e-15	Peptide chain release factor-like protein
ORF56	44984	45424	441	reverse	gp56	DAJ36219.1, TPA: MAG TPA: (p)ppGpp synthetase, RelA/SpoT family [Siphoviridae sp.], -, 43.07%, 3.80e-30	(p)ppGpp synthetase, RelA/SpoT family
ORF57	45663	46268	606	reverse	gp57	MBV8629384, hypothetical protein [Silvibacterium sp.], 61.39 %, 27.8 %, 4.07e-4	
ORF58	46354	46611	258	reverse	gp58		
ORF59	46684	47199	516	reverse	gp59		
ORF60	47210	47344	135	reverse	gp60		
ORF61	47347	47541	195	reverse	gp61		
ORF62	47616	48050	435	reverse	gp62		
ORF63	48052	48921	870	reverse	gp63	NUM33439, hypothetical protein [Candidatus Brocadiae bacterium], 73.1 %, 25.8 %, 3.06e-14	
ORF64	48992	49114	123	reverse	gp64		
ORF65	49251	49487	237	reverse	gp65		
ORF66	49600	49815	216	forward	gp66	WP_201137524, superinfection immunity protein [Pseudomonas sp. TH49], 77.78 %, 57.6 %, 7.31e-12	Superinfection immunity protein
ORF67	49817	51115	1299	reverse	gp67	PWT76391, hypothetical protein C5B59_06610 [Bacteroidetes bacterium], 62.82 %, 41.5 %, 1.54e-45	Putative exonuclease
ORF68	51165	51461	297	reverse	gp68		
ORF69	52052	53050	999	reverse	gp69	WP_158945942, acyltransferase [Granulicella sp. S190], 99.7 %, 34.1 %, 7.72e-35	Acyltransferase
ORF70	53195	53392	198	reverse	gp70		
ORF71	53389	53769	381	reverse	gp71	WP_127528935, hypothetical protein [Sinorhizobium meliloti], 98.43 %, 57.6 %, 8.05e-43	
ORF72	53832	54299	468	reverse	gp72	QOC57956, homing endonuclease [Pseudomonas phage phiK7B1], 99.36 %, 41.7 %, 1.23e-36	Homing endonuclease
ORF73	54296	54454	159	reverse	gp73		
ORF74	54462	54785	324	reverse	gp74	WP_187616641, superinfection immunity protein [Paraburkholderia sp. UCT2], 61.11 %, 33.3 %, 5.63e-4	Superinfection immunity protein
ORF75	54786	55028	243	reverse	gp75		
ORF76	55021	55389	369	reverse	gp76		
ORF77	55389	55718	330	reverse	gp77		
ORF78	55734	55934	201	reverse	gp78		
ORF79	55938	56417	480	reverse	gp79	WP_018667218, hypothetical protein [Bacteroides gallinarum], 44.38 %, 40.3 %, 1.00e-4	
ORF80	56500	58320	1821	reverse	gp80	WP_198070527, MULTISPECIES: DUF4942 domain-containing protein [Bacteria], 92.59 %, 32.8 %, 9.25e-74	
ORF81	58428	59513	1086	reverse	gp81	TDI07642, DNA polymerase III subunit beta [Acidobacteria bacterium], 86.46 %, 25.4 %, 2.70e-12	DNA polymerase III subunit beta
ORF82	59573	59905	333	reverse	gp82		
ORF83	60014	60181	168	reverse	gp83	CAB5221338, hypothetical protein UFOVP240_138 [uncultured Caudovirales phage], 94.64 %, 47.4 %, 3.73e-9	
ORF84	60187	60561	375	reverse	gp84		
ORF85	60565	60897	333	reverse	gp85		
ORF86	60927	61406	480	reverse	gp86	WP_213805773, hypothetical protein [Granulicella sp. dw_53], 88.75 %, 45.8 %, 5.09e-31	
ORF87	61450	61644	195	reverse	gp87	NUQ27246, hypothetical protein [Acidobacteriaceae bacterium], 98.46 %, 48.4 %, 8.95e-12	

EV5							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 28.8.2021)**	**Putative function:**
ORF88	61670	62140	471	reverse	gp88		
ORF89	62245	63093	849	reverse	gp89	CAB4171846, hypothetical protein UFOVP923_31 [uncultured Caudovirales phage], 96.11 %, 31.3 %, 6.37e-26	
ORF90	63193	64470	1278	reverse	gp90	CAB4151713, ATPase-like protein [uncultured Caudovirales phage], 69.48 %, 37.7 %, 1.02e-41	ATPase-like protein
ORF91	64570	64704	135	reverse	gp91		
ORF92	65135	65290	156	reverse	gp92		
ORF93	66843	67154	312	reverse	gp93		
ORF94	67220	67444	225	reverse	gp94		
ORF95	67469	67738	270	reverse	gp95	WP_179586642, hypothetical protein [Edaphobacter lichenicola], 94.44 %, 56.5 %, 2.84e-25	
ORF96	67704	68693	990	forward	gp96	WP_183977848, hypothetical protein [Edaphobacter lichenicola], 9.09 %, 76.7 %, 7.98e-4	
ORF97	68831	69295	465	reverse	gp97		
ORF98	69309	69755	447	reverse	gp98		
ORF99	69755	70570	816	reverse	gp99		
ORF100	70564	70920	357	reverse	gp100	WP_196824494, bifunctional RNase H/acid phosphatase [Corynebacterium aquatimens], 68.91 %, 38.1 %, 5.06e-7	Bifunctional RNase H/acid phosphatase
ORF101	71317	71712	396	forward	gp101		
ORF102	72087	73235	1149	forward	gp102	MBN9616244, tyrosine-type recombinase/integrase [Acidobacteriales bacterium], 98.43 %, 68.1 %, 0	Tyrosine-type recombinase/integrase
ORF103	73285	73500	216	reverse	gp103	DAO79860.1, TPA: MAG TPA: Integrase [Siphoviridae sp.], 69.7%, 31.1%, 2.01e-10	Integrase
ORF104	73497	73775	279	reverse	gp104		
ORF105	73735	74196	462	reverse	gp105	WP_105485937, dUTP diphosphatase [Candidatus Sulfotelmatomonas gaucii], 97.4 %, 56 %, 1.20e-42	dUTP diphosphatase
ORF106	74168	74770	603	reverse	gp106	WP_090336593, phosphohydrolase [Pseudomonas chengduensis], 92.4 %, 45.2 %, 3.11e-46	Phosphohydrolase
ORF107	74758	77280	2523	reverse	gp107	PWT75565, hypothetical protein C5B59_08785 [Bacteroidetes bacterium], 98.34 %, 35.6 %, 1.86e-142	DNA polymerase I
ORF108	79397	79621	225	forward	gp108		
ORF109	79625	79786	162	forward	gp109		
ORF110	79791	80057	267	forward	gp110	WP_114208534, hypothetical protein [Acidisarcina polymorpha], 97.75 %, 37.9 %, 1.29e-5	
ORF111	80094	83189	3096	forward	gp111		
ORF112	83218	83658	441	forward	gp112		
ORF113	83669	85036	1368	forward	gp113	CAB4196581.1, hypothetical protein UFOVP1290_101 [uncultured Caudovirales phage], -, 49.44%, 7.21e-54	Putative hydrolase
ORF114	85036	87342	2307	forward	gp114	PYP93179, hypothetical protein DMG65_01525 [Acidobacteriia bacterium AA117], 86.48 %, 30.1 %, 6.20e-48	Ig-like domain repeat protein
ORF115	87412	29	660	forward	gp115		

Table S5. GV1 ORFs and their putative functions.

