Genomics of bacterial and archaeal virus isolates from extreme aquatic environments

Ana Senčilo

Institute of Biotechnology and
Division of General Microbiology
Department of Biosciences
Faculty of Biological and Environmental Sciences and
Viikki Doctoral Program in Molecular Biosciences and
Doctoral Program in Microbiology and Biotechnology
University of Helsinki

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Biological and Environmental Sciences, University of Helsinki, for public examination in Lecture Hall 2 of Info Center Corona Viikinkaari 11, Helsinki, on 7th of November 2014, at 12 o'clock noon.

HELSINKI 2014

Supervisor
Docent Elina Roine
Institute of Biotechnology and
Department of Biosciences
University of Helsinki
Finland

Reviewers
Docent Petri Auvinen
Institute of Biotechnology
University of Helsinki
Finland

Professor Mikael Skurnik
Department of Bacteriology and Immunology
The Haartman Institute
University of Helsinki
Finland

Thesis committee
Docent Petri Auvinen
Institute of Biotechnology
University of Helsinki
Finland

Docent Päivi Onkamo
Department of Biosciences
University of Helsinki
Finland

Custos
Professor Benita Westerlund-Wikström
Department of Biosciences
University of Helsinki
Finland

Opponent
Professor Debbie Lindell
Faculty of Biology
Technion – Israel Institute of Technology
Israel

Cover picture: Schematic depiction of morphotypes, genome maps and GC-skews of the viruses studied in the thesis. Programs used to design the picture: GenSkew, Circos, CorelDRAW. Designed by Ana Senčilo.

ISBN 978-951-51-0133-4 (pbk.)
ISBN 978-951-51-0134-1 (PDF)

Unigrafia Oy Helsinki University Print
Helsinki 2014

ORIGINAL PUBLICATIONS

This thesis is based on the following publications, which are referred to by their Roman numerals in the text:

I Senčilo, A., Paulin, L., Kellner, S., Helm, M., and Roine, E. (2012). Related haloarchaeal pleomorphic viruses contain different genome types. Nucleic Acids Res. *40*, 5523-5534.

II Senčilo, A., Jacobs-Sera, D., Russell, D.A., Ko, C.C., Bowman, C.A., Atanasova, N.S., Österlund, E., Oksanen, H.M., Bamford, D.H., Hatfull, G.F., Roine, E. and Hendrix, R.W. (2013). Snapshot of haloarchaeal tailed virus genomes. RNA biol. *10*, 803-816.

III Senčilo, A., Luhtanen, A.-M., Saarijärvi, M., Bamford, D.H. and Roine, E. Cold-active bacteriophages from the Baltic Sea ice have diverse genomes and virus-host interactions. Environ. Microbiol., in press.

Doctoral candidate's contribution :

I AS took part in annotating and comparing the genomes and proviral regions, designing and performing experiments to describe the genome types of HRPV-3 and HGPV-1, as well as writing the manuscript.

II AS took part in annotating the genomes, performing and analyzing the results of comparative genomics studies as well as writing the manuscript.

III AS took part in sequencing and annotating the genomes, performing comparative genomics analyses and studying the genome types of the viruses as well as writing the manuscript.

ABBREVIATIONS

ACV	*Aeropyrum* coil-shaped virus
AMGs	auxiliary metabolic genes
BLAST	Basic Local Alignment Search Tool
bp	base pair
Consed	Consensus Visualization and Editing Program
CRISPR	Clustered regularly interspersed short palindromic repeat
DIG	Digoxigenin
dsDNA	Double-stranded DNA
DTRs	Direct terminal repeats
DV	Divergent regions
erf	essential recombination factor
EXPASY	Expert Protein Analysis System
FFAS03	Fold and Function Assignment System 03
HCTV-1	*Haloarcula californiae* head-tail virus 1
HCTV-2	*Haloarcula californiae* head-tail virus 2
HCTV-5	*Haloarcula californiae* head-tail virus 5
HGPV-1	*Halogeometricum* sp. pleomorphic virus 1
HGT	horizontal gene transfer
HGTV-1	*Halogranum* head-tail virus 1
HHpred	Homology Detection and Structure Prediction
HHPV-1	*Haloarcula hispanica* pleomorphiv virus 1
HHTV-1	*Haloarcula hispanica* head-tail virus 1
HHTV-2	*Haloarcula hispanica* head-tail virus 2
HRPV-1	*Halorubrum* sp. pleomorphic virus 1
HRPV-2	*Halorubrum* sp. pleomorphic virus 2
HRPV-3	*Halorubrum* sp. pleomorphic virus 3
HRPV-6	*Halorubrum* sp. pleomorphic virus 6
HRTV-4	*Halorubrum* head-tail virus 4
HRTV-5	*Halorubrum* head-tail virus 5
HRTV-7	*Halorubrum* head-tail virus 7
HRTV-8	*Halorubrum* head-tail virus 8
HSTV-1	*Haloarcula sinaiiensis* tailed virus 1
HSTV-2	*Halorubrum sodomense* tailed virus 2
HVTV-1	*Haloarcula vallismortis* tailed virus 1
ICTV	International Committee on Taxonomy of Viruses
ITRs	Inverted terminal repeats
LC-MS/MS	Liquid chromatography coupled to tandem mass spectrometry
MBN	Mung Bean Nuclease
MCM	Minichromosome maintenance
MCP	Major capsid protein
MDRs	Major different regions
MEGA	Molecular Evolutionary Genetics Analysis

Muscle	Multiple Sequence Comparison by Log- Expectation
nt	Nucleotide
ORFs	Open reading frames
PCR	Polymerase chain reaction
PCV1	Porcine circovirus 1
PesLSs	Promoter stem loop structures
Phrap	Phil's Revised Assembly Program
Phred	Phil's Read Editor
phyml	Phylogenetic Estimation using Maximum Likelihood
PHYRE2	Protein Homology/analogY Recognition Engine V 2.0
Praline	Profile Alignement
qTEM	quantitative TEM
RCR Rep	Rolling circle replication initiation protein
R-M	restriction-modification
RT-PCR	Reverse transcription PCR
SDS-PAGE	Sodium dodecyl sulphate-polyacrylamide gel electrophoresis
ssDNA	Single-stranded DNA
TEM	Transmission electron microscopy
terL	Large terminase subunits
TMHMM	Transmembrane Helices; Hidden Markov Model
TP	Terminal proteins
VLPs	Virus-like particles
wHTH	Winged helix-turn-helix domain
φ11b	Phage 11b

SUMMARY

Viruses are ubiquitous, abundant and diverse members of the biosphere. Numerous sequencing projects focusing on isolated viruses and uncultured viral communities (metaviromes) have demonstrated that viruses harbor unprecedented genotypic richness. The genomics of some viruses, for example, tailed bacteriophages infecting several widely known hosts from moderate environments, has been studied relatively well. However, viruses are known to reside in various environments, including the extreme ones, and our knowledge on the genetic make-up of these viral populations is very superficial.

In this PhD thesis, the genomics of the archaeal and bacterial viruses isolated from previously sparsely sampled extreme aquatic environments was studied. The genomes of altogether twenty haloarchaeal pleomorphic and tailed viruses from hypersaline environments as well as tailed bacteriophages from the sea ice were sequenced and analyzed. The largest portion of the genomic sequences was shown to encode proteins with no homologues in current databases emphasizing genetic distinctiveness of the studied viruses from the ones described previously.

However, all tailed viruses from both hypersaline environment and sea ice were predicted to have a cluster of genes coding for functional analogues of virion assembly and structure components of other tailed phages. Overall arrangement of this gene cluster was conserved. Haloarchaeal pleomorphic viruses were also shown to share a conserved group of genes coding for the structural and hypothetical proteins. Based on the genome organization, haloarchaeal pleomorphic viruses were classified into three subgroups. The members of one of the subgroups were demonstrated to have an unusual genome type, consisting of single-stranded and double-stranded DNA regions. In one of the viruses switches between the regions were found to be associated with a conserved DNA motif. This genome type has not been reported previously for other viruses infecting prokaryotes.

To conclude, annotation and analyses of the viral genome contents performed in this PhD thesis offered a glimpse into the diversity of putative functions of the studied viruses. Conducted comparative genomics analyses revealed different levels of relatedness among the viruses within the studied groups and similarities shared with other earlier described viruses. Overall, this work provided new insights into the genomics of understudied viruses residing in hypersaline and cold aquatic environments.

TABLE OF CONTENTS

A. LITERATURE REVIEW

1. Introduction

Virus is an infective agent typically consisting of a genome encapsulated in a proteinaceous shell termed capsid. Many viruses also have lipid constituents either in a form of a membrane or as a protein modification (King *et al.*, 2012; Pietilä *et al.*, 2013a). Viruses harbor different types of genomes including segmented or continuous, double- or single-stranded, DNA or RNA molecules (King *et al.*, 2012). The genome sizes of the viruses vary in accordance with the overall virion dimensions and complexity. One of the simplest viruses known to date is a porcine circovirus 1 (PCV1) measuring 17 nm in diameter and encoding only the capsid protein and the replicase in its 1,800 nt genome (Tischer *et al.*, 1982). "Pandoravirus" genus members are the most complex described viruses harboring 2 to 2.5 Mbp genomes in micrometer-sized particles (Philippe *et al.*, 2013). Regardless of the complexity, however, viruses lead an obligate parasitic lifestyle, i.e. they cannot reproduce outside their host cell.

It is estimated that there are 10^{30}-10^{31} viruses in the world, which is ten times higher than the estimated abundance of cellular organisms (Bergh *et al.*, 1989; Rohwer *et al.*, 2009). Viruses infect organisms from all three domains of life: Eukarya, Archaea and Bacteria. The most widely known system to classify isolated viruses was established by The International Committee on Taxonomy of Viruses (ICTV). The system is hierarchical and is based on a number of criteria. The most impactful criteria are the viral host, virion morphotype and the nature of the genome. According to the latest, ninth, report of the ICTV, 2284 virus and viroid species are classified into 349 genera, 87 families and 6 orders (King *et al.*, 2012). ICTV also deals with virus nomenclature, the standardized system of which has been proposed recently (Kropinski *et al.*, 2009).

Currently described bacterial and archaeal viruses display a variety of morphotypes (Figure 1). However, the largest portion (approximately 96%) is represented by head-and-tail (or tailed) viruses (Ackermann and Prangishvili, 2012). Viruses with tailed morphotype infect both bacteria and archaea. Regardless of the host, tailed viruses are classified into a single order Caudovirales encompassing three families. Classification into families is based on the details of viral tail morphology. Viruses with long contractile tails belong to a family *Myoviridae.* Long non-contractile tails are characteristic of *Siphoviridae* members, whereas short ones – of *Podoviridae* (King *et al.*, 2012) (Figure 1).

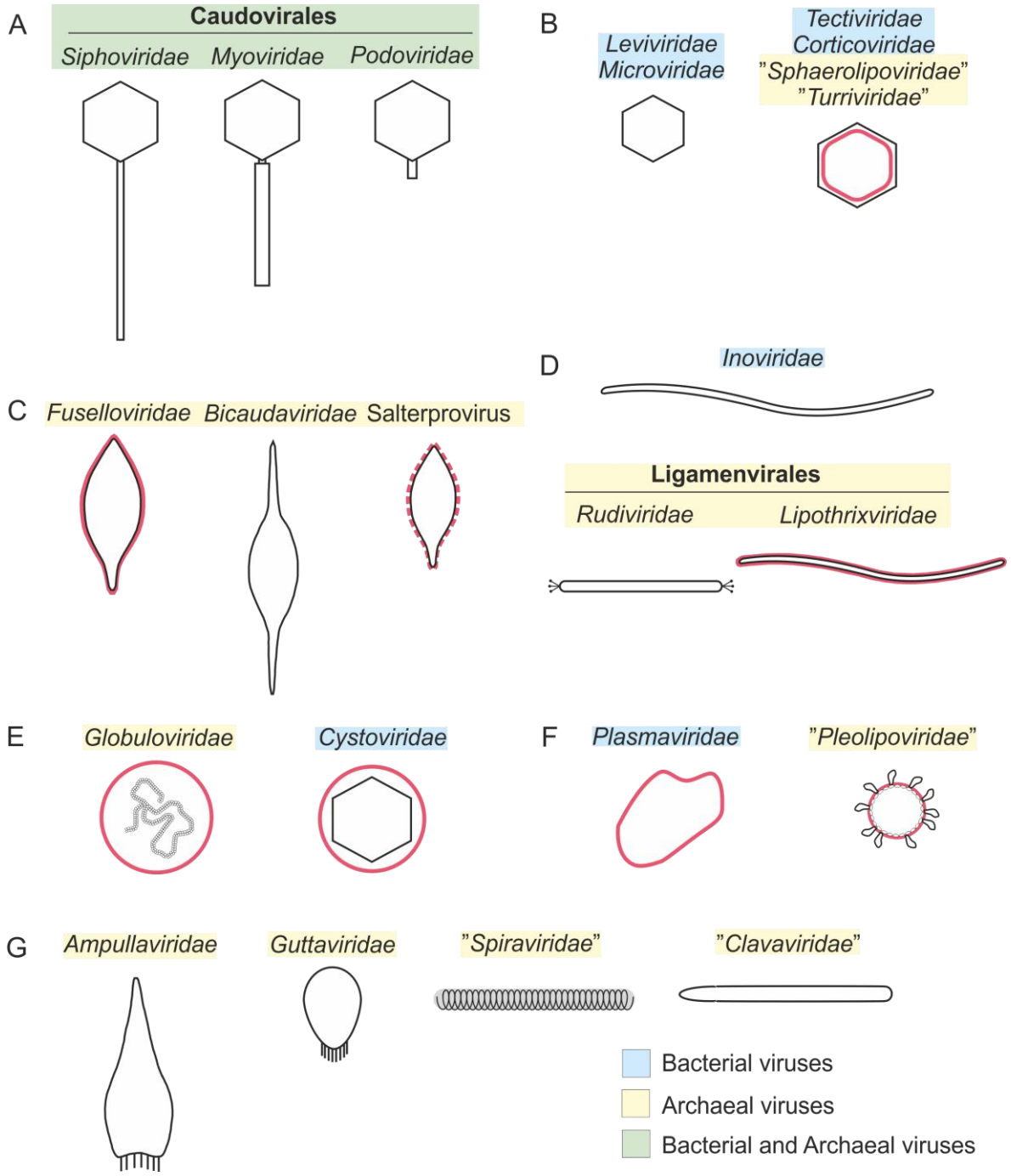

Figure 1. Morphotypes of currently described bacterial and archaeal viruses. (**A**) Tailed viruses; (**B**) Icosahedral viruses; (**C**) Spindle-shaped viruses; (**D**) Filamentous viruses; (**E**) Spherical viruses; (**F**) Pleomorphic viruses; (**G**) Bottle-shaped, droplet-shaped, coil-shaped and bacilliform viruses. Internal/external lipid membranes are marked as a solid pink line, whereas protein structures containing lipid modifications are marked with a pink dashed line. Virion sizes are not to scale. Modified from (Pina *et al.*, 2011; Ackermann and Prangishvili, 2012; Mochizuki *et al.*, 2012; Pietilä *et al.*, 2013a).

2. Prokaryotic viruses in aquatic environments

Aquatic environments and their inhabitants constitute the largest ecosystem on Earth. The two main types of aquatic environments are freshwater and marine. These environments encompass areas with extreme conditions, which are dictated by physical factors such as temperature and pressure or geochemical factors like salinity and pH. Thermal springs, deep-sea vents, alkaline and hypersaline lakes as well as ice cover are classified as extreme aquatic environments (Rothschild and Mancinelli, 2001; Pikuta *et al.*, 2007).

Both moderate and extreme aquatic environments are inhabited by organisms from the three domains of life, Bacteria, Archaea and Eukarya, as well as their viruses (Zinger *et al.*, 2012). It was estimated that one milliliter of sea water contains 10^3 eukaryotic (Brown *et al.*, 2009) and 10^5 prokaryotic cells (Whitman *et al.*, 1998). First studies on viruses from aquatic environments were conducted in 1955, when Spencer and colleagues isolated a marine bacteriophage. However, based on the inability to isolate more phages from the seawater samples, it was suggested that viruses were rare in aquatic systems (Spencer, 1955). It was not until 24 years later when the earliest microscopic observations of the particles resembling viruses (virus-like particles, VLPs) in seawater samples were made (Sieburth, 1979; Torrella and Morita, 1979). A study by Torrella and Morita (1979) gave an estimate of more than 10^4 VLPs per milliliter of seawater, which was the highest documented estimate of VLP abundance at that time. Current estimates range from less than 10^4 to more than 10^8 VLPs per milliliter in different aquatic environments varying both on the spatial and temporal scales (Wommack and Colwell, 2000).

2.1 Aquatic virus diversity

Transmission electron microscopy (TEM) was one of the first methods employed to study the diversity and abundance of viruses in different aquatic systems. Studies on viral morphotypes from the freshwater and marine environments led to rather contradictory conclusions. Some studies reported the predominance of the tailed VLPs (Proctor and Fuhrman, 1990; Demuth *et al.*, 1993), whereas others suggested that non-tailed icosahedral VLPs prevailed in some of the sampled regions (Hara *et al.*, 1991). In the latest study relying on a new quantitative TEM (qTEM) method the morphologies of marine VLPs were analyzed on a global scale. The results suggested that, overall, non-tailed icosahedral viruses constitute the largest portion of marine viral assemblages (Brum *et al.*, 2013). While tailed and non-tailed icosahedral VLPs have been commonly observed in freshwater and marine systems, studies on extreme aquatic environments such as hot springs and hypersaline lakes showed a different picture. Based on the microscopic observations, spherical and spindle-shaped VLPs were suggested to numerically dominate hypersaline environments (Guixa-Boixareu *et al.*, 1996; Oren *et al.*, 1997; Sime-Ngando *et al.*, 2011). Relative numbers of spindle-shaped VLPs were shown to increase with the increasing salinity (Guixa-Boixareu *et al.*, 1996; Santos *et al.*, 2007; Boujelben *et al.*, 2012). Several unusual morphotypes such as star-, hook-, reed- and hairpin-shaped as well as bacilliform were observed among VLPs in samples taken from Dead Sea and Lake Retba (Oren *et al.*, 1997; Sime-Ngando *et al.*, 2011). Hot springs were also shown to contain rich morphological variety of VLPs including tailed, filamentous, spindle-shaped, pleomorphic

as well as some complex morphotypes (Rachel *et al.*, 2002; Breitbart *et al.*, 2004). Even though microscopic observations reveal the variety of VLP morphotypes present in the environment, they do not provide information on the diversity of viruses having same morphotypes.