GV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF1	1	1914	1914	forward	gp1	RKY78148, hypothetical protein DRQ07_07900 [candidate division KSB1 bacterium], 83.54 %, 26.9 %, 7.40e-44	Terminase
ORF2	2002	2442	441	forward	gp2		
ORF3	3240	4220	981	forward	gp3		
ORF4	4254	4928	675	forward	gp4		
ORF5	4925	9196	4272	forward	gp5	MBR6762773, DNA polymerase I [Clostridia bacterium], 15.52 %, 30.2 %, 5.12e-15	DNA polymerase I
ORF6	9189	9899	711	forward	gp6		
ORF7	9889	12336	2448	forward	gp7	OHD24543, hypothetical protein A2Y38_08785 [Spirochaetes bacterium GWB1_59_5], 43.87 %, 38 %, 3.49e.68	Neck protein, gp14
ORF8	12373	15162	2790	forward	gp8		
ORF9	15227	16255	1029	forward	gp9		
ORF10	16279	18090	1812	forward	gp10	WP_198070527, MULTISPECIES: DUF4942 domain-containing protein [Bacteria], 91.06 %, 30.1 %, 1.36e-65	Type I restriction enzyme
ORF11	18200	18667	468	forward	gp11		
ORF12	18667	18858	192	forward	gp12		
ORF13	18858	18995	138	forward	gp13		
ORF14	19071	19736	666	forward	gp14	NCB43749, hypothetical protein [Clostridia bacterium], 86.04 %, 27.2 %, 3.60e-10	DNA repair protein RecN
ORF15	19740	20036	297	forward	gp15		
ORF16	20051	20263	213	forward	gp16	NBX19776, hypothetical protein [Bacteroidia bacterium], 71.83 %, 47.1 %, 2.20e-7	
ORF17	20432	21034	603	forward	gp17	PLX24680, hypothetical protein C0580_04530 [Candidatus Parcubacteria bacterium], 56.72 %, 38.6 %, 8.72e-15	(2Fe-2S)-binding protein
ORF18	21031	21222	192	forward	gp18		
ORF19	21219	21467	249	forward	gp19		
ORF20	21486	21890	405	forward	gp20		
ORF21	21890	22669	780	forward	gp21		
ORF22	22744	24141	1398	forward	gp22	NDJ12438, hypothetical protein [Acidobacteriia bacterium], 93.99 %, 44.2 %, 3.37e-113	
ORF23	24138	25259	1122	forward	gp23	MBC7217174, DNA polymerase III subunit gamma/tau [Candidatus Caldatribacterium sp.], 97.43 %, 33.1 %, 8.43e-36	DNA polymerase III subunit gamma/tau
ORF24	25310	26293	984	forward	gp24		
ORF25	26286	27323	1038	forward	gp25		
ORF26	27373	28008	636	forward	gp26		
ORF27	28055	28900	846	forward	gp27	DAO08328.1, TPA: MAG TPA: hypothetical protein [Siphoviridae sp.], 93.6%, 34.94%, 8.06e-20	Tail protein
ORF28	28905	33314	4410	forward	gp28	DAP54576.1, TPA: MAG TPA: hypothetical protein [Myoviridae sp.], 79%, 47.5%, 2.52e-14	Putative tail fiber protein
ORF29	33318	34175	858	forward	gp29	VVC05615, Uncharacterised protein [uncultured archaeon], 91.96 %, 28.6 %, 1.44e-06	
ORF30	34178	35023	846	forward	gp30	OFW05643, hypothetical protein A3H96_11320 [Acidobacteria bacterium RIFCSPLOWO2_02_FULL_67_36], 97.87 %, 32.3 %, 6.97e-26	
ORF31	35034	35879	846	forward	gp31	OFW05643, hypothetical protein A3H96_11320 [Acidobacteria bacterium RIFCSPLOWO2_02_FULL_67_36], 94.33 %, 34.3 %, 3.71e-30	
ORF32	35879	36736	858	forward	gp32	VVC05615, Uncharacterised protein [uncultured archaeon], 95.80 %, 28.7 %, 1.27e-19	
ORF33	36740	37057	318	forward	gp33		
ORF34	37023	38126	1104	reverse	gp34	WP_111339207, MULTISPECIES: glycosyltransferase [unclassified Streptomyces], 99.46 %, 42.7 %, 8.29e-82	Glycosyltransferase
ORF35	38140	39144	1005	reverse	gp35	QOI66578, hypothetical protein [Erwinia phage FBB1], 99.70 %, 27.1 %, 6.32e-26	Putative nucleotidyltransferase
ORF36	39163	39639	477	reverse	gp36		
ORF37	39639	39836	198	reverse	gp37		
ORF38	39872	40417	546	reverse	gp38		
ORF40	40607	41116	510	reverse	gp40		
ORF39	41113	41295	183	reverse	gp39		
ORF41	41295	42239	945	reverse	gp41	WP_138392054, hypothetical protein [Rhizobium sp. MHM7A], 91.75 %, 23.5 %, 1.08e-6	
ORF42	42293	42553	261	reverse	gp42	WP_174025717, hypothetical protein [Agrobacterium rubi], 91.95 %, 55 %, 4.36e-22	
ORF43	42573	43079	507	reverse	gp43		
ORF44	43091	43336	246	reverse	gp44	WP_211150460, hypothetical protein [Novosphingobium sp. HR1a], 97.56 %, 55 %, 2.19e-24	
ORF45	43333	43500	168	reverse	gp45	OHA92226, hypothetical protein A2723_01980 [Candidatus Zambryskibacteria bacterium], 91.07 %, 60.8 %, 2.97e-5	
ORF46	43509	43883	375	reverse	gp46		
ORF47	43939	44391	453	reverse	gp47	MBS1803810, dUTP diphosphatase [Acidobacteria bacterium], 88.08 %, 42.9 %, 3.94e-21	dUTP diphosphatase
ORF48	44490	45368	879	forward	gp48	MBT7192450, DEAD/DEAH box helicase family protein [archaeon], 92.15 %, 31.3 %, 1.88e-22	DEAD/DEAH box helicase family protein