Other methods used to study the uncultured virus diversity include genomic or amplicon fingerprint analyses, phylogenetic analyses of specific viral genes as well as metagenomics (Thurber, 2009). With the onset of sequencing era, metagenomics became the dominant tool for the assessment of viral diversity. Over 50% of the viral sequences obtained through metagenomics from multiple environments do not have any homologues in current databases (Rosario and Breitbart, 2011) suggesting that viruses harbor high unexplored genetic diversity. Indeed, using metagenomic approach it was estimated that there were more than 5000 viral genotypes in 100 liters of sea water (Breitbart *et al.*, 2002). Even though the majority of the viral sequences obtained through metagenomics cannot be attributed to any currently described viral group (Rosario and Breitbart, 2011), metagenomic studies gave invaluable insights into the composition and structure of various aquatic virus communities (Breitbart *et al.*, 2002; Angly *et al.*, 2006; Schoenfeld *et al.*, 2008; Santos *et al.*, 2010). Comparison of metaviromes retrieved from marine water in geographically distant locations showed that the same viral communities were present globally, but the abundance ratios of different virus types varied between the locations (Angly *et al.*, 2006). Studies on metaviromes from extreme aquatic environments such as hot springs and hypersaline lakes led to similar conclusions (Schoenfeld *et al.*, 2008; Sime-Ngando *et al.*, 2011).

2.2 Types of viral life cycles

Viral life cycle can be roughly divided into following stages: virus adsorption, delivery of viral genomic material into the host cell, expression and replication of the viral genome, virion assembly and virus release. Based on the variations in these stages, three main types of prokaryotic virus life cycles can be delineated: lytic, lysogenic and chronic. Characteristic feature of the lytic life cycle is propensity of the virus to redirect host metabolism towards the production of viral components upon entry and to lyse the host cells upon exit. In lysogenic life cycle virus has to undertake a "lysogenic decision" as a result of which the virus may either lyse the host or establish a proviral state. In a proviral state virus can integrate its genome into the host chromosome or be maintained in the host cell as a plasmid. Triggered by environmental factors the virus can be induced from the proviral state. This results in virion production and release via cell lysis. Viruses displaying lysogenic life cycle are termed temperate, whereas viruses with strictly lytic life cycle are designated lytic or virulent. Chronic life cycle differs from the lysogenic and lytic ones in the virus release stage. Instead of lysing the host cells, viruses exit via budding or other non-lytic extrusion often retarding the host growth (Weinbauer, 2004; Calendar, 2005).

The ability to establish a proviral state allows virus to endure times when host is not at a good capacity to support viral replication or when host abundances are low. Therefore, lysogeny is thought to be a widespread phenomenon in oligotrophic and extreme aquatic environments (Wommack and Colwell, 2000; Porter *et al.*, 2007; Paul, 2008; Anesio and Bellas, 2011). Approximately half of bacterial isolates from seawater could produce VLPs

upon chemical induction confirming that lysogeny is common in marine environment. Similar estimates of the extent of lysogeny were also obtained when available genomic sequences of marine bacteria were screened for putative prophages (Paul, 2008). Lysogeny is thought to be common in hypersaline environments since putative proviral regions were identified in numerous genomes of haloarchaeal isolates (Porter *et al.*, 2007; Krupovič *et al.*, 2010; Roine and Oksanen, 2011; Porter *et al.*, 2013).

2.3 Role of viruses in aquatic environment

Viruses can exert different effects on their prokaryotic hosts. In aquatic systems viruses were suggested to be the major cause of prokaryote mortality (Fuhrman, 1999; Wommack and Colwell, 2000). It has been estimated that 10^{23} viral infections are initiated every second in the ocean (Suttle, 2007). By lysing their microbial hosts viruses cause the release of nutrients back to the pool of dissolved organic matter and thereby have a considerable impact on global carbon and energy cycling (Wilhelm and Suttle, 1999; Suttle, 2007; Weitz and Wilhelm, 2012). In addition, through lysis viruses control the abundance of dominant competitive prokaryote populations providing a niche for other populations. In this way viruses maintain prokaryotic species richness (Weinbauer and Rassoulzadegan, 2004).

Viruses structure the genetic architecture of their hosts by serving as vectors for horizontal gene transfer (HGT) or simply integrating into the host genome. Viruses also act as a reservoir of metabolic (Sullivan *et al.*, 2006; Vidgen *et al.*, 2006) and virulence genes (Waldor and Mekalanos, 1996; Brüssow *et al.*, 2004). Upon acquisition of genes from that reservoir hosts gain new functions which can alter their phenotype. In some cases the alteration is so dramatic that it allows host to occupy a new niche and hence changes its environmental role (Brüssow *et al.*, 2004; Paul, 2008; Rohwer and Thurber, 2009).

3. Genomics and evolutionary dynamics of prokaryotic viral isolates

The genomic information on prokaryotic viruses can be retrieved by various approaches including metagenomics, sequencing the genomes of viral isolates or screening bacterial and archaeal genomes for integrated proviruses (Hatfull and Hendrix, 2011). All three approaches offer useful insights into the genomics of prokaryotic viruses. However, complete genomes of isolated viruses provide a more solid background for studies on viral genome contents and organization as well as evolutionary mechanisms shaping them. Moreover, availability of viral isolates, which were the source of the obtained genomic data, allows performing further experimental work intended to reveal the functions encoded in the viral genomes.

3.1 Genetic diversity of isolated prokaryotic viruses

Genetic information of currently described bacterial viruses is encoded in either single- or double- stranded DNA or RNA molecules (Table 1). The isolated archaeal viruses almost exclusively harbor double-stranded DNA (dsDNA) genomic molecules (King *et al.*, 2012). The only exceptions are *Halorubrum* sp. pleomorphic virus 1 (HRPV-1) and *Aeropyrum* coil-shaped virus (ACV) having single-stranded DNA (ssDNA) genomes (Pietilä *et al.*, 2009; Mochizuki *et al.*, 2012). Notably, 25 knt ACV genome is the largest ssDNA genome known (Mochizuki *et al.*, 2012). The genome sizes of prokaryotic viruses range from close to 3.5 kb in leviviruses (Inokuchi *et al.*, 1986; Kannoly *et al.*, 2012) to almost 500 kb in myovirus G, which was termed a Jumbo phage for the correspondingly large virion size (Hendrix, 2009).

Table 1. Types of genomes harbored by bacterial and archaeal viruses.

Host	Virus family (or genus)	Genome type
B	*Plasmaviridae*	dsDNA, C
	Corticoviridae	dsDNA, C
	Tectiviridae	dsDNA, L
	Inoviridae	ssDNA, C
	Microviridae	ssDNA, C
	Leviviridae	ssRNA, L
	Cystoviridae	dsRNA, L, S
B/A	*Myoviridae*	dsDNA, L
	Siphoviridae	dsDNA, L
	Podoviridae	dsDNA, L
A	*Lipothrixviridae*	dsDNA, L
	Rudiviridae	dsDNA, L
	Globuloviridae	dsDNA, L
	Ampullaviridae	dsDNA, L
	Guttaviridae	dsDNA, C
	Fuselloviridae	dsDNA, C
	Bicaudaviridae	dsDNA, C
	Salterprovirus	dsDNA, L
	”*Sphaerolipoviridae*”	dsDNA, C/L
	”*Turriviridae*”	dsDNA, C
	”*Spiraviridae*”	ssDNA, C
	”*Clavaviridae*”	dsDNA, C
	”*Pleolipoviridae*”	dsDNA, C/L ssDNA, C

C – circular; L – linear; S – segmented; B – bacterial; A – archaeal.

The number of completely sequenced genomes belonging to viral isolates infecting bacteria and archaea is approaching 2000 (estimated based on the sequences deposited in EMBL database as of 5/2014). This number is rather small in comparison with the

anticipated abundance of viruses (Rohwer *et al.*, 2009). In addition, currently available genomic sequences do not represent different viral groups equally. In accordance with the predominance of tailed morphotype among the isolated viruses, vast majority of the available genome sequences belong to tailed viruses (Krupovič *et al.*, 2011; Ackermann and Prangishvili, 2012). Tailed viruses infecting hosts from *Bacillus*, *Lactococcus*, *Mycobacterium*, *Pseudomonas*, *Staphylococcus* and several enterobacterial genera are largely overrepresented among the viruses with completely sequenced genomes. Therefore, much of what we currently know about the genomics of prokaryotic viruses comes from studies on the genomes of tailed phages infecting bacteria belonging to several genera.

However, even among the genomes of viruses infecting the same host substantial genetic diversity is observed. Mycobacteriophages, the largest group of viruses with completely sequenced genomes infecting the same host, *Mycobacterium smegmatis* mc^2155, serve as the best example for that. Despite sharing the same host, a total of 221 mycobacteriophages displayed extensive genetic variation. The phages were classified into 15 clusters encompassing members with sequence similarity along over half of their genome lengths and 8 singletons, i.e. phages having no close relatives in the described collection (Hatfull, 2012a).

Along with the growing body of research on the bacteriophage genomics, recent efforts have also focused on describing the genetic diversity of less explored territory of the virosphere, i.e. of viruses infecting archaea. Compared to the reported bacterial viruses, archaeal viruses display a variety of unusual morphotypes and mechanisms of interactions with their hosts (Pina *et al.*, 2011; Prangishvili, 2013). This is reflected by the fact that archaeal viruses have much more genomic "dark matter", i.e. genes coding for proteins having no predicted function and no homologues among the proteins of the previously described viruses (Prangishvili *et al.*, 2006). Further studies on the genetic diversity and encoded functions of these viral groups give a promise of new exciting discoveries.

3.2 Viral genome organization

The genomes of archaeal and bacterial viruses are organized in modules, in which genes are grouped according to the function of the encoded proteins. Expression of genes within the modules is often regulated by common promoters. Modules coding for proteins involved in transcription regulation, replication/recombination, assembly and structure of the virions as well as lysis/integration are commonly annotated in the genomes of prokaryotic viruses (Hatfull and Hendrix, 2011; Krupovič *et al.*, 2011). Genome modularity is not equally apparent in all viruses. For example, the genomes of giant phages seem to be less stringent in organization (Mesyanzhinov *et al.*, 2002; Cornelissen *et al.*, 2012; Šimoliūnas *et al.*, 2013). In accordance with the overall complexity, single functional modules are often comprised of several gene clusters located in different parts of giant phage genome.

Subsets of genes are conserved among the related viruses and therefore comprise the "core" set of genes in that viral group. For example, core genes of T4-like bacteriophages are coding for the proteins involved in virion structure and assembly as well as genome replication (Comeau *et al.*, 2007). The core gene set of fuselloviruses encompasses genes encoding structural proteins, integrase and putative transcriptional regulators (Redder *et al.*, 2009).

Definition of the core genes is empirical and the content of the core gene set may change as new genomes of related viruses are added to the studied group. Invariably, however, genes coding for the key structural and assembly components of the virion are present in all related viruses and generally comprise a stable constituent of the core gene set. Typically a module of structural and assembly genes displays conserved synteny even when the nucleotide and amino acid sequences have diverged beyond the significant similarities (Krupovič *et al.*, 2011). For example, order of structural and assembly genes is conserved among many archaeal and bacterial tailed viruses (Krupovič *et al.*, 2010). The arrangement and contents of gene modules coding for other functions such as lysis/integration and replication/recombination display much more variation even among tailed viruses infecting the same host (Hatfull, 2010).

Noncore genes, as definition implies, are encountered only in a number of members within the group of related viruses. Some of the noncore genes are found inserted between the genes in the conserved clusters shared among related viruses (Hendrix *et al.*, 2000; Juhala *et al.*, 2000). Other noncore genes are concentrated in hyper plastic genome regions (Comeau *et al.*, 2007).

A large portion of the noncore genes codes for the proteins of unknown function (Hatfull and Hendrix, 2011). However, for some of the noncore genes functions have been either determined or predicted. For instance, many phages are known to carry the so called auxiliary metabolic genes (AMGs), the cellular metabolic genes presumably acquired by viruses from their hosts (Breitbart *et al.*, 2007; Comeau *et al.*, 2008; Sharon *et al.*, 2011). Photosynthesis genes commonly found in cyanophage genomes are perhaps the best studied example of the AMGs (Mann, 2003; Lindell *et al.*, 2004, 2005; Sullivan *et al.*, 2006; Sharon *et al.*, 2007). Investigation of the *psbA* and *psbD* coding for photosystem components in *Prochlorococcus* phage P-SSP7 showed that these genes were expressed during the viral infection and were suggested to support host metabolism for viral particle production (Lindell *et al.*, 2005).

A number of other viral noncore genes were suggested to confer advantage to a bacterium when expressed from a provirus (Hendrix *et al.*, 2000, 2003). For example, proviral genes coding for Shiga toxin protect host bacterium from bacterivorous predators (Lainhart *et al.*, 2009). These noncore genes create a selection pressure for the bacterium to retain the integrated provirus within the genome and thereby increase viral fitness too (Hendrix *et al.*, 2000). However, generally, noncore genes are not essential for viral progeny production. Therefore, they were suggested to serve as evolution hotspots, where novel functions may be created (Hatfull and Hendrix, 2011).

3.3 Genome mosaicism and evolution

Genome regions spanning from a gene segment to entire modules can be exchanged among viruses (Haggard-Ljungquist *et al.*, 1992), plasmids and other mobile genetic elements (Osborn and Böltner, 2002; Peng, 2008) as well as host chromosome (Lindell *et al.*, 2004) via an HGT. As a result, the genomes of prokaryotic viruses appear to be mosaics consisting of segments with different evolutionary histories (Figure 2) (Hendrix *et al.*, 1999).

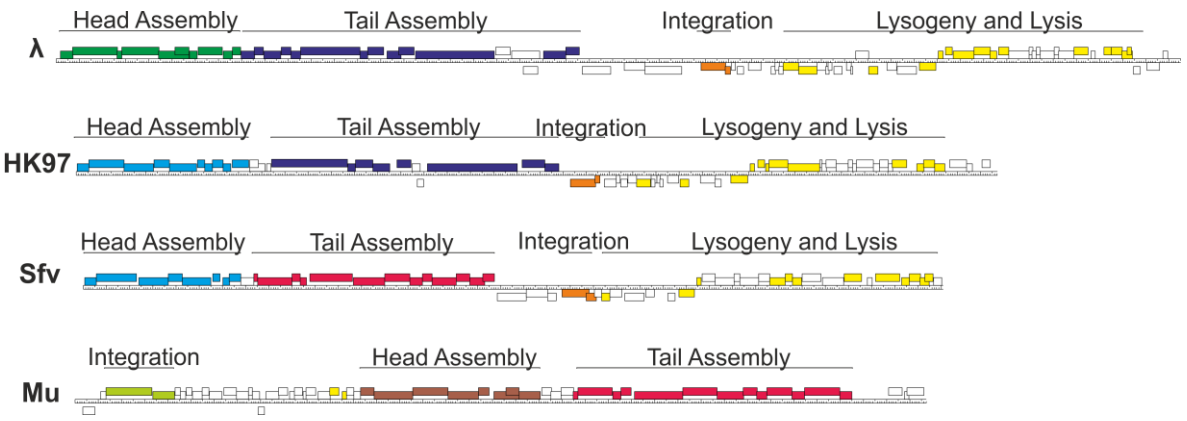

Figure 2. Mosaic genome structure of tailed bacteriophages λ, HK97, Mu and Sfv. Modules of genes coding for homologous proteins shared among the viruses are marked with the same colour. The figure was modified from (Lawrence *et al.*, 2002).

New genetic material can be incorporated into the genome by the means of homologous and non-homologous (illegitimate) recombination (Hendrix, 2002). Homologous recombination can occur both between highly similar intragenic regions (Martinsohn *et al.*, 2008; Redder *et al.*, 2009) and between the short homologous sequences at the boundaries of the genes (Clark *et al.*, 2001). Illegitimate recombination was suggested to occur randomly and give rise to a large number of viral genome variants, most of which deem the virus unable to produce a viable progeny. However, small portion of the variants, which turn out to be functional, are retained in the viral population. For that reason most of the mosaicism boundaries are found in the intergenic regions or between the gene segments coding for different protein domains (Hendrix, 2002). As a result of the illegitimate recombination, some related viruses harbor non-homologous, but functionally equivalent genes, a phenomenon termed non-orthologous gene replacement (Stassen *et al.*, 1992; Krupovič and Bamford, 2007; Krupovič *et al.*, 2010).

Mosaic genome architectures were documented in a number of bacterial and archaeal viruses including tailed viruses (Hendrix *et al.*, 1999; Juhala *et al.*, 2000), inoviruses (Lawrence *et al.*, 2002) corticoviruses (Krupovič and Bamford, 2007), fuselloviruses (Redder *et al.*, 2009), globuloviruses (Ahn *et al.*, 2006) lipothrixviruses (Vestergaard *et al.*, 2008) and rudiviruses (Peng *et al.*, 2001). The relative extent, to which illegitimate and homologous recombination events contribute to mosaicism, differs among the viral groups. Illegitimate recombination was suggested to play the most prominent role in shaping the genomes of tailed bacteriophages (Hendrix *et al.*, 2000). Although analysis of lamboid phage genomes indicated that homologous and illegitimate recombination may have an equal contribution to the genomic mosaicism of this group of phages (Martinsohn *et al.*, 2008). The genomes of fuselloviruses, which are highly mosaic with respect to each other, were also proposed to be affected by frequent homologous recombination events. Homologous recombination is likely facilitated by a tandem integration of fuselloviral genomes, which produces proviral concatemers in the host chromosome (Redder *et al.*, 2009).

Genomic mosaicism is less characteristic of smaller viruses with ss/dsRNA and ssDNA genomes. For example, only a single recombination event was predicted in over 30 studied leviviruses (Friedman *et al.*, 2012). Recombination frequency seems to be also very limited in microviruses (Rokyta *et al.*, 2006) and cystoviruses (Onodera *et al.*, 2001; Silander *et al.*, 2005). Viruses with these genome types are estimated to have higher mutation rates than dsDNA viruses (Sanjuan *et al.*, 2010) and therefore are thought to mainly diversify through the accumulation of mutations. Although tectiviruses have dsDNA genomes, PRD1-like group of tectiviruses is also thought to evolve mainly through the accumulation of mutations, since genomic mosaicism is not well pronounced in this group. This may imply that PRD1-like viruses possess an optimal gene set and any change to it could reduce the fitness of the virus (Saren *et al.*, 2005).

Unlike other bacterial and archaeal viruses, cystoviruses have fragmented genomes. Reassortment of the genomic segments was suggested to be another mechanism having significant contribution to the evolution of this group of viruses (Onodera *et al.*, 2001; Silander *et al.*, 2005).

Ability to infect the host and reproduce are the main factors driving virus evolution. Host evolution is also significantly influenced by the development of resistance mechanisms employed to prevent virus infection (Buckling and Brockhurst, 2012). These mechanisms target different stages of virus life cycle. Hosts can prevent the initial step of viral infection, an adsorption, by modifying or blocking receptors recognized by a virus or by producing extracellular matrix (Labrie *et al.*, 2010; Buckling and Brockhurst, 2012). If the adsorption has taken place, hosts can prevent viral genome internalization or recognize and destroy foreign DNA by the means of clustered regularly interspersed short palindromic repeat (CRISPR) or restriction-modification (R-M) systems (Tock and Dryden, 2005; Labrie *et al.*, 2010; Al-Attar *et al.*, 2011). CRISPR system encompasses repeat-spacer array and a set of associated *cas* genes. Sequence information of the spacers is employed to recognize and target foreign DNA for the cleavage by Cas proteins. Spacers are acquired by the host from the encountered viruses. Therefore, CRISPR systems confer adaptive and heritable immunity to prokaryotes (Al-Attar *et al.*, 2011).