GV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF49	45435	46397	963	forward	gp49	WP_106283700, dCMP deaminase family protein [Paraburkholderia sp. BL25I1N1], 49.22 %, 44.3 %, 1.74e-34	dCMP deaminase family protein
ORF50	46471	48027	1557	forward	gp50	WP_179493185, FAD-dependent thymidylate synthase [Granulicella arctica], 99.81 %, 45.5 %, 1.80e-135	FAD-dependent thymidylate synthase
ORF51	48087	48542	456	forward	gp51		
ORF52	48539	48988	450	forward	gp52		
ORF53	48997	49272	276	forward	gp53		
ORF54	49469	50038	570	forward	gp54	WP_089409443, PhoH family protein [Granulicella rosea], 97.89 %, 73.7 %, 4.01e-89	PhoH family protein
ORF55	50141	51133	993	forward	gp55		
ORF56	51133	51369	237	forward	gp56		
ORF57	51379	51729	351	forward	gp57	WP_121523675, hypothetical protein [Acinetobacter chengduensis], 48.72 %, 42.1 %, 2.66e-5	
ORF58	51791	52360	570	forward	gp58	MBO8139481, peptide deformylase [Thermosipho sp. (in: Bacteria)], 90.53 %, 36.8 %, 3.58e-26	Peptide deformylase
ORF59	52505	52996	492	forward	gp59		
ORF60	53129	53494	366	forward	gp60		
ORF61	53592	54734	1143	forward	gp61	WP_020772675, WGR domain-containing protein [Leptospira alstonii], 96.85 %, 32.4 %, 6.02e-42	
ORF62	54790	55584	795	forward	gp62	MBN9617597, phosphoadenosine phosphosulfate reductase family protein [Acidobacteriales bacterium], 94.72 %, 59 %, 3.06e-106	Phosphoadenosine phosphosulfate reductase family protein
ORF63	55585	55752	168	forward	gp63		
ORF64	55749	56120	372	forward	gp64	QIW90706.1, deoxynucleotide monophosphate kinase [Vibrio phage V07], 71.2%, 29.6%, 1.18e-12	Deoxynucleotide monophosphate kinase
ORF65	56117	56434	318	forward	gp65		
ORF66	56437	56658	222	forward	gp66		
ORF67	56655	56888	234	forward	gp67		
ORF68	56956	57750	795	forward	gp68	NDB85368, hypothetical protein [Alphaproteobacteria bacterium], 45.66 %, 30.1 %, 7.05e-7	
ORF69	57990	59429	1440	reverse	gp69		
ORF70	59426	59704	279	reverse	gp70		
ORF71	59688	60161	474	reverse	gp71		
ORF72	60210	60416	207	reverse	gp72	YP_009809412.1, cell division protein [Caulobacter phage CcrPW], -, 39.71%, 1.09e-38	Cell division protein
ORF73	60534	60875	342	reverse	gp73		
ORF74	60875	61075	201	reverse	gp74		
ORF75	61065	61292	228	reverse	gp75		
ORF76	61292	62053	762	reverse	gp76	DAT66356.1, TPA: MAG TPA: Chromatin remodeling complex ATPase [Caudovirales sp.], -, 28.65%, 4.79e-36	Chromatin remodeling complex ATPase
ORF77	62043	62696	654	reverse	gp77	WP_216845385, metallophosphoesterase [Granulicella sp. S156], 97.71 %, 58.7 %, 2.15e-84	Metallophosphoesterase
ORF78	62802	63416	615	reverse	gp78		
ORF79	63433	64077	645	reverse	gp79	VVB50871, Uncharacterised protein [uncultured archaeon], 94.42 %, 31.7 %, 1.18e-25	
ORF80	64077	64337	261	reverse	gp80		
ORF81	64394	65917	1524	reverse	gp81		
ORF82	66051	66404	354	forward	gp82	WP_167389848, hypothetical protein [Paraburkholderia acidophila], 41.53 %, 51 %, 4.06e-4	
ORF83	66500	66694	195	forward	gp83		
ORF84	66748	67182	435	forward	gp84		
ORF85	67212	67871	660	forward	gp85		
ORF86	67883	68179	297	forward	gp86	TAL55271, hypothetical protein EPN80_07945 [Pandoraea sp.], 95.96 %, 53.7 %, 2.40e-27	
ORF87	68289	68762	474	reverse	gp87		
ORF88	68759	69988	1230	reverse	gp88	RYN68654, Mitochondrial chaperone [Alternaria tenuissima], 99.02 %, 28.2 %, 3.22e-42	Mitochondrial chaperone
ORF89	70005	71540	1536	reverse	gp89		
ORF90	71537	71899	363	reverse	gp90		
ORF91	71896	72141	246	reverse	gp91		
ORF92	72253	74256	2004	reverse	gp92	NBR01213, DEAD/DEAH box helicase [Actinobacteria bacterium], 86.53 %, 32.2 %, 8.72e-67	DEAD/DEAH box helicase
ORF93	74303	74584	282	reverse	gp93		
ORF94	74619	75110	492	reverse	gp94		
ORF95	75107	75703	597	reverse	gp95		

GV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF96	75742	76419	678	reverse	gp96	MBT3298500, DEAD/DEAH box helicase family protein [archaeon], 73.89 %, 35.5 %, 1.92e-23	DEAD/DEAH box helicase family protein
ORF97	76515	77597	1083	reverse	gp97	MBT7192450, DEAD/DEAH box helicase family protein [archaeon], 75.9 %, 35.6 %, 5.08e-37	DEAD/DEAH box helicase family protein
ORF98	77600	79525	1926	reverse	gp98	WP_058465072, DEAD/DEAH box helicase family protein [Legionella cincinnatiensis], 91.9 %, 23.9 %, 2.63e-25	DEAD/DEAH box helicase family protein
ORF99	79707	79949	243	reverse	gp99		
ORF100	79952	80575	624	reverse	gp100		
ORF101	80627	80929	303	reverse	gp101	WP_047350944, hypothetical protein [Diaphorobacter sp. J5-51], 74.26 %, 41.3 %, 3.36e-9	