3.4 “Viral self” and evolutionary lineages of viruses

With the advent of the genomics era, virus classification came to rely more on the genomic nucleotide sequence and protein amino acid sequence comparisons. Meanwhile genomic information is instrumental in resolving phylogenetic relationships between closely related viruses, fast-paced virus evolution makes it impossible to use nucleotide and amino acid sequence-based phylogeny to resolve relationships between distantly related viruses.

However, virion architecture is conserved among the related viruses lacking sequence similarities. Common themes in virion architecture were also observed among the viruses infecting hosts from different domains of life (Benson *et al.*, 1999; Bamford *et al.*, 2002; Bamford, 2003). The observed similarities included principles of virion structure and assembly as well as genome packaging. These features were suggested to constitute the “viral self”, which, if shared among the viruses, may point at the common ancestry (Bamford *et al.*, 2002).

Based on these observations and assumptions, the virus lineage hypothesis was proposed (Bamford *et al.*, 2002). The hypothesis states that all extant viruses originated from a number of ancestral viruses, urviruses, which existed before the separation of living organisms into the three domains of life. In evolutionary time span viruses co-evolved with their hosts and structures as well as functions involved in virus-host interactions diverged. However, the principles of virion architecture, the "viral self", remained conserved among the related viruses. Based on the shared "self" component, all viral universe can be divided into a limited number of lineages, within which viruses are assumed to have a common ancestor (Bamford *et al.*, 2002, 2005a; Bamford, 2003).

Based on the available information on the virion structures, representatives from close to 30 viral families as well as some unclassified viruses were grouped into four lineages: PRD1-like, HK97-like, Picorna-like and BTV-like (Abrescia *et al.*, 2012). Picorna-like and BTV-like lineages encompass bacterial and eukaryotic virus members, but lack archaeal virus representatives. PRD1-like and HK97-like lineages encompass members infecting hosts from all three domains of life (Abrescia *et al.*, 2012; Pietilä *et al.*, 2013c). PRD1-like lineage unites bacterial tectiviruses and corticoviruses together with archaeal sphaerolipoviruses and turriviruses. These viruses have similar virion structures consisting of an icosahedral capsid with an internal lipid membrane enclosing either linear or circular dsDNA genome. Major capsid proteins (MCPs) of the viruses within this lineage have either double or single β-barrel fold (Krupovič and Bamford, 2008). HK97-like lineage groups bacterial and archaeal tailed viruses together. These viruses share the same virion architectures and were suggested to have similar mechanisms for virion assembly and genome packaging (Krupovič *et al.*, 2010). MCPs of both bacterial and archaeal tailed viruses have an HK97 fold (Abrescia *et al.*, 2012; Pietilä *et al.*, 2013c).

4. Genetic diveristy of prokaryotic viruses in extreme aquatic environments

Based on metagenomic studies, viruses from both moderate and extreme aquatic environments were suggested to harbor high genetic diversity (Breitbart *et al.*, 2002; Schoenfeld *et al.*, 2008; Lopez-Bueno *et al.*, 2009; Santos *et al.*, 2012). However, studies on the genomics of viruses isolated from these environments are very sparse. The current knowledge on the genomics of viruses in hypersaline (having salinity exceeding that of seawater, i.e. 3.5% total dissolved salts) (DasSarma and DasSarma, 2012) and cold (having a temperature of 4^{o}C or below) (Wells and Deming, 2006a) aquatic systems will be described below. Both of these environments are known to harbor exceptionally high numbers of VLPs. Arctic and Antarctic sea ice was estimated to have up to 10^{8} VLPs per milliliter of melted sample (Maranger *et al.*, 1994; Gowing *et al.*, 2002, 2004). Different hypersaline lakes were reported to have up to 10^{9}-10^{10} VLPs per milliliter (Boujelben *et al.*, 2012; Santos *et al.*, 2012). These estimates suggest that among other aquatic environments both hypersaline lakes and sea ice are distinguished by very high densities of viruses (Maranger *et al.*, 1994; Santos *et al.*, 2012). Since hypersaline and cold aquatic environments harbor relatively small grazer communities, viruses are assumed to exert even bigger influence on microbial populations in these environments compared to the moderate ones (Guixa-Boixareu *et al.*, 1996; Wells and Deming, 2006b).

4.1 Viruses in hypersaline environments

Up to date, close to 70 viruses were isolated from hypersaline environments (Atanasova *et al.*, 2012; Sabet, 2012). Majority of these viruses (approximately 60 isolates) infect halophilic archaea (haloarchaea), which are known to be dominant microorganisms in hypersaline systems (Oren, 2002). The rest of the isolates are halophilic bacteriophages (Atanasova *et al.*, 2012; Sabet, 2012). However, they are rather sparsely studied. Therefore, further descriptions will concentrate on viruses of haloarchaea.

Cultured haloarchaeal viruses display four morphotypes: tailed, icosahedral, spindle-shaped and pleomorphic (Sabet, 2012). Genome sequences of isolated viruses of all four morphotypes are available (Table 2) (Krupovič *et al.*, 2011).

Table 2. Characteristics of partially/completely sequenced genomes of haloarchaeal viruses.

Morphology	Family (or genus)	Virus	Genome type; type of genome termini	Genome size (bp)	Nr of ORFs	Nr of tRNAs
Tailed	*Myoviridae*	φH	dsDNA, L; Circ perm	~59,000	ND	ND
		HF1	dsDNA, L; DTR (306 bp)	75,898	112	5
		HF2	dsDNA, L; DTR (306 bp)	77,670	116	5
		φCH1	dsDNA, L; Circ perm	58,498	98	-
		HSTV-2	dsDNA, L; DTR (340 bp)	68,187	103	1
	Siphoviridae	BJ1	dsDNA, L; ND	42,271	70	1
		HVTV-1	dsDNA, L; DTR (585 bp)	101,734	173	1
	Podoviridae	HSTV-1	dsDNA, L; Circ perm	32,189	53	-
Icosahedral	"*Sphaerolipoviridae*"	SH1	dsDNA, L; ITR (309 bp)	30,898	56	-
		PH1	dsDNA, L; ITR (337 bp)	28,064	49	-
		HHIV-2	dsDNA, L; ITR (305 bp)	30,578	43	-
		SNJ1	dsDNA, C	16,341	33	-
Pleomorphic	"*Pleolipoviridae*"	HRPV-1	ssDNA, C	7,048	9	-
		HHPV-1	dsDNA, C	8,082	8	-
		His2	dsDNA, L; ITR (525 bp)	16,067	35	-
Spindle-shaped	Salterprovirus	His1	dsDNA, L; ITR (105 bp)	14,464	35	-

C – circular; L – linear; Circ perm – circularly permuted; DTR – direct terminal repeats; ITR – inverted terminal repeats; ND – not determined.

4.1.1 Tailed viruses

Majority of the isolated haloarchaeal viruses has a tailed virion morphotype. Close to 50 haloarchaeal tailed viruses have been reported to date. More than half of the isolates are myoviruses. The rest is constituted by siphoviruses and, to date, only one podovirus (Atanasova *et al.*, 2012; Sabet, 2012). Vast majority of the viruses has not been studied beyond the basic characterization. However, several viruses were described in some detail and their genomes have been either partially or completely sequenced (Table 2) (Klein *et al.*, 2002; Tang *et al.*, 2002, 2004; Pagaling *et al.*, 2007; Pietilä *et al.*, 2013b, 2013c).

4.1.1.1 Myoviruses

Myovirus ϕH was the first haloarchaeal tailed virus studied at the molecular level. The virus was isolated upon the spontaneous lysis of its host *Halobacterium salinarum* culture. ϕH is estimated to harbor approximately 59 kb genome (Schnabel *et al.*, 1982b). In a proviral state, the ϕH genome circularizes into 57 kb molecule and resides as a plasmid in the host cells (Schnabel *et al.*, 1984).

The genome of ϕH has been only partially sequenced. Approximately 10% of the ϕH virus population were shown to contain different genome variants resulting from several indels as well as an inversion of a 12 kb L segment (Schnabel *et al.*, 1982a, 1984). The invertible L segment was shown to be capable of circularization and could be maintained as the plasmid conferring immunity to the host against the ϕH infection (Schnabel, 1984). The differences between some of the ϕH variants were attributed to the insertion element ISH1.8, which was also found in *Halobacterium salinarium* genome (Schnabel *et al.*, 1984).

Another well-studied haloarchaeal tailed virus related to ϕH is a myovirus ϕCh1 infecting a haloalkaliphile *Natrialba magadii* (Witte *et al.*, 1997; Klein *et al.*, 2002). Similarly to ϕH, ϕCh1 is a temperate virus, which was isolated after the spontaneous lysis of its host culture. In contrast to ϕH, ϕCh1 was shown to be integrated into the host chromosome rather than maintained as a plasmid (Witte *et al.*, 1997). Unlike any other virus described to date, ϕCh1 harbors several 70-800 nt-long RNA species of host origin in addition to its dsDNA genome (Witte *et al.*, 1997). Despite these differences, ϕH and ϕCh1 have homologous MCPs and share highly similar genome region corresponding to the L segment in ϕH virus (Witte *et al.*, 1997; Klein *et al.*, 2000).

The other two related myoviruses, HF1 and HF2, were isolated from the Cheetham Saltworks in Australia. Both of the viruses were suggested to be virulent and were shown to have mutually exclusive host ranges based on the several tested species belonging to *Haloferax*, *Halobacterium* and *Haloarcula* genera (Nuttall and Dyall-Smith, 1993). The genome of the virus HF1 is nearly identical to HF2 genome along approximately two thirds of its length (Tang *et al.*, 2002, 2004). The variable parts of the viral genomes share some similarity, which is interrupted by multiple indels and nucleotide substitutions. Sequence changes are mostly concentrated in two genomic regions termed the major different regions (MDRs). The most likely scenario explaining the divergence of these viruses includes a very recent recombination event(s) with MDRs serving as recombinational hot spots (Tang *et al.*, 2004).

The most recent detailed report on haloarchaeal myoviruses has focused on a *Halorubrum sodomense* tailed virus 2 (HSTV-2) (Pietilä *et al.*, 2013b). This virus was isolated from the sample taken from hypersaline lake in Eilat, Israel (Atanasova *et al.*, 2012). The genome of HSTV-2 shares close to 70% and 60% of overall nucleotide sequence similarity with the genomes of myoviruses HF1 and HF2, respectively. The major difference between the viruses results from a putative deletion of over 10 kb fragment in the left end of HSTV-2 genome relative to the genomes of HF1 and HF2 (Tang *et al.*, 2002, 2004; Pietilä *et al.*, 2013b). In HF1 and HF2 this fragment codes for enzymes involved in nucleotide metabolism (Tang *et al.*, 2002, 2004).

4.1.1.2 Siphoviruses and podoviruses

Siphovirus BJ1 was isolated from the hypersaline lake located in Inner Mongolia, China. The virus infects *Halorubrum* sp. host and is likely to be temperate as it was found to encode a putative integrase (Pagaling *et al.*, 2007). BJ1 shares a number of putative homologues with two proviral regions termed Hlac-Pro1 (approximately 29 kb-long) and Nmag-Pro1 (approximately 50 kb-long), which were identified in the genomes of *Halorubrum lacusprofundi* ATCC 49239 and *Natrialba magadii* ATVV 43099, respectively. Most of the genes coding for the shared homologous proteins are arranged in a collinear cluster, which follows the genome region coding for head structural and assembly proteins. However, in Nmag-Pro1 this cluster is interrupted by indels. BJ1 and Hlac-Pro1 share another cluster of putative genes, one of which is coding for minichromosome maintenance (MCM) helicase (Krupovič *et al.*, 2010).

Another siphovirus, a *Haloarcula vallismortis* tailed virus 1 (HVTV-1), was isolated from Samut Sakhon in Thailand (Atanasova *et al.*, 2012). At a genomic level HVTV-1 did not bear resemblance to any other described haloarchaeal virus (Pietilä *et al.*, 2013b).

Up to date, *Haloarcula sinaiiensis* tailed virus 1 (HSTV-1) is the only isolated and described haloarchaeal podovirus. The virus originates from Margherita di Savoia, Italy (Atanasova *et al.*, 2012). As inferred from the genomic sequence, HSTV-1 is rather divergent from other described viruses including the haloarchaeal tailed ones. Cryo-EM and 3D reconstruction of HSTV-1 particles yielded the structure at a 8.9 Å resolution, which is the highest obtained resolution of the haloarchaeal tailed virus structure to date. Fitting of the HK97 fold within the obtained 3D reconstruction suggested that HSTV-1 MCP indeed adopted the HK97 fold. The result gave a justification for classifying archaeal tailed viruses into HK97 virus lineage (Pietilä *et al.*, 2013c).

4.1.1.3 Genome contents, organization and mosaicism of haloarchaeal tailed viruses

From the analyses of the small set of complete and partial genome sequences of haloarchaeal tailed viruses as well as related proviral regions several tendencies emerge. In all of the viruses predicted coding regions constitute over 90% of the genomes. However, function can be predicted for only up to one fourth of the putative encoded proteins. Most of the proteins have either no similarity to the proteins in current databases or are similar to

hypothetical proteins. Those proteins for which function can be predicted generally fall into two groups. One group includes proteins involved in DNA metabolism and replication. Genes coding for such proteins as ribonucleotide reductase, thymidylate synthase, DNA polymerase, helicase, integrase/recombinase and methylase were encountered in more than one described haloarchaeal tailed virus genome. Another group of commonly annotated proteins includes proteins responsible for the genome packaging (small and large terminase subunits, portal protein) as well as structural and assembly proteins of the virion head (prohead protease, major and minor capsid proteins) and tail (tails assembly chaperones, tape measure protein) (Klein *et al.*, 2002; Tang *et al.*, 2002, 2004; Pagaling *et al.*, 2007; Krupovič *et al.*, 2010; Pietilä *et al.*, 2013b, 2013c). These proteins have functional analogues in tailed bacteriophages (Krupovič *et al.*, 2011).

Similarly to tailed bacteriophages, the genomes of haloarchaeal tailed viruses are organized in functional modules and display high degree of mosaicism (Klein *et al.*, 2002; Tang *et al.*, 2002, 2004; Pagaling *et al.*, 2007; Krupovič *et al.*, 2010; Pietilä *et al.*, 2013b, 2013c). For instance BJ1 shares several homologues with haloarchaeal myoviruses ϕH, ϕCh1 and HF1 (Pagaling *et al.*, 2007). HF1 and HF2 genomes were shown to code for proteins having homologues in a wide range of bacteria and bacteriophages (Tang *et al.*, 2002, 2004). It was suggested that the genes coding for these homologues were transferred by the HGT from the organisms adapted to moderately saline environments through the progressively halophilic organisms and to the extreme halophiles and their viruses (Tang *et al.*, 2002).

4.1.2 Icosahedral viruses

Up to date, four haloarchaeal icosahedral viruses have been described. Three of them, SH1 (Dyall-Smith *et al.*, 2003), PH1 (Porter *et al.*, 2013) and HHIV-2(Jaakkola *et al.*, 2012), are virulent viruses propagating on *Haloarcula hispanica* hosts. The fourth icosahedral virus, SNJ1, is a temperate virus isolated from the mitomycin C-treated *Natrinema* sp. J7-1 strain, in which it was maintained as a plasmid (Mei *et al.*, 2007; Zhang *et al.*, 2012). All four viruses are united into a single proposed family "*Sphaerolipoviridae*" (Dyall-Smith *et al.*, 2013). The particles of sphaerolipoviruses have no tail and consist of an icosahedral capsid with an internal lipid membrane enclosing the genomic molecule (Dyall-Smith *et al.*, 2003; Jaakkola *et al.*, 2012; Zhang *et al.*, 2012; Porter *et al.*, 2013). Structural studies on SH1 particles showed that the five-fold vertices of the capsid contain horn-like spikes, which were suggested to be involved in host recognition (Jäälinoja *et al.*, 2008).

SH1, PH1 and HHIV-2 have approximately 30 kb-long linear dsDNA genomes with inverted terminal repeats (ITRs) (Table 2) (Bamford *et al.*, 2005b; Jaakkola *et al.*, 2012; Porter *et al.*, 2013). The SH1 ITRs were shown to be covalently bound by proteins suggested to prime the genome replication (Porter and Dyall-Smith, 2008). The genomes of all three viruses are collinear (Bamford *et al.*, 2005b; Jaakkola *et al.*, 2012; Porter *et al.*, 2013). SH1 and PH1 are the most closely related sphaerolipoviruses. They share 74% nucleotide sequence identity along the whole genome length on average. Most of the differences between the two genomes are attributable to indels and replacements concentrated in the 18 divergent regions (DV1-DV18). Proteins involved in host recognition (spike proteins), which generally display high variability in other viruses (Saren

et al., 2005; Hatfull *et al.*, 2006), have relatively high amino acid similarity in SH1 and PH1, which is consistent with the shared host ranges of the two viruses (Porter *et al.*, 2013). HHIV-2 genome shares 55% and 59% nucleotide sequence identity with the genomes of PH1 and SH1, respectively (Jaakkola *et al.*, 2012; Porter *et al.*, 2013). The genome regions of HHIV-2 and SH1 coding for the putative ATPase, two MCPs and a major membrane protein share some of the highest similarities (Jaakkola *et al.*, 2012).

SNJ1 is the most disparate member of the "*Sphaerolipoviridae"*. This virus harbors circular dsDNA genome, which is half the size of the other shaerolipoviral genomes (Table 2). None of the 10 structural proteins identified in SNJ1 have homologues in SH1, PH1 or HHIV-2 (Zhang *et al.*, 2012).

Genomic comparisons of the sphaerolipoviruses with the related proviral regions in different haloarchaea suggested a high degree of recombination between them. This indicates the mosaic nature of these genetic elements (Porter *et al.*, 2013).

4.1.3 Pleomorphic viruses

Although first haloarchaeal pleomorphic virus was described only in 2009 (Pietilä *et al.*, 2009), this group of viruses quickly became the second largest of all studied haloarchaeal viruses. The seven reported haloarchaeal pleomorphic viruses were isolated from hypersaline environments located in Australia, Italy, Spain, Israel and Thailand. All seven viruses are classified into a proposed family "*Pleolipoviridae*" (Bath *et al.*, 2006; Pietilä *et al.*, 2009, 2012; Roine *et al.*, 2010; Atanasova *et al.*, 2012). Four of the viruses infect *Halorubrum* sp. hosts and were correspondingly named *Halorubrum* sp. pleomorphic viruses 1, 2, 3 and 6 (HRPV-1, HRPV-2, HRPV-3 and HRPV-6) (Pietilä *et al.*, 2009; Atanasova *et al.*, 2012). One of the viruses infects *Halogeometricum* sp. and was termed *Halogeometricum* sp. pleomorphic virus 1 (HGPV-1)(Atanasova *et al.*, 2012). The remaining two viruses, *Haloarcula hispanica* pleomorphic virus 1 (HHPV-1) and His2, infect *Haloarcula hispanica* (Bath *et al.*, 2006; Roine *et al.*, 2010). Infection by haloarchaeal pleomorphic viruses does not cause host cell lysis. The viruses do, however, retard the growth of the host cultures. It is thought that the viruses exit host cells by budding, but, evidence using electron microscopy has never been obtained (Pietilä *et al.*, 2009, 2012; Roine *et al.*, 2010).

Pleomorphic virus particles are lipid vesicles of 40 to 70 nm in diameter harboring two to three different major structural proteins and enclosing the genomic molecule. Lipid composition of the particles is similar to that of the host membrane showing that haloarchaeal pleomorphic viruses acquire lipids relatively unselectively. The viruses have two types of major structural proteins: the smaller membrane proteins and the larger spike proteins. The smaller membrane proteins are mainly embedded in the membrane, whereas the larger spike proteins are predominantly exposed on the outer surface of the virions and are thought to be responsible for host recognition and mediation of viral and host membrane fusion (Pietilä *et al.*, 2009, 2012; Roine *et al.*, 2010). Spike proteins of HGPV-1 and His2 are lipid-modified (Pietilä *et al.*, 2012). In HRPV-1 the spike proteins are N-glycosylated (Pietilä *et al.*, 2010; Kandiba *et al.*, 2012). The glycan moiety was shown to be a pentasaccharide accommodating 5-N-formyl-legionaminic acid as the terminal monosaccharide (Kandiba *et al.*, 2012). This was the first report of legionaminic acid-

containing glycan modification of archaeal protein. Much of the knowledge on the spike structures comes from the studies on HRPV-1 virus. HRPV-1 particles were studied using biochemical dissociation combined with cryo-electron microscopy and cryo-electron tomography. It was revealed that the spikes are club-shaped 7 nm-long structures randomly distributed on the surface of the virions (Pietilä *et al.*, 2012).

The genomes of three haloarchaeal pleomorphic viruses, HRPV-1, HHPV-1 and His2, were sequenced (Table 2). All three viruses have different genome types. HRPV-1, HHPV-1 and His2, have circular ssDNA, circular dsDNA and linear dsDNA genomes, respectively (Bath *et al.*, 2006; Pietilä *et al.*, 2009; Roine *et al.*, 2010). Notably, HRPV-1 was the first described archaeal virus containing an ssDNA genome. Being only 7048 nt-long, HRPV-1 genome is also the smallest of all sequenced haloarchaeal virus genomes (Pietilä *et al.*, 2009).

Despite different genome types, HRPV-1 and HHPV-1 share a collinear cluster of seven genes coding for homologous proteins. Besides the two identified major structural proteins, the cluster codes for a putative rolling circle replication initiation protein (RCR Rep), an ATPase and a number of hypothetical proteins (Pietilä *et al.*, 2009; Roine *et al.*, 2010). ATPase was identified as a minor structural protein in HRPV-1 (Pietilä *et al.*, 2009).

Virus His2 is a more divergent member of this group. It shares only four protein homologues with HHPV-1 and HRPV-1. Two of the homologues are spike protein and ATPase, whereas the other two have no predicted function. The genome of His2 is almost twice larger than the genomes of other haloarchaeal pleomorphic viruses (Bath *et al.*, 2006; Pietilä *et al.*, 2009; Roine *et al.*, 2010). The ends of the genome contain inverted terminal repeats ITRs, the 5' ends of which were shown to be covalently bound by terminal proteins (TP) (Bath *et al.*, 2006; Porter and Dyall-Smith, 2008). These proteins were suggested to prime the His2 genome replication. Since the virus was also predicted to encode a putative type B polymerase, it is likely that His2 employs protein-primed genome replication (Porter and Dyall-Smith, 2008).

A number of genetic elements reminiscent of haloarchaeal pleomorphic viruses was identified in the genomes of different haloarchaea belonging to genera *Haloferax*, *Haloarcula* and *Natromonas* suggesting that pleolipoviruses are wide-spread in hypersaline systems (Pietilä *et al.*, 2009; Roine *et al.*, 2010; Roine and Oksanen, 2011).

4.1.4 Spindle-shaped viruses

Spindle-shaped virion morphotype is rather common among the studied archaeal viruses. Most of the described spindle-shaped viruses belong to family *Fuselloviridae* and infect extreme thermophiles belonging to phylum *Crenarchaeaota* (Pina *et al.*, 2011). Currently there is only one haloarchaeal spindle-shaped virus described. His1 infecting *Haloarcula hispanica* was isolated from Australian saltern (Bath and Dyall-Smith, 1998). During the infection in laboratory conditions the release of His1 virions did not cause lysis of the host cells (Pietilä *et al.*, 2013a).

His1 virions are spindle-shaped particles with a tail at one of the tapered ends (Bath and Dyall-Smith, 1998). The particles consist of one lipid-modified MCP and three minor structural proteins (Pietilä *et al.*, 2013a). The MCP of His1 was shown to share 27% amino

acid identity with the MCP of fusellovirus SSV1 adding strength to the suggested relatedness of these two morphologically similar viruses (Pietilä *et al.*, 2013a).

His1 virions package linear dsDNA genome terminated by imperfect inverted repeats (Table 2), which were suggested to be bound by as-yet-unidentified proteins (Bath *et al.*, 2006). Annotations of His1 genome revealed that it encodes type B DNA polymerase, which is thought to be implicated in protein-primed replication of the viral genome (Bath *et al.*, 2006). Apart from the polymerase, putative genes coding for an ATPase and a glycosyltransferase were annotated in His1 genome (Bath *et al.*, 2006).

4.1.5 Uncultured viral diversity

The genetic diversity of uncultured viruses was studied in Californian, Spanish, Tunisian and Senegalese hypersaline systems (Santos *et al.*, 2007, 2010; Dinsdale *et al.*, 2008; Sime-Ngando *et al.*, 2011; Boujelben *et al.*, 2012; Garcia-Heredia *et al.*, 2012). In some of these studies the metaviromes were obtained by direct sequencing of viral DNA retrieved from the environment (Dinsdale *et al.*, 2008; Santos *et al.*, 2010; Boujelben *et al.*, 2012). Other studies utilized cloning of environmental viral DNA into fosmids followed by sequencing in order to obtain more complete viral genomic sequences (Santos *et al.*, 2007; Garcia-Heredia *et al.*, 2012).

Common to all these studies, most of the metaviromic sequences lacked matches in current databases and therefore function could be predicted for the minority of the annotated genes. The predicted functions of the largest portion of genes were related to DNA metabolism (Santos *et al.*, 2007, 2010; Dinsdale *et al.*, 2008). Samples from San Diego salterns were also predicted to contain substantial amount of genes coding for virulence factors and carbohydrate metabolism (Dinsdale *et al.*, 2008). The 42 almost complete viral genomes reconstructed from the Santa-Pola solar saltern in Spain contained the genes coding for hallmark tailed virus proteins such as large terminase subunit, portal, capsid and tail sheath proteins suggesting that the obtained sequences belong to tailed viruses (Garcia-Heredia *et al.*, 2012).

Even though metavirome sequences from all sampled hypersaline environments had very few matches in databases, there was a considerable sequence overlap between different metaviromes (Boujelben *et al.*, 2012; Santos *et al.*, 2012). The overlap was bigger between the metaviromes originating from the environments of similar salinity (Boujelben *et al.*, 2012). Relatedness of metavirome sequences retrieved from geographically distant sources indicated that some of the viral communities were globally distributed (Boujelben *et al.*, 2012; Santos *et al.*, 2012).

4.2 Cold-active viruses

Cold environments of the deep sea, polar oceans, high-latitude lakes as well as sea ice constitute a substantial part of the aquatic ecosystems. These environments are populated by psychrophilic archaeal, bacterial and eukaryal microorganisms as well as their viruses (D'Amico *et al.*, 2006; Säwström *et al.*, 2008; Siddiqui *et al.*, 2013). Available metaviromic sequence data indicates that, compared to temperate aquatic ecosystems, cold environments

may harbour exceptionally high genotypic and taxonomic diversity of viruses (Lopez-Bueno *et al.*, 2009). However, studies on the genetic diversity of viruses from cold environments are extremely sparse.

Up to date, only a handful of cold-active viruses (viruses able to infect hosts at temperatures of 4^{o}C and below) have been cultured including tailed bacteriophages from cold marine waters, sea ice and nepheloid layer (particle-rich layer above the ocean floor) (Borriss *et al.*, 2003; Wells and Deming, 2006a; Wells, 2008; Luhtanen *et al.*, 2014).

4.2.1 Sea-ice bacteriophages

First report of virus isolation from the sea ice dates back to 2003. Then three phage-host systems were isolated from the samples of Arctic sea ice and melt ponds collected close to Svalbard. Myovirus 1a and siphoviruses 11b and 21c were shown to infect bacterial isolates belonging to genera *Shewanella*, *Flavobacterium* and *Colwellia*, respectively (Borriss *et al.*, 2003). Recently eight more phage-host systems originating from the Baltic Sea ice in the Finnish coastal area were described. The set included one phage infecting *Flavobacterium* sp. and seven *Shewanella* sp. phages (Luhtanen *et al.*, 2014).

Out of a total of 11 sea-ice phage isolates, the genomic sequence of only one was determined. The genome of *Flavobacterium* phage 11b (φ11b) was shown to be circularly permuted dsDNA molecule of 36,012 bp in size. Among the 65 annotated open reading frames (ORFs) several were predicted to encode proteins involved in DNA packaging and head morphogenesis including terminase small and large subunits, portal, minor head protein, protease and MCPs. All these proteins, except for the protease, were also detected by mass spectrometry as structural components of φ11b particles. Besides these structural proteins, φ11b was also predicted to encode a putative endolysin, essential recombination factor (*erf*), methylase and endonuclease. Majority of predicted and identified φ11b genes had similarities to ORFs of other phages and different species of *Bacteroides/Chlorobi*, *Firmicutes* as well as α-, β- and γ-proteobacteria. Phylogenetic trees based on the whole φ11b genome and separate φ11b proteins were incongruent, thereby demonstrating the mosaic nature of the viral genome (Borriss *et al.*, 2007).

4.2.3 Bacteriophage from nepheloid layer

Another cold-active bacteriophage, a siphophage 9A, was isolated from Arctic nepheloid layer. The phage infected *Colwellia psychrerythraea* and *Colwellia demingiae*, however, temperature and salinity ranges of plaque formation were different on the two hosts (Wells and Deming, 2006a).

A 104,936 bp-long 9A genome was predicted to encompass 149 ORFs. Similarly to φ11b, 9A was predicted to encode virion structure and assembly proteins, nucleic acid metabolism proteins as well as lysis enzymes. However, compared to φ11b, many more putative nucleic acid metabolism genes were annotated in 9A genome consistent with its larger size.

Many of the annotated 9A ORFs had homologues in γ-proteobacteria and their phages. Based on the amino acid composition, majority of the 9A proteins were suggested to be

non-psychrophilic relative to their homologues from databases. Strong psychrophilic characteristics expressed by the avoidance of charged residues and aromaticity were predicted for *RNaseHI*, DNA adenine methylase and transglycosylase leading to the speculation that these proteins play an important role in cold-adaptation of bacteriophage 9A.

Analogously to φ11b, 9A genome was also demonstrated to be mosaic by the incongruousness of phylogenetic trees constructed based on several of the viral proteins (Colangelo-Lillis and Deming, 2013).

B. AIMS OF STUDY

Viruses are astoundingly numerous and extremely important components of the biosphere (Suttle, 2007; Rohwer *et al.*, 2009). Sequencing of uncultured viral communities (metaviromes) indicated that, in addition to being numerous and ubiquitous, viruses harbor unprecedented genetic diversity (Breitbart *et al.*, 2002; Angly *et al.*, 2006). However, full potential of this data cannot be explored due to, among other reasons, the inability to relate the obtained sequences to the existing viral taxonomic groups. This results from the lack of cultured viral representatives with completely sequenced genomes. Moreover, the sequencing data obtained through metagenomic approach does not allow characterization and further manipulations of the organism it originates from. These points underline the great need and importance of the research on the genomics of the isolated viruses. While there are certain advancements made in this area (Krupovič *et al.*, 2011; Hatfull, 2014), not all environments have been sampled for viruses equally.

This work focused on the genomics of bacterial and archaeal viruses residing in sparsely sampled extreme aquatic environments – hypersaline systems and sea ice. The studied set of viruses included haloarchaeal tailed and pleomorphic viruses originating from different European and Asian hypersaline lakes and ponds (Atanasova *et al.*, 2012) and tailed bacteriophages isolated from the Baltic Sea ice (Luhtanen *et al.*, 2014).

The overall aim was to gain insights into the genomic diversity and dynamics of these viral groups. For that the complete genomes of the viral isolates were sequenced. In order to increase the studied dataset, proviral regions related to the sequenced viruses were identified in archaeal and bacterial genome sequences deposited in databases. The main aims of the study were:

1. To expand the known gene pool of the studied viral groups by annotating and analyzing the contents of their genomes.
2. To investigate the relatedness between the viruses within the three studied groups by performing comparative genomics analyses. These studies were also intended to answer additional individual questions for each of the viral groups.
    a. For haloarchaeal pleomorphic viruses, which make up a very recently described viral group (Pietilä *et al.*, 2009; Roine *et al.*, 2010), it was important to get an overview of the conserved versus variable genome regions and to develop a scheme to classify the viruses according to their genome contents and organization.
    b. For haloarchaeal tailed viruses the goal was to investigate whether the previously reported features shared with tailed bacteriophages (Krupovič *et al.*, 2011) are conserved among the studied viruses and to explore if there are any other common features, which have not been described before.
    c. Since prior to this study there was only one sea-ice phage genome sequenced (Borriss *et al.*, 2007), the comparative genomics analysis of the sequenced isolates provided a first glimpse into the genetic diversity of these viruses. In this case it was important to get an idea of the adaptation of the sea-ice phages in the environment by investigating if they share conserved gene modules and whether they are related to any other previously described viruses.

3. To investigate the physical nature of the viral genomes. In the course of the studies the genomes of two haloarchaeal pleomorphic viruses were revealed to have unusual characteristics. The aim was to study the physical nature of these genomes in more detail.

C. MATERIALS AND METHODS

The genomes of four haloarchaeal pleomorphic viruses, ten haloarchaeal tailed viruses and six tailed bacteriophages from the sea ice were analyzed in the study. The names, hosts and origins of the viruses are listed in Table 3. The employed laboratory methods and bioinformatics tools are presented in Tables 4 and 5, respectively.

Table 3. Viruses used in this study.

Virus	Host	Origin	Reference
Halorubrum sp. pleomorphic virus 2 (HRPV-2)	*Halorubrum* sp. SS5-4	SS	(Atanasova *et al.*, 2012)
Halorubrum sp. pleomorphic virus 3 (HRPV-3)	*Halorubrum* sp. SP3-3	SP	(Atanasova *et al.*, 2012)
Halorubrum sp. pleomorphic virus 6 (HRPV-6)	*Halorubrum* sp. SS7-4	SS	(Pietilä *et al.*, 2012)
Halogeometricum sp. pleomorphic virus 1 (HGPV-1)	*Halogeometricum* sp. CG-9	CG	(Atanasova *et al.*, 2012)
Haloarcula californiae head-tail virus 1 (HCTV-1)	*'Har. californiae'*	MdS	(Kukkaro and Bamford, 2009)
Haloarcula californiae head-tail virus 2 (HCTV-2)	*'Har. californiae'*	SS	(Atanasova *et al.*, 2012)
Haloarcula californiae head-tail virus 5 (HCTV-5)	*'Har. californiae'*	SS	(Atanasova *et al.*, 2012)
Halogranum head-tail virus 1 (HGTV-1)	*Halogranum* sp. SS5-1	SS	(Atanasova *et al.*, 2012)
Haloarcula hispanica head-tail virus 1 (HHTV-1)	*Har. hispanica*	MdS	(Kukkaro and Bamford, 2009)
Haloarcula hispanica head-tail virus 2 (HHTV-2)	*Har. hispanica*	SS	(Atanasova *et al.*, 2012)
Halorubrum head-tail virus 4 (HRTV-4)	*Halorubrum* sp. s5a-3	MdS	(Atanasova *et al.*, 2012)
Halorubrum head-tail virus 5 (HRTV-5)	*Halorubrum* sp. s5a-3	MdS	(Atanasova *et al.*, 2012)
Halorubrum head-tail virus 7 (HRTV-7)	*Halorubrum* sp. B2-2	MdS	(Atanasova *et al.*, 2012)
Halorubrum head-tail virus 8 (HRTV-8)	*Halorubrum* sp. B2-2	SS	(Atanasova *et al.*, 2012)
1/4	*Shewanella sp.*, 4	BSi	(Luhtanen *et al.*, 2014)
1/40	*Shewanella sp.*, 40	BSi	(Luhtanen *et al.*, 2014)
1/32	*Flavobacterium sp.*, 32	BSi	(Luhtanen *et al.*, 2014)
1/41	*Shewanella sp.*, 41	BSi	(Luhtanen *et al.*, 2014)
1/44	*Shewanella sp.*, 44	BSi	(Luhtanen *et al.*, 2014)
3/49	*Shewanella sp.*, 49	BSi	(Luhtanen *et al.*, 2014)

CG – Cabo de Gata, Spain; MdS – Margherita di Savoia, Italy; SP – Sedom Ponds, Israel; SS – Samut Sakhon, Thailand; BSi – Baltic Sea, Finland.

Table 4. Methods used in this study

Method	**Used in**
Plaque assay	I, II, III
Virus growth and purification	I, II, III
N-terminal sequencing and mass spectrometry of proteins*	I, III
Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)	I, III
Agarose gel electrophoresis (conventional, alkaline or 2D)	I, III
DNA extraction using phenol-chloroform method	I, II, III
RNA isolation	III
DNA enzymatic digestions	I, III
Polymerase chain reaction (PCR)	I, III
Reverse transcription PCR (RT-PCR)	III
Molecular cloning	I, III
Digoxigenin-11-dUTP (DIG-11-dUTP) end-labeling and detection	I
DNA sequencing*	I, II, III

***** The works utilizing these methods were either done exclusively by collaborators or ordered as a commercial service.