ORF102	81014	81598	585	reverse	gp102	WP_125486822, hypothetical protein [Edaphobacter aggregans], 97.95 %, 32 %, 4.44e-19	
ORF103	81730	82320	591	reverse	gp103		
ORF104	82313	82888	576	reverse	gp104	WP_037459016, hypothetical protein [Skermanella stibiiresistens], 89.58 %, 34.3 %, 1.15e-21	
ORF105	82885	83460	576	reverse	gp105		
ORF106	83460	84413	954	reverse	gp106	HEL91508, NAD-dependent DNA ligase LigA [Ignavibacteria bacterium], 94.34 %, 30.6 %, 1.14e-25	DNA ligase
ORF107	84416	84676	261	reverse	gp107		
ORF108	84759	85370	612	reverse	gp108	QIG70776, P-loop NTPase domain-containing protein [Rhizobium phage RHph_I1_18], 81.86 %, 40.6 %, 1.78e-32	AAA family ATPase
ORF109	85857	86258	402	reverse	gp109		
ORF110	86255	86824	570	reverse	gp110		
ORF111	86915	87145	231	reverse	gp111		
ORF112	87126	87374	249	reverse	gp112		
ORF113	87621	88763	1143	reverse	gp113	WP_134101819, NAD-dependent DNA ligase LigA [Kribbella sp. VKM Ac-2573], 99.21 %, 34.1 %, 7.53e-58	NAD-dependent DNA ligase LigA
ORF114	89220	89609	390	reverse	gp114	MBA2497452, J domain-containing protein [Acidimicrobiia bacterium], 73.85 %, 30.9 %, 5.13e-5	Molecular chaperone DnaJ
ORF115	89606	89995	390	reverse	gp115		
ORF116	90028	90648	621	reverse	gp116		
ORF117	90648	90815	168	reverse	gp117		
ORF118	90815	91279	465	reverse	gp118	MBW2560194, hypothetical protein [Deltaproteobacteria bacterium], 68.39 %, 32.1 %, 8.13e-4	
ORF119	91338	91820	483	reverse	gp119		
ORF120	91985	92182	198	reverse	gp120		
ORF121	92192	93121	930	reverse	gp121	DAF33691.1, TPA: MAG TPA: activating signal cointegrator [Siphoviridae sp.], 72%, 44.16%, 9.60e-13	Activating signal cointegrator
ORF122	93208	93648	441	reverse	gp122	NP_037688.1, hypothetical protein HK022p35 [Escherichia virus HK022], 71.2%, 31.34%, 3.05e-13	Ead/Ea22-like protein
ORF123	93697	93966	270	reverse	gp123		
ORF124	93956	97258	3303	reverse	gp124	WP_207545753, hypothetical protein [Achromobacter insolitus], 98.46 %, 37.3 %, 0	
ORF125	97360	98091	732	reverse	gp125		
ORF126	98127	98327	201	reverse	gp126		
ORF127	98327	99025	699	reverse	gp127	WP_142185981, hypothetical protein [Rhizobium cellulosilyticum], 98.28 %, 34.5 %, 1.68e-33	
ORF128	99029	99436	408	reverse	gp128	APU88927, hypothetical protein Rctr197k_121 [Virus Rctr197k], 63.97 %, 41.4 %, 3.37e-4	
ORF129	99439	99747	309	reverse	gp129		
ORF130	99744	100844	1101	reverse	gp130	WP_216327480, RNA ligase (ATP) [Deinococcus sp. SYSU M49105], 99.73 %, 42.3 %, 8.61e-75 %	RNA ligase
ORF131	100901	101155	255	reverse	gp131		
ORF132	101149	101925	777	reverse	gp132		
ORF133	101918	102196	279	reverse	gp133		
ORF134	102189	102488	300	reverse	gp134		
ORF135	102557	103123	567	reverse	gp135		
ORF136	103270	103611	342	reverse	gp136		
ORF137	103608	104402	795	reverse	gp137	WP_198152145, hypothetical protein [Granulicella tundricola], 92.45 %, 27 %, 3.71e-8	
ORF138	104456	104647	192	forward	gp138	WP_183793025, hypothetical protein [Edaphobacter lichenicola], 84.38 %, 72.2 %, 2.01e-19	
ORF139	104763	105257	495	forward	gp139	NTV10739, YbjN domain-containing protein [Zoogloea sp.], 93.94 %, 24.1 %, 1.14e-7	
ORF140	105263	105379	117	forward	gp140		
ORF141	105521	105715	195	reverse	gp141		
ORF142	105790	106254	465	reverse	gp142		

GV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF143	106290	106493	204	reverse	gp143		
ORF144	106495	106629	135	reverse	gp144	MBV8771096, KH domain-containing protein [Deltaproteobacteria bacterium], 84.44 %, 50 %, 5.43e-4	
ORF145	106816	107115	300	reverse	gp145	WP_184224078, hypothetical protein [Granulicella aggregans], 78 %, 38.5 %, 4.67e-10	
ORF146	107108	107623	516	reverse	gp146		
ORF147	107732	108145	414	reverse	gp147	WP_183974943, hypothetical protein [Edaphobacter lichenicola], 60.14 %, 41.8 %, 5.50e-12	
ORF148	108147	108362	216	reverse	gp148	WP_184217366, hypothetical protein [Granulicella aggregans], 97.22 %, 62.9 %, 7.08e-23	
ORF149	108369	108638	270	reverse	gp149	MBB5328540, hypothetical protein [Edaphobacter lichenicola], 93.33 %, 35.3 %, 9.60e-12	
ORF150	109170	109640	471	reverse	gp150		
ORF151	109927	110292	366	reverse	gp151		
ORF152	110888	111913	1026	reverse	gp152	WP_165420403, hypothetical protein [Edaphobacter modestus], 96.49 %, 41.1 %, 5.02e-83	RepA
ORF153	112123	112716	594	reverse	gp153		