Table 5. Bioinformatic tools used in this study

Program	**Reference or web site of application (if not published)**	**Used in**
Genome assembly and annotation		
Phil's Read Editor (Phred)/ Phil's Revised Assembly Program (Phrap)/ Consensus Visualization and Editing Program (Consed)	(Gordon *et al.*, 1998)	II, III
DNAMaster	http://cobamide2.bio.pitt.edu/	I, II, III
pDRAW32	http://www.acaclone.com/	I
GeneMark.hmm	(Lukashin and Borodovsky, 1998)	I, II, III
Homology detection		
Basic Local Alignment Search Tool (BLAST)	(Altschul *et al.*, 1990)	I, II, III
Fold and Function Assignment System 03 (FFAS03)	(Jaroszewski *et al.*, 2005)	II
Homology Detection and Structure Prediction (HHpred)	(Söding *et al.*, 2005)	II, III
Protein Homology/analogY Recognition Engine V 2.0 (PHYRE2)	(Kelley and Sternberg, 2009)	II, III
Phamerator	(Cresawn *et al.*, 2011)	II
Sequence alignment		
T-Coffee	(Notredame *et al.*, 2000)	I, II, III
Multiple Sequence Comparison by Log-Expectation (Muscle)	(Edgar, 2004)	I
Profile Alignement (Praline)	(Bawono and Heringa, 2014)	I
Needle	https://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html	I, II, III
Phylogenetic analysis		
Gblocks	(Castresana, 2000)	I
Phylogenetic Estimation using Maximum Likelihood (phyml)	(Guindon *et al.*, 2010)	I
Molecular Evolutionary Genetics Analysis (MEGA)	(Tamura *et al.*, 2011)	III

Properties of the DNA sequence		
Tandem repeats finder	(Benson, 1999)	II, III
CRISPRFinder	(Grissa *et al.*, 2007)	II
Properties of the proteins		
Expert Protein Analysis System (EXPASY) tools	(Artimo *et al.*, 2012)	I
Search for protein motifs		
SignalP	(Nielsen *et al.*, 1997)	I
TatFind	(Rose *et al.*, 2002)	I
InterProScan	(Zdobnov and Apweiler, 2001)	I
Transmembrane Helices; Hidden Markov Model (TMHMM)	(Krogh *et al.*, 2001)	I
TMPred	(Hofmann and Stoffel, 1993)	I
Coils	(Lupas *et al.*, 1991)	I
Visualization software		
WebLogo	(Crooks *et al.*, 2004)	I, II
Phylodendron	http://iubio.bio.indiana.edu/treeapp/treeprint-form.html	I
Circos	(Krzywinski *et al.*, 2009)	III
Gepard	(Krumsiek *et al.*, 2007)	II

D. RESULTS AND DISCUSSION

1. Genomics of viral isolates from hypersaline environments

Haloarchaeal tailed and pleomorphic viruses are the most common isolates originating from hypersaline environments (Roine and Oksanen, 2011; Atanasova *et al.*, 2012; Sabet, 2012). Prior to this study complete genomic sequences of only four tailed (Klein *et al.*, 2002; Tang *et al.*, 2002, 2004; Pagaling *et al.*, 2007) and three pleomorphic viruses (Bath *et al.*, 2006; Pietilä *et al.*, 2009; Roine *et al.*, 2010) were available. Almost concomitantly with our study the genomes of three more haloarchaeal tailed viruses were published (Pietilä *et al.*, 2013b, 2013c). In addition, putative proviral regions related to the described haloarchaeal tailed and pleomorphic viruses were detected in the genomes of several haloarchaeal species (Pietilä *et al.*, 2009; Krupovič and Bamford, 2010; Roine *et al.*, 2010; Roine and Oksanen, 2011). Nevertheless, available sequences were not sufficient to get a comprehensive view of the genomic diversity and dynamics of these viral groups. Therefore, we sequenced and analyzed the genomes of ten more tailed and four more pleomorphic viruses isolated from samples taken during the recent global survey of hypersaline environments (Atanasova *et al.*, 2012; Pietilä *et al.*, 2012).

1.1 The genomes of haloarchaeal pleomorphic viruses

Up to date, seven haloarchaeal pleomorphic viruses have been isolated (Bath *et al.*, 2006; Pietilä *et al.*, 2009, 2012; Roine *et al.*, 2010; Atanasova *et al.*, 2012). The genomes of haloarchaeal pleomorphic viruses His2 (Bath *et al.*, 2006), HRPV-1 (Pietilä *et al.*, 2009) and HHPV-1 (Roine *et al.*, 2010) have been analyzed previously. We sequenced the genomes of the remaining four viruses: HRPV-2, HRPV-3 and HRPV-6 and HGPV-1 (**I**). The four sequenced genomes ranged from approximately 8,500 to 10,700 nt in size and were predicted to contain from 10 to 15 ORFs (Table 6). Some of the genes were identified by N-terminal sequencing of the encoded structural proteins (**I**) (Pietilä *et al.*, 2012).

Table 6. The genomes of haloarchaeal pleomorphic viruses

Virus	Genome type	Genome size (nt/bp)	Nr. of ORFs
HRPV-2	ssDNA, C	10,656	15
HRPV-3	dsDNA, C, D	8,771	12
HRPV-6	ssDNA, C	8,549	10
HGPV-1	dsDNA, C, D	9,694	15

C – circular; D – discontinuous (See section 1.1.1).

1.1.1 Physical nature of the genomes

Similarly to the previously described HRPV-1 (Pietilä *et al.*, 2009) and HHPV-1 (Roine *et al.*, 2010), the genomes of the four studied haloarchaeal pleomorphic viruses were shown to be circular DNA molecules. Based on the sensitivity to ssDNA-specific endonuclease, Mung Bean Nuclease (MBN), HRPV-2 and HRPV-6 genomes were shown to be single-stranded molecules (See Figure 2A in **I**). Digestion of HRPV-3 and HGPV-1 genomes with MBN yielded a number of resistant fragments (See Figure 2A in **I**) suggesting that the genomic molecules of these viruses consist of both single-stranded and double-stranded regions (**I**).

Putative ssDNA regions in HRPV-3 and HGPV-1 genomes were further studied by the MBN analysis of the genomic molecules repaired using *Sulfolobus* DNA polymerase IV in combination with a DNA ligase or using the DNA ligase alone. Since only the polymerase in combination with the ligase and not the ligase alone were able to yield both genomes uniformly resistant to MBN (See Figure 2B in **I**), it was confirmed that HRPV-3 and HGPV-1 have partially dsDNA genomes interspersed by ssDNA regions (**I**).

In order to determine the positions of the ssDNA regions in the viral genomes two different approaches were utilized. First, restriction digestion analysis of the DIG-11-dUTP end-labeled genomic molecules was employed to study the approximate location of the free 3'-ends of DNA, which should be present at the junctions between dsDNA and ssDNA parts of the genomes. The label was detected in all fragments resulting from the digestion analyses suggesting that the single-stranded regions are distributed roughly throughout the genomes of HRPV-3 and HGPV-1. Another approach used to elucidate more precise locations of the single-stranded regions involved sequencing of the genomic fragments resistant to MBN digestion. The obtained sequences were overlapping, thereby, suggesting that there is a degree of heterogeneity in the distribution of the ssDNA regions in both of the viral genomes. However, alignment of the sequenced MBN-resistant genomic fragments of HRPV-3 showed that the majority of the fragments terminated with the conserved DNA motif "GCCCA" (Figure 3A). This result suggested that the switches from the double-stranded to single-stranded regions in HRPV-3 genome were associated with the motif. In order to corroborate this assumption, the genome sites containing "GCCCA" were sequenced using Sanger method. HRPV-3 genome has a total of 27 such sites (Figure 3C). Majority of the reactions displayed an abrupt drop of sequencing signal following the "GCCCA" motif (Figure 3B, C). The drop can be explained by the absence of the template resulting from the switch of dsDNA to ssDNA region in the HRPV-3 genome. This result confirmed that the single-stranded regions in HRPV-3 genome were indeed associated with the "GCCCA" motif and showed that the discontinuities are present in the strand opposite from the one containing the motif. Discontinuities in HGPV-1 genome were not found to be associated with any DNA motif (**I**).

Collectively, the results indicated that HRPV-3 and HGPV-1 have dsDNA genomes with single-stranded regions, which are localized to specific positions in HRPV-3 genome. This is the first report of this genome type found among the viruses of prokaryotes (**I**). Earlier studies showed that "phiKMV-like" and T5 bacteriophages also have discontinuities in their dsDNA genomes (Wang *et al.*, 2005; Kulakov *et al.*, 2009). However, in both cases, discontinuities were nicks and not longer ssDNA regions, which seems to be the case for HRPV-3 and HGPV-1 (**I**).

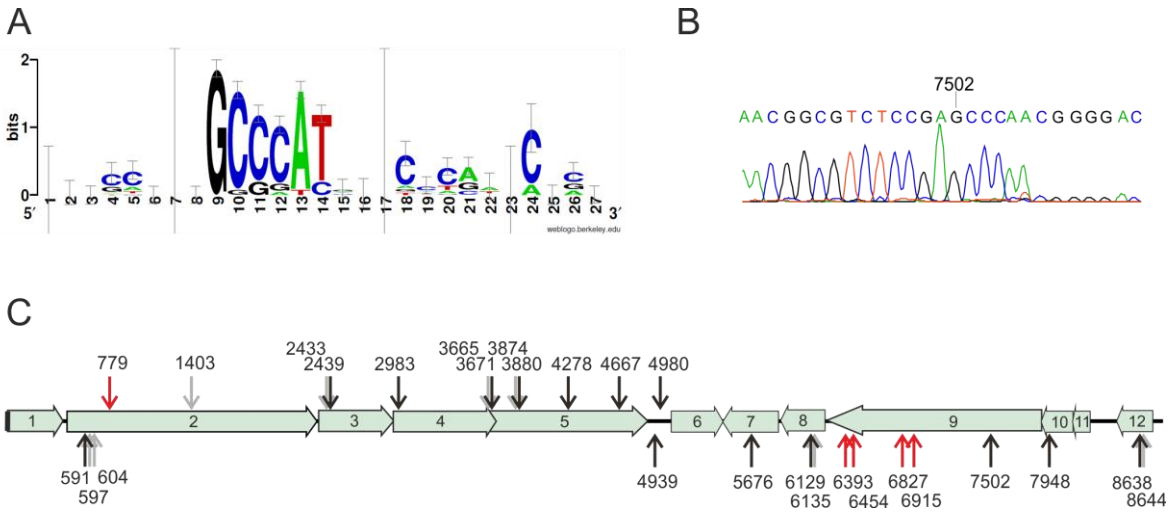

Figure 3. Mapping of the switches from double-stranded to single-stranded regions in HRPV-3 genome. (**A**) Conserved DNA motif at the termini of MBN-resistant genomic fragments of HRPV-3. (**B**) Sanger sequencing of HRPV-3 genome region containing "GCCCA" motif. Nucleotide coordinates are indicated above the graph. (**C**) The genome map of HRPV-3 with locations of "GCCCA" motifs indicated. Positions of "GCCCA" motifs are marked with black, red and grey arrows. Nucleotide coordinates are indicated above the arrows. Black arrows indicate that there was a sudden drop of signal observed during Sanger sequencing of these genome positions. Red arrows mark locations where no such drop was observed, whereas grey arrows indicate the places which could not be sequenced.

1.1.2 Cluster of conserved genes

Among the described haloarchaeal pleomorphic viruses only HRPV-2 and HRPV-6 display high nucleotide sequence identity (more than 90%) over longer genomic regions (up to 4000 nt). However, a collinear cluster of five genes is shared among all haloarchaeal pleomorphic viruses with an exception of His2 (Bath *et al.*, 2006) (Figure 4). This cluster codes for two major structural proteins, two hypothetical proteins with predicted transmembrane domains and a putative ATPase. Based on the similarity to HRPV-1 (Pietilä *et al.*, 2009), these proteins were termed VP3-like and VP4-like proteins, ORF6-like and ORF7-like products and VP8-like proteins, respectively (**I**). VP3-like and VP4-like are major structural proteins. In HRPV-1 VP3 protein was shown to reside mainly in the membrane with the small portion facing the virion inside, whereas VP4 was shown to form club-shaped spikes on the virion surface suggested to have a role in host recognition (Pietilä *et al.*, 2012). Shared major structural proteins reflect similar virion architectural principles of haloarchaeal pleomorphic viruses (Pietilä *et al.*, 2012). The most diverged member of the haloarchaeal pleomorphic virus group, His2 (Bath *et al.*, 2006), has homologues of VP4-like protein, ORF6-like and ORF7-like products as well as VP8-like protein, but not VP3-like protein. His2 and HGPV-1 share one more small major structural protein (VP3 in HGPV-1 and VP27 in His2, Figure 4) indicating that there may be additional subtle structural similarities between these two viruses (**I**) (Pietilä *et al.*, 2012).

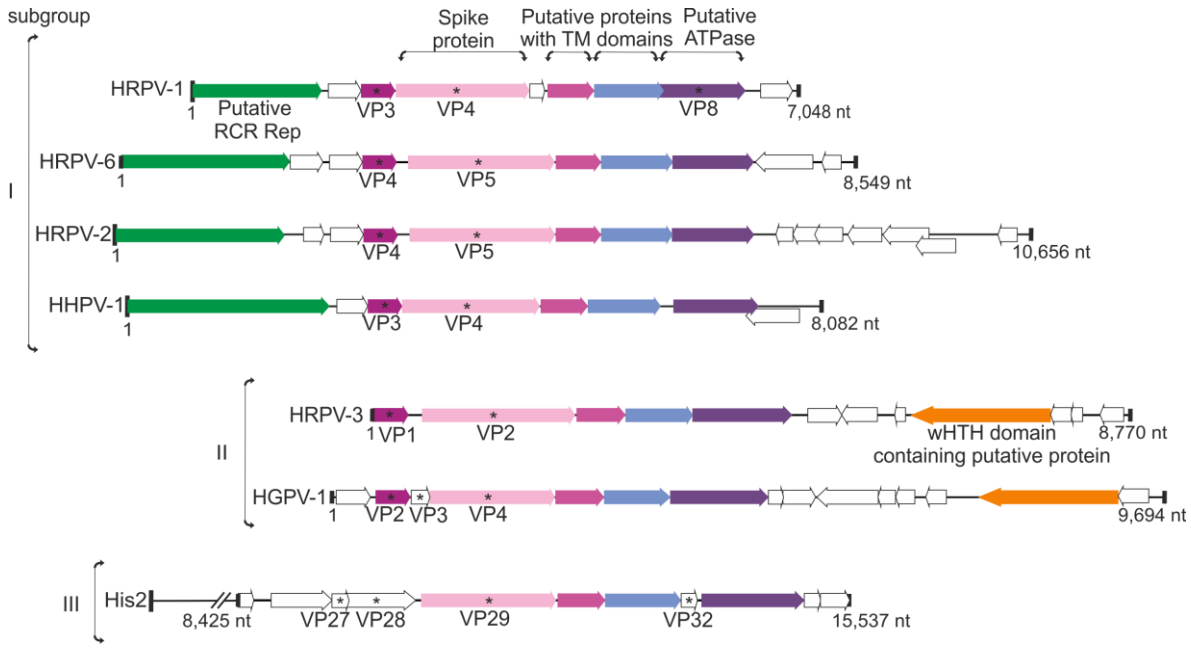

Figure 4. The genome maps of haloarchaeal pleomorphic viruses. ORFs encoding putative homologues have the same color. Genes marked with asterisks code for the major structural proteins, which were identified by N-terminal sequencing. The annotations of putative and identified proteins belonging to the conserved cluster are indicated on the top of the figure. Designation of the viruses into subgroups based on the genome contents are indicated on the left of the figure. Modified from (Pietilä *et al.*, 2009, 2012; Roine *et al.*, 2010).

Judged by the amino acid identity VP3-like and VP8-like proteins as well as ORF7-like gene product are the most conserved proteins shared among haloarchaeal pleomorphic viruses. Phylogenetic trees reconstructed based on these proteins were congruent suggesting that the conserved cluster of genes is inherited vertically in pleomorphic viruses (See Figures S3B, S4B, S5B in I) (**I**). Vertical inheritance indicates that these genes may constitute the "viral self" (Bamford, 2003) of the haloarchaeal pleomorphic viruses.

1.1.3 Genome-based classification of haloarchaeal pleomorphic viruses

Based on the genome organization and shared homologues the group of the four studied (HRPV-2, HRPV-3, HRPV-6 and HGPV-1) and three earlier reported (HRPV-1, HHPV-1 and His2) (Bath *et al.*, 2006; Pietilä *et al.*, 2009; Roine *et al.*, 2010) haloarchaeal pleomorphic viruses was divided into three subgroups. Viruses belonging to the first subgroup have circular ssDNA (HRPV-1, HRPV-2, HRPV-6) or dsDNA (HHPV-1) genomes and, in addition to the conserved cluster, share genes coding for RCR-Rep (**I**) (Pietilä *et al.*, 2009; Roine *et al.*, 2010). HRPV-3 and HGPV-1 having discontinuous dsDNA genomes comprise the second subgroup. No gene coding for replication-related protein could be predicted in these genomes. However, both viruses share a homologue

with the putative role in DNA binding suggested by the predicted winged helix-turn-helix domain (wHTH). The third subgroup contains a single member, His2 virus (**I**) (Bath *et al.*, 2006). Unlike all other haloarchaeal pleomorphic viruses (Pietilä *et al.*, 2009; Roine *et al.*, 2010), His2 has a linear dsDNA genome almost twice the size of the other haloarchaeal pleomorphic virus genomes (Bath *et al.*, 2006). His2 encodes a putative type B polymerase and was suggested to replicate its genome via a protein-primed mechanism (Porter and Dyall-Smith, 2008).

Therefore, while sharing common virion architecture, haloarchaeal pleomorphic viruses appear to be less stringent in preserving the module of genes coding for functions related to the genome replication. This is not the only example demonstrating that the evolution of the module coding for virion structure can be uncoupled from the evolution of the module coding for replication-related proteins. For example, tailed viruses are also known to employ different genome replication strategies despite the shared virion architecture (Weigel and Seitz, 2006). However, all tailed viruses have dsDNA genomes, whereas haloarchaeal pleomorphic viruses possess a variety of genome types including linear and circular, single-stranded and double-stranded DNA molecules. Such flexibility in choosing not only the genome replication strategy, but also the genome type to be packaged into the virions is remarkable and has not been observed in other groups of viruses described to date.

1.1.4 Related proviral regions

Haloferax lucentense plasmid pHK2 and proviral regions in numerous haloarchaeal species were reported to show similarities to the genomes of haloarchaeal pleomorphic viruses (Pietilä *et al.*, 2009; Roine *et al.*, 2010; Roine and Oksanen, 2011). In this study we identified three more related proviral regions in *Haloarcula hispanica* ATCC 33960 and *Halopiger xanaduensis* SH-6 genomes (**I**). Overall, a total of 15 putative proviruses and proviral remnants were discovered (**I**) (Pietilä *et al.*, 2009; Roine *et al.*, 2010; Roine and Oksanen, 2011). The gene organization of 11 of these proviral regions and plasmid pHK2 was studied (See Table 3 in **I**). According to the organization and shared homologues, two proviral regions and pHK2 plasmid could be classified into the first subgroup of haloarchaeal pleomorphic viruses with the remaining nine proviral regions falling into the second subgroup. The fact that proviral regions adhered to the genome organization of delineated subgroups supports the proposed classification of haloarchaeal pleomorphic viruses and strengthens the assumption that the genome organization is conserved among these viruses (**I**).

1.2 The genomes of haloarchaeal tailed viruses

We sequenced and analyzed the genomes of six siphoviruses (HCTV-1, HCTV-2, HCTV-5, HHTV-1, HHTV-2 and HRTV-4) and four myoviruses (HGTV-1, HRTV-5, HRTV-7 and HRTV-8) infecting different haloarchaeal species including *Haloarcula hispanica*, *Haloarcula californiae*, *Halogranum* sp. and *Halorubrum* sp. (**II**) (Kukkaro and Bamford, 2009; Atanasova *et al.*, 2012). The genomes are linear dsDNA molecules ranging in size

from approximately 35 kb to 144 kb and terminating with either direct terminal repeats or circular permutation (Table 7) (**II**).

Two of the previously sequenced haloarchaeal tailed viruses, HF1 and HF2 (Tang *et al.*, 2002, 2004), were shown to be closely related at the nucleotide sequence level to viruses HRTV-5 and HRTV-8 from the studied dataset. Therefore, HF1 and HF2 were also included in the following analyses (**II**).