ORF154	112898	113446	549	forward	gp154	WP_179587261, hypothetical protein [Edaphobacter lichenicola], 99.45 %, 83.5 %, 6.82e-92	
ORF155	113541	114206	666	forward	gp155	WP_179587262, ParA family protein [Edaphobacter lichenicola], 99.55 %, 86 %, 6.50e-132	ParA family protein
ORF156	114344	114586	243	forward	gp156	WP_179587264, hypothetical protein [Edaphobacter lichenicola], 98.77 %, 69.1 %, 2.19e-29	
ORF157	114777	116639	1863	reverse	gp157		
ORF158	116950	117117	168	reverse	gp158		
ORF159	117114	117689	576	reverse	gp159		
ORF160	117686	118090	405	reverse	gp160		
ORF161	118108	118416	309	reverse	gp161	WP_102856293, hypothetical protein [Phaeobacter inhibens], 99.03 %, 39.8 %, 1.71e-15	
ORF162	118391	119368	978	reverse	gp162		
ORF163	119387	119611	225	reverse	gp163		
ORF164	119728	120090	363	reverse	gp164		
ORF165	120087	120737	651	reverse	gp165	WP_211091050, DUF1643 domain-containing protein [Sphingomonas sp. S2M10], 96.77 %, 36.6 %, 8.36e-30	
ORF166	120730	121089	360	reverse	gp166	WP_095666301, hypothetical protein [Vibrio coralliilyticus], 97.5 %, 27.4 %, 2.71e-14	Chaperonin GroEl, gp228
ORF167	121130	122128	999	reverse	gp167	DAE83794, TPA: MAG TPA: putative Fe-S-cluster redox enzyme [Bacteriophage sp.], 95.20 %, 34.8 %, 3.08e-58	Fe-S-cluster redox enzyme
ORF168	122177	122437	261	reverse	gp168		
ORF169	122430	122723	294	reverse	gp169		
ORF170	122726	122839	114	reverse	gp170		
ORF171	122958	123716	759	reverse	gp171		
ORF172	123730	123963	234	reverse	gp172	WP_090634797, hypothetical protein [Nitrosomonas marina], 65.38 %, 59.6 %, 8.69e-8	
ORF173	123966	124397	432	reverse	gp173		
ORF174	124406	125410	1005	reverse	gp174	NBR68336, hypothetical protein [Actinobacteria bacterium], 88.66 %, 32.8 %, 1.18e-27	Metallophosphoesterase
ORF175	125415	125816	402	reverse	gp175		
ORF176	125878	126516	639	reverse	gp176	MBS1811881, PEP-CTERM sorting domain-containing protein [Acidobacteria bacterium], 99.53 %, 28.1 %, 4.60e-11	Putative secreted protein with PEP-CTERM sorting signal
ORF177	126646	126939	294	reverse	gp177		
ORF178	126949	128406	1458	reverse	gp178	VVB50766, Uncharacterised protein [uncultured archaeon], 57 %, 38.5 %, 3.00e-53	Chromosome segregation protein
ORF179	128417	128584	168	reverse	gp179		
ORF180	128623	128847	225	reverse	gp180		
ORF181	129027	130322	1296	reverse	gp181	VVB50762, Protein RecA [uncultured archaeon], 90.51 %, 40.9 %, 1.82e-90	Recomdinase RecA
ORF182	130336	131241	906	reverse	gp182	WP_130419071, hypothetical protein [Edaphobacter modestus], 38.41 %, 94 %, 1.07e-72	
ORF183	131342	132121	780	reverse	gp183	RZU39177, hypothetical protein BDD14_0524 [Edaphobacter modestus], 71.54 %, 74.9 %, 2.84e-99	
ORF184	132226	133005	780	reverse	gp184	WP_130417424, hypothetical protein [Edaphobacter modestus], 63.08 %, 73.3 %, 2.04e-80	
ORF185	133030	133365	336	reverse	gp185	WP_130417423, hypothetical protein [Edaphobacter modestus], 46.43 %, 65.4 %, 1.62e-14	
ORF186	133392	134483	1092	reverse	gp186		
ORF187	134501	135217	717	reverse	gp187		
ORF188	135234	135422	189	reverse	gp188		
ORF189	135533	137074	1542	reverse	gp189	WP_013580312, hypothetical protein [Granulicella tundricola], 89.3 %, 25.9 %, 4.6e-29	Replicative DNA helicase
ORF190	137078	138700	1623	reverse	gp190	HGF05425, hypothetical protein [bacterium], 98.89 %, 30.3 %, 3.41e-66	ATP-dependent DNA helicase

GV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF191	138745	139410	666	reverse	gp191		
ORF192	139407	139724	318	reverse	gp192		
ORF193	139776	139994	219	reverse	gp193		
ORF194	140030	140326	297	reverse	gp194		
ORF195	140435	140698	264	reverse	gp195		
ORF196	140711	140962	252	reverse	gp196		
ORF197	140974	141117	144	reverse	gp197		
ORF198	141131	141364	234	reverse	gp198	MAZ56790, hypothetical protein [bacterium], 94.87 %, 37.8 %, 6.57e-9	
ORF199	141519	141926	408	reverse	gp199		
ORF200	142090	142377	288	forward	gp200	WP_026441557, hypothetical protein [Acidobacterium ailaaui], 98.96 %, 46.3 %, 4.34e-17	
ORF201	142374	143408	1035	reverse	gp201	OGU54782, hypothetical protein A2V66_01610 [Ignavibacteria bacterium RBG_13_36_8], 24.64 %, 56.5 %, 4.32e-22	N-acetyltransferase
ORF202	143383	144243	861	reverse	gp202	WP_117297806, hypothetical protein [Acidipila sp. 4G-K13], 89.2 %, 26.1 %, 2.45e-5	
ORF203	144708	147269	2562	reverse	gp203	WP_176331407, hypothetical protein [Burkholderia vietnamiensis], 13.58 %, 69 %, 1.05e-41	N-6 DNA methylase