Table 7. The genomes of haloarchaeal tailed viruses.

Virus	Type of genome termini	Genome size (bp)	Nr. of ORFs	Nr. of tRNAs
HCTV-1	739 bp DTR	103,257	160	1
HCTV-2	Circ perm	54,291	86	0
HCTV-5	583 bp DTR	102,105	166	1
HGTV-1	Circ perm	143,855	281	36
HHTV-1	Circ perm	49,107	74	0
HHTV-2	Circ perm	52,643	88	0
HRTV-4	Circ perm	35,722	73	0
HRTV-5	271 bp DTR	76,134	118	4
HRTV-7	340 bp DTR	69,048	105	1
HRTV-8	346 bp DTR	74,519	124	4

DTR – direct terminal repeats; Circ perm – circularly permuted.

1.2.1 Genome annotations and protein phamilies

A total of 1,275 ORFs, 47 tRNAs and 3 ribozymes were annotated in close to 760 kb of the new genomic sequence of 10 haloarchaeal tailed viruses. However, function could be predicted for approximately 20% of all annotated ORFs based on similarities to ORFs in tailed bacteriophages. Most of the proteins with predicted function were either tailed virion structural and assembly proteins or enzymes involved in nucleic acid metabolism (**II**). This situation is reminiscent of what was observed in the case of few other described archaeal tailed viruses and proviral regions (Klein *et al.*, 2002; Tang *et al.*, 2002, 2004; Pagaling *et al.*, 2007; Krupovič and Bamford, 2010; Pietilä *et al.*, 2013b, 2013c).

A total of 1491 ORFs obtained by adding HF1 and HF2 annotations to the dataset were analyzed by the Phamerator (**II**). Phamerator (Cresawn *et al.*, 2011) is a tool performing pairwise alignments and grouping genes coding for related proteins into the so called "phamilies". The analyzed 1491 ORFs were assembled into 966 phamilies, of which 726 were composed of single members, i.e. genes coding for proteins having no significant similarity to other proteins in the analyzed dataset. This result corroborates high diversity harbored in the 12 studied genomes. Some of the largest phamilies having seven to eight members contained genes coding for DNA polymerase elongation subunits, RtcB-like proteins as well as key enzymes involved in deoxyribonucleotide synthesis: ribonucleotide reductases, thymidylate synthases and dUTPases. Wide distribution of these genes among the haloarchaeal tailed viruses argues that their products play an important role in the life cycles of the viruses. High similarity shared among these proteins suggests that their encoding genes may be frequently exchanged among the viruses (**II**). Phage-to-host and

host-mediated phage-to-phage transfer of genes encoding enzymes involved in deoxyribonucleotide synthesis has been demonstrated previously by phylogenetic analyses of these genes from marine phages and environmental metaviromes (Huang *et al.*, 2012; Dwivedi *et al.*, 2013). Our observation supports this idea and suggests that it may be also extended to archaeal viruses.

1.2.2 Common themes with tailed bacteriophages

In agreement with other studies on the genomes of haloarchaeal tailed viruses and related proviral regions (Klein *et al.*, 2002; Tang *et al.*, 2002, 2004; Pagaling *et al.*, 2007; Krupovič and Bamford, 2010; Pietilä *et al.*, 2013b, 2013c), the genomes described in this study (**II**) are similar in organization to the genomes of tailed bacteriophages (Krupovič *et al.*, 2011; Hatfull, 2012b). Genes coding for functional analogues of virion structural and assembly components are arranged in a conserved order, according to which, block of genes coding for DNA packaging machinery is followed by genes encoding head and then tail assembly and structural proteins (**II**).

As is the case for previously described archaeal and bacterial tailed viruses (Krupovič *et al.*, 2011), the genomes of the studied haloarchaeal tailed viruses are mosaic structures (**II**). In one of the viruses, HGTV-1, a specific process responsible for the gene shuffling and consequent genomic mosaicism was suggested. Separate ORFs in the right-hand side of HGTV-1 genome were flanked by 50 bp-long repeat sequences containing two putative elements – TATA box-like sequence and inverted repeats (See Figure 6 in **II**). Similar intergenic repeats, termed promoter stem loop structures (PesLSs), were reported previously in T4-like bacteriophages (Arbiol *et al.*, 2010). PesLSs contain σ^{70} promoter sequence and inverted repeats. Besides the proposed role in transcription regulation, PesLSs were suggested to mediate gene shuffling in T4-like bacteriophages. Indeed, obtained experimental evidence showed that the recombination between PesLSs led to the excision of the flanked gene and formation of mini-circles containing this gene, which may be subsequently packaged into virus particles and delivered to the next host (Arbiol *et al.*, 2010). Structural similarity of HGTV-1 repeats and T4-like PesLSs suggests that they may serve similar roles in both viruses (**II**).

1.2.3 Comparative genomics

Different levels of relatedness were observed between the 12 analyzed haloarchaeal tailed viruses (ten viruses from this study, HF1 and HF2) (**II**). Some of the viruses were closely related showing almost uninterrupted alignment along the whole genome (Figure 5). Two clusters of such closely related viruses were delineated. One of the clusters encompassed myoviruses HRTV-5, HRTV-8, HF1 and HF2. The other cluster included siphoviruses HCTV-1 and HCTV-5. These two viruses had a weaker similarity than the viruses from the first cluster. First of all, right-hand parts of the genomes, which code for tail structural and assembly proteins, are rather diverged as is evident from the gaps in the genome alignment (Figure 5). Second, closer inspection of the genome contents showed that compared to

HCTV-1, HCTV-5 was heavily invaded by homing endonucleases interrupting the alignment of these two genomes (see Figure 4 in **II**) (**II**).

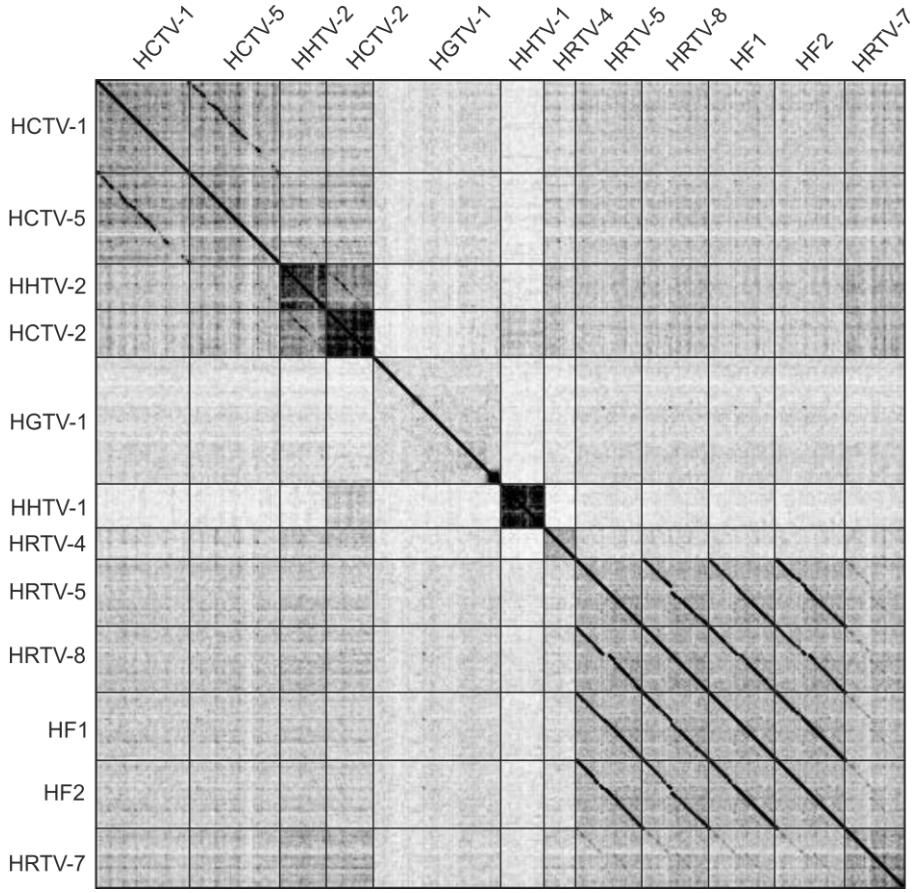

Figure 5. Dotplot alignment of haloarchaeal tailed virus genomes. Alignments of concatenated genomic sequences of the studied viruses as well as HF1 and HF2 were done using Gepard software (Krumsiek *et al.*, 2007).

Notably, closely related viruses from both clusters were isolated from geographically distant locations arguing for the dispersal of these viral types across large distances. For instance, HRTV-5, HRTV-8, HF1 and HF2 were isolated from different samples taken in Italy, Thailand and Australia in a timespan of almost 20 years (Nuttall and Dyall-Smith, 1993; Atanasova *et al.*, 2012). This is consistent with the idea of the global virus distribution suggested by the studies both on isolated viruses (Snyder *et al.*, 2007; Atanasova *et al.*, 2012) and on environmental metaviromes (Angly *et al.*, 2006; Schoenfeld *et al.*, 2008; Sime-Ngando *et al.*, 2011).

Some of the other studied viruses were more distantly related, which was apparent from the few patches of sequence similarity seen in the genome alignments (Figure 5). The example of such relatedness is a pair of siphoviruses HHTV-2 and HCTV-2. Also a myovirus HRTV-7 shares some similarities with HRTV-5, HRTV-8, HF1 and HF2 viruses. Majority of the studied viruses, however, showed no similarities extending beyond very short genomic regions (**II**). This emphasizes the necessity of the further sampling of haloarchaeal tailed virus genetic diversity.

2. Genomics of viral isolates from the Baltic Sea ice

Close to 40% of Baltic Sea surface area is covered by the seasonal ice for a time period of up to six months annually (Granskog *et al.*, 2006). However, the first and, up to date, the only report on bacteriophage isolates from the Baltic Sea ice was published only in 2014 (Luhtanen *et al.*, 2014). Luhtanen and others isolated one siphovirus infecting *Flavobacterium* sp. as well as five myoviruses and two siphoviruses infecting *Shewanella* sp. hosts from the coastal ice samples collected approximately two and half months after the freeze-up (Luhtanen *et al.*, 2014). In this study the genomes of six of the phages, i.e. myoviruses 1/4, 1/40, 1/41 and siphoviruses 1/32, 1/44, 3/49, were analyzed (**III**).

Table 8. The genomes of bacteriophages from the Baltic Sea ice.

Virus	Type of genome termini	Genome size (bp)	Nr. of ORFs	Nr. of tRNAs
1/4	381 bp DTR	133,824	235	3
1/40	381 bp DTR	139,004	236	3
1/32	Circ perm, pac at ~ 3.1 kb	42,252	63	-
1/41	Circ perm, pac at ~7.7 kb	43,510	69	-
1/44	Circ perm, pac at ~4 kb	49,640	75	-
3/49	Circ perm, pac at ~5.6 kb	40,161	70	-

DTR – direct terminal repeats; Circ perm – circularly permuted.

The determined genomic sequences of the sea-ice bacteriophages range from approximately 40 to 140 kb in size (Table 8). The genomes of 1/4 and 1/40 are linear dsDNA molecules with direct terminal repeats (DTRs). The other four phages have circularly permuted dsDNA genomes packaged from the pac sites. The pac sites of 1/32, 1/41 and 3/49 were estimated to reside within or close to the coding region of terminase small subunit gene. Phage 1/44 was suggested to contain pac site in the non-coding region densely populated by short tandem repeats (**III**).

2.1 Genome annotations

The genomes of the sea-ice bacteriophages were predicted to encode DNA-packaging machinery components as well as virion structural and assembly proteins reminiscent of those found in other tailed bacteriophages. In many cases genes coding for virion structural proteins were highly divergent and could not be annotated based on similarities to previously described viral genes. Therefore, some of the annotations were done based on the N-terminal sequencing and/or liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) analyses of virion constituent proteins. In total, from 11 up to 26 genes coding for structural proteins were annotated in the sea-ice phage genomes (**III**). The order of structural and assembly genes was conserved and similar to the organization observed previously in tailed viruses (Krupovič *et al.*, 2011; Hatfull, 2012b), i.e. genes encoding small and large terminase subunits, portal, minor and major capsid components, tail assembly chaperones, tape measure and other structural tail proteins were arranged in a sequential order (**III**). However, 1/4 and 1/40 genomes displayed one unusual feature. In

these viruses genome regions coding for large terminase subunits (*terL*) were predicted to contain introns and homing endonuclease genes. Analysis of *terL*-targeted RT-PCR product showed that at least in phage 1/4 mRNA is processed producing an uninterrupted transcript of *terL* gene (See Figure S10 in **III**). To our knowledge, this is the first report of tailed bacteriophage *terL* genes accommodating introns and HNH homing endonuclease genes (**III**). However, the presence of introns in a number of other phage genes, majority of which have a function linked to DNA metabolism, has been reported previously (Edgell *et al.*, 2000).

Besides virion structural and assembly proteins, sea-ice phages were predicted to encode a number of proteins with a putative role in DNA replication, recombination and repair processes (**III**). These genes are commonly annotated in other tailed bacteriophages too (Krupovič *et al.*, 2011; Hatfull, 2012b).

2.2 Comparative genomics

Of the six studied sea-ice bacteriophages only 1/4 and 1/40 were shown to be closely related having approximately 70% of their annotated proteins homologous (See Figure 4 in **III**). These two viruses were also shown to be related to a number of *Vibrio*-specific phages including ICP1 (Seed *et al.*, 2011), helene 12B3 and PWH3a-P1, which were chosen as representatives for the comparative genomics analyses (**III**). MCPs of the sea-ice phages shared over 50% amino acid identity with the MCPs of *Vibrio* phages arguing for the recent common ancestry of these two viral groups. Overall, approximately 25% of phage 1/4 and 1/40 proteins have homologues in at least one of the three *Vibrio* phages. Genes coding for these shared homologues are not clustered suggesting that HGT was a significant force shaping these genomes (See Figure 5A in **III**).

Several putative prophages related to siphophage 1/44 were detected in the genomes of *Shewanella baltica*, *Shewanella frigidimarina* and *Shewanella denitrificans* (See Figure 5B in **III**). Shared homologues are encoded in a collinear cluster of genes and include minor capsid protein, major tail protein, tape measure protein, baseplate protein as well as five hypothetical proteins (**III**).

Apart from the above mentioned examples, the studied sea-ice bacteriophages are rather diverged from each other and any tailed bacteriophage with sequenced genome known to date. Based on the ICTV bacteriophage genus definition, which states that the phages within the same genus have at least 40% of the proteins homologous, the studied sea-ice phages can be classified into five new genera: four genera with single members (1/41, 1/32, 1/44 and 3/49) and one genus encompassing two phages, 1/4 and 1/40 (**III**).

E. CONCLUDING REMARKS

This study focused on the genomics of three viral groups, the members of which were underrepresented among the viruses with sequenced genomes. The obtained genomic sequences of haloarchaeal pleomorphic and tailed viruses more than doubled the previously available genomic information on these viral groups. The genomic sequences of the Baltic Sea ice bacteriophages provided first insight into the genome-level variation among the bacterial viruses in the sea ice, since prior to this study only a single sea-ice bacteriophage has been sequenced (Borriss *et al.*, 2007).

The studied viruses are rather diverged from the previously described ones as the majority of the genes annotated in the viral genomes did not have significant matches in the current databases. As a consequence, function could not be assigned for the largest portion of the predicted genes. In the case of haloarchaeal pleomorphic viruses, besides the determined virion structural components, function was predicted for only two encoded putative proteins, an ATPase and RCR-Rep. The genomes of tailed viruses isolated from hypersaline environments and sea ice were annotated to contain a number of genes coding for virion structural and assembly as well as nucleic acid metabolism proteins. Even though the content of nucleic acid metabolism genes was different among the viruses, the fact that the genes coding for deoxyribonucleotide synthesis and replication functions were present in many of the genomes corroborates their importance in the life cycles of tailed viruses regardless of the environment they reside or the hosts they infect. In agreement with the previous studies (Klein *et al.*, 2002; Tang *et al.*, 2002, 2004; Borriss *et al.*, 2007; Pagaling *et al.*, 2007; Pietilä *et al.*, 2013b, 2013c), the virion structural and assembly components were similar and, therefore, presumably functionally analogous to the proteins in previously described tailed viruses. In addition to the functional analogues shared with bacteriophages, haloarchaeal tailed virus HGTV-1 was suggested to have a genome shuffling mechanism reminiscent of that in T4-like phages (Arbiol *et al.*, 2010). Further studies may uncover more common lines between bacterial and archaeal tailed viruses.

Despite the general divergence of the studied viruses, closely related representatives were detected within each of the studied viral groups. Comparison of the related viruses showed that the shared similarity is patchy suggesting mosaic nature of the studied genomes, which goes in accordance with the previous studies (Krupovič *et al.*, 2011). Comparative genomics analyses of haloarchaeal pleomorphic viruses showed that they have a conserved cluster of genes, which was suggested to be inherited vertically in this group of viruses. Based on the genome organization beyond the conserved cluster, viruses could be subdivided into three subgroups, which also reflected different genome types of the viruses. One of the subgroups containing HRPV-3 and HGPV-1 viruses was distinguished by the unusual type of the genomes, which were dsDNA molecules interspersed by localized short ssDNA regions. This genome type has not been previously found in other viruses of prokaryotes.

Even though the studied dataset was too small for drawing conclusions, it gave some hints on the distribution of viral groups in the environment. Haloarchaeal tailed and pleomorphic viruses were isolated from a variety of geographic sites (Atanasova *et al.*, 2012; Sabet, 2012). Based on the ability of viruses to cross-infect the hosts originating from different locations, it was concluded that hypersaline environments function as a single global system (Atanasova *et al.*, 2012). If it was so, then the related viruses would be

found in geographically distant locations. Comparative genomics analyses of both haloarchaeal pleomorphic and tailed viruses confirmed that this was indeed the case. The isolation and genomic data available for the sea-ice phages is not sufficient for this type of analyses since the studied phages were isolated from a single site and only few related viruses and proviral regions were found in the databases.

To conclude, this study provided an overview of the genomic diversity of the three understudied viral groups – haloarchaeal pleomorphic and tailed viruses and sea-ice tailed bacteriophages. The obtained information gave some insights into the contents and nature of the genomes as well as genomic dynamics of the studied viral groups. This study provides one of the initial steps towards understanding of the role and function of these viruses in the environment in the light of their genomic diversity.

The obtained genomic data serves as a good starting point for further experimental work. For instance, studies sought to elucidate the cause and the role of the unusual genome types found in the two haloarchaeal pleomorphic viruses would be of particular interest. As the largest portion of the annotated genes in all three studied viral groups has no predicted function, the research focusing on the functional annotations of these genes is in great demand.

F. ACKNOWLEDGEMENTS

This work was performed under the Programme on Molecular Virology, at the Institute of Biotechnology and Department of Biosciences, Faculty of Biological and Environmental Sciences, University of Helsinki, under the supervision of Doc. Elina Roine.

I owe my deepest gratitude to my supervisor, Elina. I am very grateful to Elina for providing me with exciting projects and interesting topics to study. I have greatly benefited from her guidance and I am very happy that I had a thesis supervisor, who is an extremely enthusiastic and committed scientist as well as kind and caring person. Thank you, Elina, for always having time for me, for your patience and support throughout my PhD work.