ORF204	147405	148355	951	forward	gp204		
ORF205	148389	150284	1896	reverse	gp205	QQS52182, UvrD-helicase domain-containing protein [Bacteroidetes bacterium], 99.37 %, 30.4 %,1.73e-65	DNA helicase
ORF206	150353	150757	405	reverse	gp206		
ORF207	150768	151469	702	reverse	gp207		
ORF208	151486	151827	342	reverse	gp208	MBI3565737, HEAT repeat domain-containing protein [Elusimicrobia bacterium], 97.37 %, 32.5 %, 1.90e-4	
ORF209	151918	152616	699	reverse	gp209	WP_214553905, RNA polymerase sigma factor RpoE [Enterobacter cloacae], 74.68 %, 33.90 %, 2.65e-17	RNA polymerase
ORF210	152618	153211	594	reverse	gp210	NDD55038, hypothetical protein [bacterium], 97.98 %, 37.3 %, 3.81e-27	N-glycosylase/DNA lyase
ORF211	153358	153681	324	reverse	gp211		
ORF212	153721	154245	525	reverse	gp212		
ORF213	154333	154608	276	reverse	gp213		
ORF214	154718	155101	384	reverse	gp214		
ORF215	155196	156596	1401	reverse	gp215	MBA3732966, nicotinate phosphoribosyltransferase [Patescibacteria group bacterium], 98.72 %, 46.2 %, 1.9e-142	Nicotinate phosphoribosyltransferase
ORF216	156714	157763	1050	reverse	gp216	WP_199654207, NUDIX domain-containing protein [Chitinophaga sp. OAE873], 86.57 %, 40.7 %, 1.90e-60	ADP-ribose pyrophosphatase
ORF217	158431	158778	348	reverse	gp217		
ORF218	158840	159796	957	reverse	gp218	QIG58829.1, hypothetical protein SEA_DATBOI_165 [Gordonia phage DatBoi], 48.5%, 29.63%, 6.10e-04	
ORF219	159808	160227	420	reverse	gp219		
ORF220	160283	160867	585	reverse	gp220	DAL94903.1, TPA: MAG TPA: baseplate protein [Myoviridae sp.], 47.8%, 34.57%, 1.81e-04	Baseplate protein
ORF221	160852	161307	456	reverse	gp221		
ORF222	161335	163401	2067	reverse	gp222	WP_196807281, PcfJ domain-containing protein [Solirubrobacter sp. URHD0082], 27.58 %, 25.6 %, 5.08e-4	GNAT family N-acetyltransferase, partial
ORF223	163409	167197	3789	reverse	gp223	DAF43683.1, TPA: MAG TPA: Nucleoside 2-deoxyribosyltransferase like protein [Myoviridae sp. ctNQV2], -, 48.30%, 1.36e-41	Nucleoside 2-deoxyribosyltransferase like protein
ORF224	167342	167599	258	forward	gp224		
ORF225	167596	168684	1089	forward	gp225	GGA63356, acyltransferase [Edaphobacter acidisoli], 94.21 %, 46.3 %, 7.36e-71	Acyltransferase
ORF226	168750	169001	252	forward	gp226	NOS67793, hypothetical protein [Candidatus Peribacteraceae bacterium], 84.52 %, 56.3 %, 2.46e-23	
ORF227	169068	169412	345	forward	gp227	WP_026441556, hypothetical protein [Acidobacterium ailaaui], 99.13 %, 46.5 %, 3.13e-26	
ORF228	169590	170072	483	forward	gp228	WP_130425268, hypothetical protein [Edaphobacter modestus], 88.20 %, 43.4 %, 4.46e-35	
ORF229	170135	170659	525	forward	gp229	WP_199032189, hypothetical protein [Ralstonia sp. ASV6], 48.57 %, 44.3 %, 7.36e-7	GNAT family N-acetyltransferase
ORF230	170659	170913	255	forward	gp230		
ORF231	170921	171298	378	forward	gp231		
ORF232	171349	171825	477	forward	gp232		
ORF233	171885	172127	243	reverse	gp333		
ORF234	172200	172685	486	forward	gp334	WP_214688291, MULTISPECIES: hypothetical protein [unclassified Exiguobacterium], 79.63 %, 29 %, 9.97e-6	
ORF235	172766	173320	555	forward	gp335	MBK6616643, hypothetical protein [Ottowia sp.], 43.24 %, 33.8 %, 5.21e-6	
ORF236	173593	173811	219	forward	gp336		
ORF237	173815	174849	1035	reverse	gp337		

GV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF238	174852	178571	3720	reverse	gp338	CAB4196506.1, Baseplate protein J-like [uncultured Caudovirales phage], 60.5%, 29.86%, 2.20e-06	Baseplate protein
ORF239	178578	179339	762	reverse	gp339		
ORF240	179404	180021	618	reverse	gp340		
ORF241	180030	181286	1257	reverse	gp341		
ORF242	181296	184616	3321	reverse	gp342	CAB4221170.1, Putative Ig [uncultured Caudovirales phage], 63.5%, 27.11%, 2.61e-07	Major tail protein
ORF243	184663	190572	5910	reverse	gp343	QFG05090.1, major tail protein [Gordonia phage Gibbous], 59.3%, 38.38%, 2.36e-06	Major tail protein
ORF244	190582	194322	3741	reverse	gp344	DAP71525.1, TPA: MAG TPA: protein of unknown function (DUF4815) [Siphoviridae sp.], -, 29.72%, 7.71e-20	Baseplate protein, gp211
ORF245	194391	194888	498	forward	gp345		
ORF246	194885	195646	762	reverse	gp346		
ORF247	195690	195917	228	reverse	gp347		
ORF248	195914	196291	378	reverse	gp348	DAQ75556.1, TPA: MAG TPA: lipoprotein [Myoviridae sp.], 95.1%, 47.83%, 5.51e-24	Lipoprotein
ORF249	196646	199657	3012	reverse	gp349		
ORF250	199806	202091	2286	reverse	gp350		
ORF251	202105	204756	2652	reverse	gp351		
ORF252	204753	205613	861	reverse	gp352		
ORF253	205614	206933	1320	reverse	gp353		
ORF254	206933	210253	3321	reverse	gp354		
ORF255	210257	210859	603	reverse	gp355		
ORF256	210868	211890	1023	reverse	gp356		
ORF257	211910	212389	480	reverse	gp357		
ORF258	212840	213661	822	forward	gp358		
ORF259	213691	214983	1293	forward	gp359	WP_109486015, ATP-binding protein [Occallatibacter savannae], 63.34 %, 22.1 %, 8.15e-4	ATP-binding protein
ORF260	215144	218233	3090	forward	gp360	RZD42887, ribonucleoside-diphosphate reductase [Euryarchaeota archaeon], 77.38 %, 60.8 %, 0	Ribonucleoside-diphosphate reductase
ORF261	218339	218470	132	forward	gp361	WP_038367787, hypothetical protein [Bosea sp. UNC402CLCol], 95.45 %, 57.1 %, 7.44e-7	
ORF262	218521	219033	513	forward	gp362	OGF34684, hypothetical protein A2482_00850 [Candidatus Falkowbacteria bacterium RIFOXYC2_FULL_48_21], 86.55 %, 58.1 %, 2.29e-54	Nucleoside 2-deoxyribosyltransferase
ORF263	219470	219838	369	forward	gp363	MBK7397504, hypothetical protein [Myxococcales bacterium], 91.87 %, 40.7 %, 1.75e-24	GxxExxY protein
ORF264	219885	221600	1716	forward	gp364	WP_011831632, helicase SNF2 [Methylibium petroleiphilum], 99.30 %, 52.3 %, 0	Helicase SNF2
ORF265	221597	222013	417	forward	gp365		
ORF266	222130	223764	1635	reverse	gp366	QDP54516.1, hypothetical protein Unbinned2514contig1000_8 [Prokaryotic dsDNA virus sp.], -, 42.11%, 5.11e-122	DNA polymerase II small subunit
ORF267	223875	224360	486	forward	gp367	PWT80209, hypothetical protein C5B44_05740 [Acidobacteria bacterium], 85.80 %, 32.2 %, 3.56e-9	
ORF268	224482	225729	1248	forward	gp368	CAB5220467, hypothetical protein UFOVP236_66 [uncultured Caudovirales phage], 58.17 %, 46.4 %, 4.37e-50	Phage protein Gp37/Gp68-like protein
ORF269	225787	230271	4485	forward	gp369	WP_155903073, hypothetical protein [Marinobacter gelidimuriae], 22.94 %, 30.6 %, 1.51e-24	Polymerase
ORF270	230255	230476	222	reverse	gp370		
ORF271	230580	230783	204	forward	gp371	WP_130419072, hypothetical protein [Edaphobacter modestus], 97.06 %, 78.8 %, 3.98e-27	
ORF272	230845	231102	258	forward	gp372		
ORF273	231768	231995	228	forward	gp373	MBV8136090, hypothetical protein [Deltaproteobacteria bacterium], 82.89 %, 77.8 %, 1.75e-29	
ORF274	232650	232763	114	forward	gp374		
ORF275	232764	233102	339	reverse	gp375		
ORF276	233135	234583	1449	reverse	gp376	MBV8113382, CCA tRNA nucleotidyltransferase [Silvibacterium sp.], 89.65 %, 44.1 %, 2.15e-116	CCA tRNA nucleotidyltransferase
ORF277	234567	235004	438	reverse	gp377		
ORF278	235246	235440	195	reverse	gp378		
ORF279	235461	235925	465	reverse	gp379	DAG81687.1, TPA: MAG TPA: hypothetical protein [Siphoviridae sp.], -, 55.32%, 7.90e-56	Endopeptidase
ORF280	235952	236398	447	reverse	gp380	MBK9497279, hypothetical protein [Xanthomonadales bacterium], 83.89 %, 45.9 %, 1.65e-145	
ORF281	236466	236801	336	reverse	gp381	WP_173929609, hypothetical protein [Pseudomonas syringae], 71.43 %, 30.9 %, 9.30e-4	
ORF282	236837	237166	330	reverse	gp382		
ORF283	237261	237809	549	reverse	gp383	WP_130425373, hypothetical protein [Edaphobacter modestus], 59.56 %, 27.7 %, 7.41e-4	
ORF284	237875	238129	255	reverse	gp384	NBT35917, hypothetical protein [Betaproteobacteria bacterium], 98.82 %, 44 %, 2.62e-15	
ORF285	238147	238446	300	reverse	gp385	MBF0388372, hypothetical protein [Candidatus Omnitrophica bacterium], 99 %, 56.6 %, 2.86e-29	Putative YahA protein
ORF286	238446	238631	186	reverse	gp386		
ORF287	238748	239134	387	forward	gp387	WP_123850854, hypothetical protein [Chryseobacterium shandongense], 86,05 %, 37.8 %, 7.89e-14	
ORF288	240148	240669	522	forward	gp388	MBV8731637, HNH endonuclease [Acidobacteriia bacterium], 99.43 %, 53.8 %, 2.08e-62	HNH endonuclease
ORF289	241095	241493	399	forward	gp389		
ORF290	242495	242806	312	forward	gp390		
ORF291	243432	243698	267	forward	gp391	YP_009876900, hypothetical protein HYP10_gp231 [Tenacibaculum phage pT24], 75.28 %, 50.7 %, 8.20e-14	
ORF292	243762	244943	1182	forward	gp392	MBV8136091, RtcB family protein [Deltaproteobacteria bacterium], 99.24 %, 78.3 %, 0	RtcB family protein