I am sincerely grateful to Prof. Dennis Bamford for giving me an opportunity to work in the DB-lab starting from my Bachelor degree studies and on. It has been such a perfectly organized and excellent environment to work in! Thank you, Dennis, for your care and insightful advice you provided throughout these years.

I would like to thank Doc. Petri Auvinen and Prof. Mikael Skurnik for critically reviewing my thesis and giving constructive feedback to it. I thank the members of my thesis follow-up group Doc. Petri Auvinen and Doc. Päivi Onkamo for giving useful advice concerning my studies and research work.

I would like to acknowledge Viikki Graduate School in Biosciences and Doctoral Program in Mirobiology and Biotechnology for supporting my thesis work and participation in conferences and internship abroad as well as organizing useful courses and nice social events. Academy of Finland (grants 271413 and 272853) and University of Helsinki are thanked for the support to EU ESFRI Instruct Centre for Virus Production used in this study. The University of Helsinki three year grant 2010–2012 (to E.R.) and Academy of Finland Centre of Excellence Program in Virus Research grant 11296841 (to D.B) (2006–2011) are thanked for funding this study.

I am profoundly indebted to all my collaborators and co-authors, without whom this work would not have been possible. I am particularly grateful to Prof. Roger Hendrix and Deborah Jacobs-Sera for hosting me in Pittsburgh as well as teaching me how to use the DNAMaster and the Phamerator. I thank Dr. Nina Atanasova and Dr. Hanna Oksanen for providing me with interesting viruses to study as well as for valuable discussions and advice. Anne-Mari Luhtanen and Mikko Saarijärvi are thanked for their input into the studies on the sea-ice phages.

I would like to thank all former and present members of the Programme on Molecular Virology. It is a great pleasure to work among nice and helpful people. I thank Doc. Janne Ravantti for his help with computer-related questions and for making every day seem brighter with his optimistic contemplations about life. I am sincerely grateful to a very warm and kind person, who also happens to be my very good friend, Dr. Nina Atanasova, for all the great time we have had together. I thank Dr. Alesia Romanovskaya, Heli Mönttinen, Dr. Gabija Žiedaitė, Dr. Mart Krupovič, Dr. Virginija Cvirkaitė-Krupovič, Dr.

Xiaoyu Sun, Tatiana Demina, Salla Jaakkola, Julija Svirskaitė and, of course, my dear office neighbor, Outi Lyytinen, for your nice company both in the lab and during the coffee breaks. Alesia, thank you for inviting me to a knowledge game and a number of wonderful dinners. Heli, thank you for our common outdoor activities.

I am very grateful to our skilled technical personnel for excellent assistance. I would like to especially thank Sari Korhonen for the numerous virus purifications and Riitta Tarkiainen for her help with DNA and RNA work.

I would like to thank the members of The Cyanobacteria Group, where I have also done some work during my PhD studies. Thank you for your hospitality and willingness to help.

I would like to express my sincere gratitude to Prof. Rimantas Daugelavičius for being a very patient and kind mentor during my Bachelor degree studies as well as for introducing me to the DB-lab and for supporting my further study decisions.

I thank my much loved friends and family for bringing lots of joy into my life and for being there for me during difficult times. Mantas, thank you for being a very attentive and good friend throughout our stay in Helsinki. Kristina, thank you for being by my side, listening, advising and not judging. Nastia and Janne: although we do not see and talk to each other that often, I know that I have you and this is important.

I have so much to be grateful for to my beloved Martin. Most of all, thank you for making me happy. Also, thank you for providing me with a firm shoulder and invaluable academic help, even when you were extremely busy with your own work. And finally, there is no good enough words in the world to express my gratitude to dear parents, Galina and Viktor, grandmother Liubov and brother Michail. Thank you for your unconditional love and everlasting support.

Ana
Helsinki, October 2014

G. REFERENCES

Abrescia, N.G., Bamford, D.H., Grimes, J.M., and Stuart, D.I. (2012). Structure unifies the viral universe. Annu Rev Biochem *81*, 795-822.

Ackermann, H.W., and Prangishvili, D. (2012). Prokaryote viruses studied by electron microscopy. Arch Virol *157*, 1843-1849.

Ahn, D.G., Kim, S.I., Rhee, J.K., Kim, K.P., Pan, J.G., and Oh, J.W. (2006). TTSV1, a new virus-like particle isolated from the hyperthermophilic crenarchaeote *Thermoproteus tenax*. Virology *351*, 280-290.

Al-Attar, S., Westra, E.R., van der Oost, J., and Brouns, S.J. (2011). Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biol Chem *392*, 277-289.

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.

Anesio, A.M., and Bellas, C.M. (2011). Are low temperature habitats hot spots of microbial evolution driven by viruses? Trends Microbiol *19*, 52-57.

Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., *et al.* (2006). The marine viromes of four oceanic regions. PLoS Biol *4*, e368.

Arbiol, C., Comeau, A.M., Kutateladze, M., Adamia, R., and Krisch, H.M. (2010). Mobile regulatory cassettes mediate modular shuffling in T4-type phage genomes. Genome biology and evolution *2*, 140-152.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., *et al.* (2012). ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res *40*, W597-603.

Atanasova, N.S., Roine, E., Oren, A., Bamford, D.H., and Oksanen, H.M. (2012). Global network of specific virus-host interactions in hypersaline environments. Environ Microbiol *14*, 426-440.

Bamford, D.H. (2003). Do viruses form lineages across different domains of life? Res Microbiol *154*, 231-236.

Bamford, D.H., Burnett, R.M., and Stuart, D.I. (2002). Evolution of viral structure. Theor Popul Biol *61*, 461-470.

Bamford, D.H., Grimes, J.M., and Stuart, D.I. (2005a). What does structure tell us about virus evolution? Curr Opin Struct Biol *15*, 655-663.

Bamford, D.H., Ravantti, J.J., Rönnholm, G., Laurinavičius, S., Kukkaro, P., Dyall-Smith, M., Somerharju, P., Kalkkinen, N., and Bamford, J.K. (2005b). Constituents of SH1, a novel lipid-containing virus infecting the halophilic euryarchaeon *Haloarcula hispanica*. J Virol *79*, 9097-9107.

Bath, C., Cukalac, T., Porter, K., and Dyall-Smith, M.L. (2006). His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. Virology *350*, 228-239.

Bath, C., and Dyall-Smith, M.L. (1998). His1, an archaeal virus of the *Fuselloviridae* family that infects *Haloarcula hispanica*. J Virol *72*, 9392-9395.

Bawono, P., and Heringa, J. (2014). PRALINE: a versatile multiple sequence alignment toolkit. Methods Mol Biol *1079*, 245-262.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res *27*, 573-580.

Benson, S.D., Bamford, J.K., Bamford, D.H., and Burnett, R.M. (1999). Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. Cell *98*, 825-833.

Bergh, O., Borsheim, K.Y., Bratbak, G., and Heldal, M. (1989). High abundance of viruses found in aquatic environments. Nature *340*, 467-468.

Borriss, M., Helmke, E., Hanschke, R., and Schweder, T. (2003). Isolation and characterization of marine psychrophilic phage-host systems from Arctic sea ice. Extremophiles *7*, 377-384.

Borriss, M., Lombardot, T., Glöckner, F.O., Becher, D., Albrecht, D., and Schweder, T. (2007). Genome and proteome characterization of the psychrophilic *Flavobacterium* bacteriophage 11b. Extremophiles *11*, 95-104.

Boujelben, I., Yarza, P., Almansa, C., Villamor, J., Maalej, S., Anton, J., and Santos, F. (2012). Virioplankton community structure in Tunisian solar salterns. Appl Environ Microbiol *78*, 7429-7437.

Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., and Rohwer, F. (2002). Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A *99*, 14250-14255.

Breitbart, M., Thompson, L.R., Suttle, C.A., and Sullivan, M.B. (2007). Exploring the vast diversity of marine viruses. Oceanography *20*, 135-139.

Breitbart, M., Wegley, L., Leeds, S., Schoenfeld, T., and Rohwer, F. (2004). Phage community dynamics in hot springs. Appl Environ Microbiol *70*, 1633-1640.

Brown, M.V., Philip, G.K., Bunge, J.A., Smith, M.C., Bissett, A., Lauro, F.M., Fuhrman, J.A., and Donachie, S.P. (2009). Microbial community structure in the North Pacific ocean. The ISME journal *3*, 1374-1386.

Brum, J.R., Schenck, R.O., and Sullivan, M.B. (2013). Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. The ISME journal *7*, 1738-1751.

Brüssow, H., Canchaya, C., and Hardt, W.D. (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev *68*, 560-602.

Buckling, A., and Brockhurst, M. (2012). Bacteria-virus coevolution. Adv Exp Med Biol *751*, 347-370.

Calendar, R.L. (2005). The Bacteriophages, 2nd edn (Oxford: Oxford University Press).

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol *17*, 540-552.

Clark, A.J., Inwood, W., Cloutier, T., and Dhillon, T.S. (2001). Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. J Mol Biol *311*, 657-679.

Colangelo-Lillis, J.R., and Deming, J.W. (2013). Genomic analysis of cold-active Colwelliaphage 9A and psychrophilic phage-host interactions. Extremophiles *17*, 99-114.

Comeau, A.M., Bertrand, C., Letarov, A., Tetart, F., and Krisch, H.M. (2007). Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. Virology *362*, 384-396.

Comeau, A.M., Hatfull, G.F., Krisch, H.M., Lindell, D., Mann, N.H., and Prangishvili, D. (2008). Exploring the prokaryotic virosphere. Res Microbiol *159*, 306-313.

Cornelissen, A., Hardies, S.C., Shaburova, O.V., Krylov, V.N., Mattheus, W., Kropinski, A.M., and Lavigne, R. (2012). Complete genome sequence of the giant virus OBP and comparative genome analysis of the diverse PhiKZ-related phages. J Virol *86*, 1844-1852.

Cresawn, S.G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R.W., and Hatfull, G.F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinformatics *12*, 395.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

D'Amico, S., Collins, T., Marx, J.C., Feller, G., and Gerday, C. (2006). Psychrophilic microorganisms: challenges for life. EMBO reports *7*, 385-389.

DasSarma, S., and DasSarma, P. (2012). Halophiles. In Encyclopedia of Life Sciences (Chichester: John Wiley & Sons Ltd.), pp. 1-9.

Demuth, J., Neve, H., and Witzel, K.P. (1993). Direct electron microscopy study on the morphological diversity of bacteriophage populations in lake Plußsee. Appl Environ Microbiol *59*, 3378-3384.

Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., Furlan, M., Desnues, C., Haynes, M., Li, L., *et al.* (2008). Functional metagenomic profiling of nine biomes. Nature *452*, 629-632.

Dwivedi, B., Xue, B., Lundin, D., Edwards, R.A., and Breitbart, M. (2013). A bioinformatic analysis of ribonucleotide reductase genes in phage genomes and metagenomes. BMC Evol Biol *13*, 33.

Dyall-Smith, M., Porter, K., and Tang, S.L. (2013). Official website of the international committee on taxonomy of viruses (ICTV). http://talk.ictvonline.org/files/proposals/taxonomy_proposals_prokaryote1/m/bact01/4633.aspx.

Dyall-Smith, M., Tang, S.L., and Bath, C. (2003). Haloarchaeal viruses: how diverse are they? Res Microbiol *154*, 309-313.

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

Edgell, D.R., Belfort, M., and Shub, D.A. (2000). Barriers to intron promiscuity in bacteria. J Bacteriol *182*, 5281-5289.

Friedman, S.D., Snellgrove, W.C., and Genthner, F.J. (2012). Genomic sequences of two novel levivirus single-stranded RNA coliphages (family *Leviviridae*): evidence for recombinationin environmental strains. Viruses *4*, 1548-1568.

Fuhrman, J.A. (1999). Marine viruses and their biogeochemical and ecological effects. Nature *399*, 541-548.

Garcia-Heredia, I., Martin-Cuadrado, A.B., Mojica, F.J., Santos, F., Mira, A., Anton, J., and Rodriguez-Valera, F. (2012). Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. PloS one *7*, e33802.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome Res *8*, 195-202.

Gowing, M.M., Garrison, D.L., Gibson, A.H., Krupp, J.M., Jeffries, M.O., and Fritsen, C.H. (2004). Bacterial and viral abundance in Ross Sea summer pack ice communities. Mar Ecol Prog Ser *279*, 3–12.

Gowing, M.M., Riggs, B.E., Garrison, D.L., Gibson, A.H., and Jeffries, M.O. (2002). Large viruses in Ross Sea late autumn pack ice habitats. Mar Ecol Prog Ser *241*, 1–11.

Granskog, M., Kaartokallio, H., Kuosa, H., Thomas, D.N., and Vainio, J. (2006). Sea ice in the Baltic Sea – A review. Estuar Coast Shelf S *70*, 145-160.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res *35*, W52-57.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol *59*, 307-321.

Guixa-Boixareu, N., Calderón-Paz, J.I., Heldal, M., Bratbak, G., and Pedrós-Alió, C. (1996). Viral lysis and bacterivory as prokaryotic loss factors along a salinity gradient. Aquat Microb Ecol *11*, 215-227.

Haggard-Ljungquist, E., Halling, C., and Calendar, R. (1992). DNA sequences of the tail fiber genes of bacteriophage P2: evidence for horizontal transfer of tail fiber genes among unrelated bacteriophages. J Bacteriol *174*, 1462-1477.

Hara, S., Terauchi, K., and Koike, I. (1991). Abundance of viruses in marine waters: assessment by epifluorescence and transmission electron microscopy. Appl Environ Microbiol *57*, 2731-2734.

Hatfull, G.F. (2010). Mycobacteriophages: genes and genomes. Annu Rev Microbiol *64*, 331-356.

Hatfull, G.F. (2012a). Complete genome sequences of 138 mycobacteriophages. J Virol *86*, 2382-2384.

Hatfull, G.F. (2012b). The secret lives of mycobacteriophages. Adv Virus Res *82*, 179-288.

Hatfull, G.F. (2014). Mycobacteriophages: windows into tuberculosis. PLoS Pathog *10*, e1003953.

Hatfull, G.F., and Hendrix, R.W. (2011). Bacteriophages and their genomes. Current opinion in virology *1*, 298-303.

Hatfull, G.F., Pedulla, M.L., Jacobs-Sera, D., Cichon, P.M., Foley, A., Ford, M.E., Gonda, R.M., Houtz, J.M., Hryckowian, A.J., Kelchner, V.A., *et al.* (2006). Exploring the

mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet *2*, e92.
Hendrix, R.W. (2002). Bacteriophages: evolution of the majority. Theor Popul Biol *61*, 471-480.
Hendrix, R.W. (2009). Jumbo bacteriophages. Curr Top Microbiol Immunol *328*, 229-240.
Hendrix, R.W., Hatfull, G.F., and Smith, M.C. (2003). Bacteriophages with tails: chasing their origins and evolution. Res Microbiol *154*, 253-257.
Hendrix, R.W., Lawrence, J.G., Hatfull, G.F., and Casjens, S. (2000). The origins and ongoing evolution of viruses. Trends Microbiol *8*, 504-508.
Hendrix, R.W., Smith, M.C., Burns, R.N., Ford, M.E., and Hatfull, G.F. (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci U S A *96*, 2192-2197.
Hofmann, K., and Stoffel, W. (1993). TMbase - A database of membrane spanning proteins segments. Biol Chem Hoppe-Seyler *374*.
Huang, S., Wang, K., Jiao, N., and Chen, F. (2012). Genome sequences of siphoviruses infecting marine Synechococcus unveil a diverse cyanophage group and extensive phage-host genetic exchanges. Environ Microbiol *14*, 540-558.
Inokuchi, Y., Takahashi, R., Hirose, T., Inayama, S., Jacobson, A.B., and Hirashima, A. (1986). The complete nucleotide sequence of the group II RNA coliphage GA. J Biochem *99*, 1169-1180.
Jaakkola, S.T., Penttinen, R.K., Vilen, S.T., Jalasvuori, M., Rönnholm, G., Bamford, J.K., Bamford, D.H., and Oksanen, H.M. (2012). Closely related archaeal *Haloarcula hispanica* icosahedral viruses HHIV-2 and SH1 have nonhomologous genes encoding host recognition functions. J Virol *86*, 4734-4742.
Jäälinoja, H.T., Roine, E., Laurinmäki, P., Kivelä, H.M., Bamford, D.H., and Butcher, S.J. (2008). Structure and host-cell interaction of SH1, a membrane-containing, halophilic euryarchaeal virus. Proc Natl Acad Sci U S A *105*, 8008-8013.
Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., and Godzik, A. (2005). FFAS03: a server for profile--profile sequence alignments. Nucleic Acids Res *33*, W284-288.
Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., and Hendrix, R.W. (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol *299*, 27-51.
Kandiba, L., Aitio, O., Helin, J., Guan, Z., Permi, P., Bamford, D.H., Eichler, J., and Roine, E. (2012). Diversity in prokaryotic glycosylation: an archaeal-derived N-linked glycan contains legionaminic acid. Mol Microbiol *84*, 578-593.
Kannoly, S., Shao, Y., and Wang, I.N. (2012). Rethinking the evolution of single-stranded RNA (ssRNA) bacteriophages based on genomic sequences and characterizations of two R-plasmid-dependent ssRNA phages, C-1 and Hgal1. J Bacteriol *194*, 5073-5079.
Kelley, L.A., and Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nature protocols *4*, 363-371.
King, A.M.Q., Adams, M.J., Carstens, E.B., and Lefkowitz, E.J. (2012). Ninth Report of the International Committee on Taxonomy of Viruses. (San Diego, California, USA: Elsevier Academic Press).
Klein, R., Baranyi, U., Rössler, N., Greineder, B., Scholz, H., and Witte, A. (2002). *Natrialba magadii* virus φCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. Mol Microbiol *45*, 851-863.
Klein, R., Greineder, B., Baranyi, U., and Witte, A. (2000). The structural protein E of the archaeal virus ϕCh1: evidence for processing in *Natrialba magadii* during virus maturation. Virology *276*, 376-387.
Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol *305*, 567-580.
Kropinski, A.M., Prangishvili, D., and Lavigne, R. (2009). Position paper: the creation of a rational scheme for the nomenclature of viruses of Bacteria and Archaea. Environ Microbiol *11*, 2775-2777.

Krumsiek, J., Arnold, R., and Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics *23*, 1026-1028.
Krupovič, M., and Bamford, D.H. (2007). Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. BMC Genomics *8*, 236.
Krupovič, M., and Bamford, D.H. (2008). Virus evolution: how far does the double beta-barrel viral lineage extend? Nature reviews Microbiology *6*, 941-948.
Krupovič, M., and Bamford, D.H. (2010). Order to the viral universe. J Virol *84*, 12476-12479.
Krupovič, M., Forterre, P., and Bamford, D.H. (2010). Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. J Mol Biol *397*, 144-160.
Krupovič, M., Prangishvili, D., Hendrix, R.W., and Bamford, D.H. (2011). Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev *75*, 610-635.
Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Res *19*, 1639-1645.
Kukkaro, P., and Bamford, D.H. (2009). Virus-host interactions in environments with a wide range of ionic strenghts. Envrionmental Microbiology Reports *1*, 71-77.
Kulakov, L.A., Ksenzenko, V.N., Shlyapnikov, M.G., Kochetkov, V.V., Del Casale, A., Allen, C.C., Larkin, M.J., Ceyssens, P.J., and Lavigne, R. (2009). Genomes of "ϕKMV-like viruses" of *Pseudomonas aeruginosa* contain localized single-strand interruptions. Virology *391*, 1-4.
Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nature reviews Microbiology *8*, 317-327.
Lainhart, W., Stolfa, G., and Koudelka, G.B. (2009). Shiga toxin as a bacterial defense against a eukaryotic predator, *Tetrahymena thermophila*. J Bacteriol *191*, 5116-5122.
Lawrence, J.G., Hatfull, G.F., and Hendrix, R.W. (2002). Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches. J Bacteriol *184*, 4891-4905.
Lindell, D., Jaffe, J.D., Johnson, Z.I., Church, G.M., and Chisholm, S.W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. Nature *438*, 86-89.
Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004). Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A *101*, 11013-11018.
Lopez-Bueno, A., Tamames, J., Velazquez, D., Moya, A., Quesada, A., and Alcami, A. (2009). High diversity of the viral community from an Antarctic lake. Science *326*, 858-861.
Luhtanen, A.M., Eronen-Rasimus, E., Kaartokallio, H., Rintala, J.M., Autio, R., and Roine, E. (2014). Isolation and characterization of phage-host systems from the Baltic Sea ice. Extremophiles *18*, 121-130.
Lukashin, A.V., and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res *26*, 1107-1115.
Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. Science *252*, 1162-1164.
Mann, N.H. (2003). Phages of the marine cyanobacterial picophytoplankton. FEMS Microbiol Rev *27*, 17-34.
Maranger, R., Bird, D.F., and Juniper, S.K. (1994). Viral and bacterial dynamics in Arctic sea ice during the spring algal bloom near Resolute, N.W.T., Canada. Mar Ecol Prog Ser *111*, 121-127.
Martinsohn, J.T., Radman, M., and Petit, M.A. (2008). The lambda red proteins promote efficient recombination between diverged sequences: implications for bacteriophage genome mosaicism. PLoS Genet *4*, e1000065.
Mei, Y., Chen, J., Sun, D., Chen, D., Yang, Y., Shen, P., and Chen, X. (2007). Induction and preliminary characterization of a novel halophage SNJ1 from lysogenic *Natrinema* sp. F5. Can J Microbiol *53*, 1106-1110.

Mesyanzhinov, V.V., Robben, J., Grymonprez, B., Kostyuchenko, V.A., Bourkaltseva, M.V., Sykilinda, N.N., Krylov, V.N., and Volckaert, G. (2002). The genome of bacteriophage phiKZ of Pseudomonas aeruginosa. J Mol Biol *317*, 1-19.

Mochizuki, T., Krupovič, M., Pehau-Arnaudet, G., Sako, Y., Forterre, P., and Prangishvili, D. (2012). Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. Proc Natl Acad Sci U S A *109*, 13386-13391.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng *10*, 1-6.

Notredame, C., Higgins, D.G., and Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol *302*, 205-217.

Nuttall, S.D., and Dyall-Smith, M.L. (1993). HF1 and HF2: novel bacteriophages of halophilic archaea. Virology *197*, 678-684.

Onodera, S., Sun, Y., and Mindich, L. (2001). Reverse genetics and recombination in Phi8, a dsRNA bacteriophage. Virology *286*, 113-118.

Oren, A. (2002). Molecular ecology of extremely halophilic Archaea and Bacteria. FEMS Microbiol Ecol *39*, 1-7.

Oren, A., Bratbak, G., and Heldal, M. (1997). Occurrence of virus-like particles in the Dead Sea. Extremophiles : life under extreme conditions *1*, 143-149.

Osborn, A.M., and Böltner, D. (2002). When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. Plasmid *48*, 202-212.

Pagaling, E., Haigh, R.D., Grant, W.D., Cowan, D.A., Jones, B.E., Ma, Y., Ventosa, A., and Heaphy, S. (2007). Sequence analysis of an archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. BMC Genomics *8*, 410.

Paul, J.H. (2008). Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? The ISME journal *2*, 579-589.

Peng, X. (2008). Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of *Sulfolobus* and the implications for plasmid survival. Microbiology *154*, 383-391.

Peng, X., Blum, H., She, Q., Mallok, S., Brugger, K., Garrett, R.A., Zillig, W., and Prangishvili, D. (2001). Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. Virology *291*, 226-234.

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

Pietilä, M.K., Atanasova, N.S., Manole, V., Liljeroos, L., Butcher, S.J., Oksanen, H.M., and Bamford, D.H. (2012). Virion architecture unifies globally distributed pleolipoviruses infecting halophilic archaea. J Virol *86*, 5067-5079.

Pietilä, M.K., Atanasova, N.S., Oksanen, H.M., and Bamford, D.H. (2013a). Modified coat protein forms the flexible spindle-shaped virion of haloarchaeal virus His1. Environ Microbiol *15*, 1674-1686.

Pietilä, M.K., Laurinavičius, S., Sund, J., Roine, E., and Bamford, D.H. (2010). The single-stranded DNA genome of novel archaeal virus *Halorubrum* pleomorphic virus 1 is enclosed in the envelope decorated with glycoprotein spikes. J Virol *84*, 788-798.

Pietilä, M.K., Laurinmäki, P., Russell, D.A., Ko, C.C., Jacobs-Sera, D., Butcher, S.J., Bamford, D.H., and Hendrix, R.W. (2013b). Insights into head-tailed viruses infecting extremely halophilic archaea. J Virol *87*, 3248-3260.

Pietilä, M.K., Laurinmäki, P., Russell, D.A., Ko, C.C., Jacobs-Sera, D., Hendrix, R.W., Bamford, D.H., and Butcher, S.J. (2013c). Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. Proc Natl Acad Sci U S A *110*, 10604-10609.

Pietilä, M.K., Roine, E., Paulin, L., Kalkkinen, N., and Bamford, D.H. (2009). An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. Mol Microbiol *72*, 307-319.

Pikuta, E.V., Hoover, R.B., and Tang, J. (2007). Microbial extremophiles at the limits of life. Crit Rev Microbiol *33*, 183-209.

Pina, M., Bize, A., Forterre, P., and Prangishvili, D. (2011). The archeoviruses. FEMS Microbiol Rev *35*, 1035-1054.

Porter, K., and Dyall-Smith, M.L. (2008). Transfection of haloarchaea by the DNAs of spindle and round haloviruses and the use of transposon mutagenesis to identify non-essential regions. Mol Microbiol *70*, 1236-1245.

Porter, K., Russ, B.E., and Dyall-Smith, M.L. (2007). Virus-host interactions in salt lakes. Curr Opin Microbiol *10*, 418-424.

Porter, K., Tang, S.L., Chen, C.P., Chiang, P.W., Hong, M.J., and Dyall-Smith, M. (2013). PH1: an archaeovirus of *Haloarcula hispanica* related to SH1 and HHIV-2. Archaea *2013*, 456318.

Prangishvili, D. (2013). The wonderful world of archaeal viruses. Annu Rev Microbiol *67*, 565-585.

Prangishvili, D., Garrett, R.A., and Koonin, E.V. (2006). Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. Virus Res *117*, 52-67.

Proctor, L.M., and Fuhrman, J.A. (1990). Viral mortality of marine bacteria and cyanobacteria. Nature *343*, 60-62.

Rachel, R., Bettstetter, M., Hedlund, B.P., Haring, M., Kessler, A., Stetter, K.O., and Prangishvili, D. (2002). Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. Arch Virol *147*, 2419-2429.

Redder, P., Peng, X., Brügger, K., Shah, S.A., Roesch, F., Greve, B., She, Q., Schleper, C., Forterre, P., Garrett, R.A., *et al.* (2009). Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. Environ Microbiol *11*, 2849-2862.

Rohwer, F., Prangishvili, D., and Lindell, D. (2009). Roles of viruses in the environment. Environ Microbiol *11*, 2771-2774.

Rohwer, F., and Thurber, R.V. (2009). Viruses manipulate the marine environment. Nature *459*, 207-212.

Roine, E., Kukkaro, P., Paulin, L., Laurinavičius, S., Domanska, A., Somerharju, P., and Bamford, D.H. (2010). New, closely related haloarchaeal viral elements with different nucleic acid types. J Virol *84*, 3682-3689.

Roine, E., and Oksanen, H.M. (2011). Viruses from the hypersaline environments: current research an future trends. In Halophiles and Hypersaline Environments, A. Ventosa, A. Oren, and Y. Ma, eds. (Heidelberg: Springer), pp. 153-172.

Rokyta, D.R., Burch, C.L., Caudle, S.B., and Wichman, H.A. (2006). Horizontal gene transfer and the evolution of microvirid coliphage genomes. J Bacteriol *188*, 1134-1142.

Rosario, K., and Breitbart, M. (2011). Exploring the viral world through metagenomics. Current opinion in virology *1*, 289-297.

Rose, R.W., Brüser, T., Kissinger, J.C., and Pohlschröder, M. (2002). Adaptation of protein secretion to extremely high-salt conditions by extensive use of the twin-arginine translocation pathway. Mol Microbiol *45*, 943-950.

Rothschild, L.J., and Mancinelli, R.L. (2001). Life in extreme environments. Nature *409*, 1092-1101.

Sabet, S. (2012). Halophilic viruses. In Advances in Understanding the Biology of Halophilic Microorganisms, R. Vreeland, ed. (New York: Springer), pp. 81-116.

Sanjuan, R., Nebot, M.R., Chirico, N., Mansky, L.M., and Belshaw, R. (2010). Viral mutation rates. J Virol *84*, 9733-9748.

Santos, F., Meyerdierks, A., Pena, A., Rossello-Mora, R., Amann, R., and Anton, J. (2007). Metagenomic approach to the study of halophages: the environmental halophage 1. Environ Microbiol *9*, 1711-1723.

Santos, F., Yarza, P., Parro, V., Briones, C., and Anton, J. (2010). The metavirome of a hypersaline environment. Environ Microbiol *12*, 2965-2976.

Santos, F., Yarza, P., Parro, V., Meseguer, I., Rossello-Mora, R., and Anton, J. (2012). Culture-independent approaches for studying viruses from hypersaline environments. Appl Environ Microbiol *78*, 1635-1643.

Saren, A.M., Ravantti, J.J., Benson, S.D., Burnett, R.M., Paulin, L., Bamford, D.H., and Bamford, J.K. (2005). A snapshot of viral evolution from genome analysis of the *Tectiviridae* family. J Mol Biol *350*, 427-440.

Säwström, C., Lisle, J., Anesio, A.M., Priscu, J.C., and Laybourn-Parry, J. (2008). Bacteriophage in polar inland waters. Extremophiles *12*, 167-175.

Schnabel, H. (1984). An immune strain of *Halobacterium halobium* carries the invertible L segment of phage ϕH as a plasmid. Proc Natl Acad Sci U S A *81*, 1017-1020.

Schnabel, H., Palm, P., Dick, K., and Grampp, B. (1984). Sequence analysis of the insertion element ISH1.8 and of associated structural changes in the genome of phage ϕH of the archaebacterium *Halobacterium halobium*. EMBO J *3*, 1717-1722.

Schnabel, H., Schramm, E., Schnabel, R., and Zillig, W. (1982a). Structural variability in the genome of phage ϕH of *Halobacterium halobium*. Mol Gen Genet *188*, 370-377.

Schnabel, H., Zillig, W., Pfaffle, M., Schnabel, R., Michel, H., and Delius, H. (1982b). *Halobacterium halobium* phage ϕH. EMBO J *1*, 87-92.

Schoenfeld, T., Patterson, M., Richardson, P.M., Wommack, K.E., Young, M., and Mead, D. (2008). Assembly of viral metagenomes from yellowstone hot springs. Appl Environ Microbiol *74*, 4164-4174.

Seed, K.D., Bodi, K.L., Kropinski, A.M., Ackermann, H.W., Calderwood, S.B., Qadri, F., and Camilli, A. (2011). Evidence of a dominant lineage of *Vibrio cholerae*-specific lytic bacteriophages shed by cholera patients over a 10-year period in Dhaka, Bangladesh. mBio *2*, e00334-10.

Sharon, I., Battchikova, N., Aro, E.M., Giglione, C., Meinnel, T., Glaser, F., Pinter, R.Y., Breitbart, M., Rohwer, F., and Beja, O. (2011). Comparative metagenomics of microbial traits within oceanic viral communities. The ISME journal *5*, 1178-1190.

Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D.B., Yooseph, S., Zeidner, G., Golden, S.S., Mackey, S.R., *et al.* (2007). Viral photosynthetic reaction center genes and transcripts in the marine environment. The ISME journal *1*, 492-501.

Siddiqui, K.S., Williams, T.J., Wilkins, D., Yau, S., Allen, M.A., Brown, M.V., Lauro, F.M., and Cavicchioli, R. (2013). Psychrophiles. Earth and Planetary Sciences *41*, 87-115.

Sieburth, J.M. (1979). Sea microbes (New York, N. Y.: Oxford University Press).

Silander, O.K., Weinreich, D.M., Wright, K.M., O'Keefe, K.J., Rang, C.U., Turner, P.E., and Chao, L. (2005). Widespread genetic exchange among terrestrial bacteriophages. Proc Natl Acad Sci U S A *102*, 19009-19014.

Sime-Ngando, T., Lucas, S., Robin, A., Tucker, K.P., Colombet, J., Bettarel, Y., Desmond, E., Gribaldo, S., Forterre, P., Breitbart, M., *et al.* (2011). Diversity of virus-host systems in hypersaline Lake Retba, Senegal. Environ Microbiol *13*, 1956-1972.

Šimoliūnas, E., Kalinienė, L., Truncaitė, L., Zajančkauskaitė, A., Staniulis, J., Kaupinis, A., Ger, M., Valius, M., and Meškys, R. (2013). Klebsiella phage vB_KleM-RaK2 - a giant singleton virus of the family Myoviridae. PloS one *8*, e60717.

Snyder, J.C., Wiedenheft, B., Lavin, M., Roberto, F.F., Spuhler, J., Ortmann, A.C., Douglas, T., and Young, M. (2007). Virus movement maintains local virus population diversity. Proc Natl Acad Sci U S A *104*, 19102-19107.

Söding, J., Biegert, A., and Lupas, A.N. (2005). The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res *33*, W244-248.

Spencer, R. (1955). A marine bacteriophage. Nature *175*, 690-691.

Stassen, A.P., Schoenmakers, E.F., Yu, M., Schoenmakers, J.G., and Konings, R.N. (1992). Nucleotide sequence of the genome of the filamentous bacteriophage I2-2: module evolution of the filamentous phage genome. J Mol Evol *34*, 141-152.

Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol *4*, e234.

Suttle, C.A. (2007). Marine viruses--major players in the global ecosystem. Nature reviews Microbiology *5*, 801-812.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol *28*, 2731-2739.
Tang, S.L., Nuttall, S., and Dyall-Smith, M. (2004). Haloviruses HF1 and HF2: evidence for a recent and large recombination event. J Bacteriol *186*, 2810-2817.
Tang, S.L., Nuttall, S., Ngui, K., Fisher, C., Lopez, P., and Dyall-Smith, M. (2002). HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome. Mol Microbiol *44*, 283-296.
Thurber, R.V. (2009). Current insights into phage biodiversity and biogeography. Curr Opin Microbiol *12*, 582-587.
Tischer, I., Gelderblom, H., Vettermann, W., and Koch, M.A. (1982). A very small porcine virus with circular single-stranded DNA. Nature *295*, 64-66.
Tock, M.R., and Dryden, D.T. (2005). The biology of restriction and anti-restriction. Curr Opin Microbiol *8*, 466-472.
Torrella, F., and Morita, R.Y. (1979). Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, oregon: ecological and taxonomical implications. Appl Environ Microbiol *37*, 774-778.
Vestergaard, G., Aramayo, R., Basta, T., Haring, M., Peng, X., Brügger, K., Chen, L., Rachel, R., Boisset, N., Garrett, R.A., *et al.* (2008). Structure of the acidianus filamentous virus 3 and comparative genomics of related archaeal lipothrixviruses. J Virol *82*, 371-381.
Vidgen, M., Carson, J., Higgins, M., and Owens, L. (2006). Changes to the phenotypic profile of *Vibrio harveyi* when infected with the *Vibrio harveyi* myovirus-like (VHML) bacteriophage. J Appl Microbiol *100*, 481-487.
Waldor, M.K., and Mekalanos, J.J. (1996). Lysogenic conversion by a filamentous phage encoding cholera toxin. Science *272*, 1910-1914.
Wang, J., Jiang, Y., Vincent, M., Sun, Y., Yu, H., Bao, Q., Kong, H., and Hu, S. (2005). Complete genome sequence of bacteriophage T5. Virology *332*, 45-65.
Weigel, C., and Seitz, H. (2006). Bacteriophage replication modules. FEMS Microbiol Rev *30*, 321-381.
Weinbauer, M.G. (2004). Ecology of prokaryotic viruses. FEMS Microbiol Rev *28*, 127-181.
Weinbauer, M.G., and Rassoulzadegan, F. (2004). Are viruses driving microbial diversification and diversity? Environ Microbiol *6*, 1-11.
Weitz, J.S., and Wilhelm, S.W. (2012). Ocean viruses and their effects on microbial communities and biogeochemical cycles. F1000 biology reports *4*, 17.
Wells, L.E. (2008). Cold-active viruses. In Psychrophiles: from Biodiversity to Biotechnology, R. Margesin, F. Schinner, J.C. Marx, and C. Gerday, eds. (Heidelberg, Germany: Springer), pp. 157-173.
Wells, L.E., and Deming, J.W. (2006a). Characterization of a cold-active bacteriophage on two psychrophilic marine hosts. Aquat Microb Ecol *45*, 15–29.
Wells, L.E., and Deming, J.W. (2006b). Modelled and measured dynamics of viruses in Arctic winter sea-ice brines. Environ Microbiol *8*, 1115-1121.
Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998). Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A *95*, 6578-6583.
Wilhelm, S.W., and Suttle, C.A. (1999). Viruses and nutrient cycles in the sea - viruses play critical roles in the structure and function of aquatic food webs. Bioscience *49*, 781-788.
Witte, A., Baranyi, U., Klein, R., Sulzner, M., Luo, C., Wanner, G., Krüger, D.H., and Lubitz, W. (1997). Characterization of *Natronobacterium magadii* phage ϕCh1, a unique archaeal phage containing DNA and RNA. Mol Microbiol *23*, 603-616.
Wommack, K.E., and Colwell, R.R. (2000). Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev *64*, 69-114.
Zdobnov, E.M., and Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.

Zhang, Z., Liu, Y., Wang, S., Yang, D., Cheng, Y., Hu, J., Chen, J., Mei, Y., Shen, P., Bamford, D.H., *et al.* (2012). Temperate membrane-containing halophilic archaeal virus SNJ1 has a circular dsDNA genome identical to that of plasmid pHH205. Virology *434*, 233-241.

Zinger, L., Gobet, A., and Pommier, T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. Mol Ecol *21*, 1878-1896.