ORF293	245172	245951	780	forward	gp393	WP_207054928, ATP-grasp domain-containing protein [Corallococcus macrosporus], 99.62 %, 39.5 %, 7.32e-56	
ORF294	245942	246325	384	forward	gp394		
ORF295	246322	246462	141	forward	gp395		
ORF296	246471	246680	210	forward	gp396		
ORF297	246958	247725	768	forward	gp397	MBU0651168, polysaccharide pyruvyl transferase family protein [bacterium], 92.19 %, 35.7 %, 1.12e-38	Polysaccharide pyruvyl transferase family protein
ORF298	247764	248696	933	forward	gp398	WP_213613931, glycosyltransferase [Paenibacillus lactis], 85.21 %, 33 %, 8.33e-33	Glycosyltransferase
ORF299	248688	248807	120	reverse	gp399		
ORF300	248839	249570	732	forward	gp400	PYU88141, histidine phosphatase family protein [Acidobacteria bacterium], 93.03 %, 60.8 %, 5.58e-62	Histidine phosphatase family protein

GV1							
Name	**Start**	**End**	**Length**	**Direction**	**ORF product**	**Best blast hit: acc. No, organism, % query cover, % identity, e-value, (searche date: 10.8.2021)**	**Putative function:**
ORF301	249563	249904	342	forward	gp401		
ORF302	249901	250071	171	forward	gp402		
ORF303	250237	250383	147	forward	gp403		
ORF304	250944	251495	552	forward	gp404	DAY62547.1, TPA: MAG TPA: hypothetical protein [Myoviridae sp.], 62.8%, 29.55%, 8.12e-10	RdRp
ORF305	251573	251722	150	forward	gp405		
ORF306	253531	253740	210	forward	gp406		
ORF307	254083	254943	861	forward	gp407		
ORF308	255181	255366	186	forward	gp408		
ORF309	255363	255728	366	reverse	gp409	WP_208889125, hypothetical protein [Polaribacter sejongensis], 84.43 %, 33 %, 1.76e-8	HNH endonuclease
ORF310	255903	256082	180	reverse	gp410	WP_179586589, hypothetical protein [Edaphobacter lichenicola], 98.33 %, 45.8 %, 5.69e-8	
ORF311	256523	257020	498	forward	gp411	NDE17795, hypothetical protein [bacterium], 48.8 %, 59.8 %, 1.40e-18	
ORF312	257077	257415	339	forward	gp412		
ORF313	257412	257828	417	reverse	gp413		
ORF314	257825	258073	249	reverse	gp414		
ORF315	258070	258900	831	reverse	gp415	HCB04515, hypothetical protein [Nocardioides sp.], 98.18 %, 39.4 %, 3.39e-57	
ORF316	258911	260272	1362	reverse	gp416	WP_199028442, ankyrin repeat domain-containing protein [Ralstonia sp. ASV6], 95.59 %, 26.2 %, 8.99e-28	
ORF317	260369	261220	852	forward	gp417	NDB59445, hypothetical protein [bacterium], 98.94 %, 28.1 %, 1.39e-24	Tail sheet protein, gp165
ORF318	261291	261461	171	forward	gp418		
ORF319	261501	265589	4089	forward	gp419	MBK8168264, hypothetical protein [bacterium], 21.94 %, 34.1 %, 2.63e-29	DNA repair protein
ORF320	265599	266870	1272	forward	gp420	MBD3262625, hypothetical protein [Candidatus Altiarchaeales archaeon], 97.64 %, 29.4 %, 2.78e-22	
ORF321	266939	267418	480	forward	gp421		
ORF322	267432	269021	1590	forward	gp422		
ORF323	269118	269483	366	forward	gp423		
ORF324	269610	273224	3615	forward	gp424	YP_009015481.1, gp178 [Bacillus virus G], 77.8%, 24.62%, 6.78e-12	Capside vertex protein, gp178
ORF325	273309	273797	489	forward	gp425		
ORF326	273847	276330	2484	forward	gp426		
ORF327	276441	277136	696	forward	gp427	YP_009015483.1, gp180 [Bacillus virus G], 55.1%, 25.81%, 8.85e-07	Cardoxypeptidase D, gp180
ORF328	277218	277718	501	forward	gp428		
ORF329	277718	278581	864	forward	gp429	DAT66482.1, TPA: MAG TPA: hypothetical protein [Caudovirales sp.], -, 36.67%, 4.65e-32	Tail tube protein, gp19
ORF330	278679	280265	1587	forward	gp430		
ORF331	280535	281032	498	forward	gp431		
ORF332	281069	282028	960	forward	gp432	DAL00741.1, TPA: MAG TPA: major capsid protein [Myoviridae sp.], -, 35.96%, 2.77e-58	Major capsid protein
ORF333	282134	282925	792	forward	gp433		
ORF334	283022	284056	1035	forward	gp434	DAT66381.1, TPA: MAG TPA: hypothetical protein [Caudovirales sp.], 52.8%, 33.11%, 1.75e-05	
ORF335	284114	288259	4146	forward	gp435		
ORF336	288265	290073	1809	forward	gp436		
ORF337	290137	290499	363	forward	gp437		
ORF338	290513	290920	408	forward	gp438		
ORF339	290996	294877	3882	forward	gp439		
ORF340	294888	298448	3561	forward	gp440		
ORF341	298534	299334	801	forward	gp441		
ORF342	299402	299602	201	forward	gp442		
ORF343	299635	301638	2004	forward	gp443		
ORF344	301753	302487	735	forward	gp444		
ORF345	302637	302915	279	forward	gp445		
ORF346	302926	304518	1593	forward	gp446		
ORF347	304523	305035	513	forward	gp447	VVB52321, Uncharacterised protein [uncultured archaeon], 83.04 %, 29.1 %, 1.43e-11	
ORF348	305039	308281	3243	forward	gp448		