Going to extremes – a metagenomic journey into the dark matter of life.

Arnthór Aevarsson[1,*], Anna-Karina Kaczorowska[2], Björn Thor Adalsteinsson[1], Josefin Ahlqvist[3], Salam Al-Karadaghi[4], Joseph Altenbuchner[5], Hasan Arsin[6], Úlfur Áugúst Átlasson[1], David Brandt[7], Magdalena Cichowicz-Cieślak[8], Katy A. S. Cornish[9], Jérémy Courtin[1], Slawomir Dabrowski[10], Håkon Dahle[6,11], Samia Djeffane[1], Sebastian Dorawa[8], Julia Dusaucy[1], Francois Enault[12], Anita-Elin Fedøy[6], Stefanie Freitag-Pohl[9], Olafur H. Fridjonsson[1], Clovis Galiez[13], Eirin Glomsaker[14], Mickael Guérin[1], Sigurd E. Gundesø[14], Elisabet E. Gudmundsdóttir[1], Hördur Gudmundsson[1], Maria Håkansson[4], Christian Henke[7,15], Alexandra Helleux[1], Jørn Remi Henriksen[14], Sigrídur Hjörleifdóttir[1], Gudmundur O. Hreggvidsson[1,16], Andrius Jasilionis[3], Annika Jochheim[13], Ilmur Jónsdóttir[1], Lilja Björk Jónsdóttir[1], Agata Jurczak-Kurek[17], Tadeusz Kaczorowski[8], Jörn Kalinowski[7], Lukasz P. Kozlowski[13,23], Mart Krupovic[18], Karolina Kwiatkowska-Semrau[8], Olav Lanes[14], Joanna Lange[19], Julien Lebrat[1], Javier Linares-Pastén[3], Ying Liu[18], Steffen A. Lorentsen[14], Tobias Lutterman[7], Thibaud Mas[12], William Merré[1], Milot Mirdita[13], Agnieszka Morzywołek[8], Eric Olo Ndela[12], Eva Nordberg Karlsson[3], Edda Olgudóttir[1], Cathrine Pedersen[14], Francine Perler[22], Sólveig K. Pétursdóttir[1], Magdalena Plotka[8], Ehmke Pohl[9,20], David Prangishvili[18], Jessica L. Ray[6,21], Birkir Reynisson[1], Tara Róbertsdóttir[1], Ruth-Anne Sandaa[6], Alexander Sczyrba[7,15], Sigurlaug Skírnisdóttir[1], Johannes Söding[13], Terese Solstad[14], Ida H. Steen[6], Sigmar Karl Stefánsson[1], Martin Steinegger[13], Katrine Stange Overå[14], Bernd Striberny[14], Anders Svensson[4], Monika Szadkowska[8], Emma J. Tarrant[9], Paul Terzian[12], Mathilde Tourigny[1], Tom van den Bergh[19], Justine Vanhalst[1], Jonathan Vincent[12], Bas Vroling[19], Björn Walse[4], Lei Wang[5], Hildegard Watzlawick[5], Martin Welin[4], Olesia Werbowy[8], Ewa Wons[8], Ruoshi Zhang[13].

[1]Matis ohf, Vinlandsleid 12, Reykjavik 113, Iceland, [2]Collection of Plasmids and Microorganisms, Faculty of Biology, University of Gdansk, Wita Stwosza 59, Gdansk 80-308, Poland, [3]Biotechnology, Department of Chemistry, Lund University, PO Box 124, SE-221 00 Lund, Sweden, [4]SARomics Biostructures, Scheelevägen 2, SE-223 81 Lund, Sweden, [5]Institute for Industrial Genetics, University of Stuttgart, Allmandring 31, 70569 Stuttgart, Germany, [6]Department of Biological Sciences, University of Bergen, PO Box 7803, N-5020 Bergen, Norway, [7]Center for Biotechnology, Bielefeld University, Bielefeld 33615, Germany, [8]Laboratory of Extremophiles Biology, Department of Microbiology, Faculty of Biology, University of Gdansk, Wita Stwosza 59, Gdansk 80-308, Poland, [9]Department of Chemistry, Durham University, South Road, Durham DH1 3LE, United Kingdom, [10]A&A Biotechnology, Al. Zwycięstwa 96/98, Gdynia 84-451, Poland, [11]Department of Informatics, University of Bergen, PO Box 7803, N-5020 Bergen, Norway, [12]Université Clermont Auvergne, CNRS, Laboratoire Microorganismes: Génome et Environnement, UMR 6023, Clermont-Ferrand, France, [13]Quantitative and Computational Biology, Max-Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany, [14]ArcticZymes Technologies PO Box 6463, 9294 Tromsø, Norway, [15] Computational Metagenomics, Bielefeld University, 30501 Bielefeld, Germany, [16] Faculty of Life and Environmental Sciences, University of Iceland, Reykjavik, Iceland, [17]Department of Molecular Evolution, Faculty of Biology, University of Gdansk, Wita Stwosza 59, Gdansk 80-308, Poland,

[18]Institute Pasteur, Department of Microbiology, 75015 Paris, France, [19]Bio-Prodict, Nieuwe Marktstraat 54E 6511AA Nijmegen, Netherlands, [20]Department of Biosciences, Durham University, South Road, Durham DH1 3LE, United Kingdom, [21]NORCE Environment, NORCE Norwegian Research Centre AS, Nygårdsgaten 112, 5008 Bergen, Norway, [22]Perls of Wisdom Biotech Consulting, 74 Fuller Street, Brookline, MA 02446, USA, [23]Institute of Informatics, Faculty of Mathematics, Informatics, and Mechanics,University of Warsaw, Banacha 2, Warsaw 02-097, Poland.

*Correspondence:

Arnthor Aevarsson, Matis ohf, Vinlandsleid 12, Reykjavik 113, Iceland, e-mail: arnthor@matis.is

Keywords: virus, virosphere, archaea, metagenomics, bioprospecting, thermophiles.

Abstract

The Virus-X – Viral Metagenomics for Innovation Value – project was a scientific expedition to explore and exploit uncharted territory of genetic diversity in extreme natural environments such as geothermal hot springs and deep-sea ocean ecosystems. Specifically, the project was set to analyse and exploit viral metagenomes with the ultimate goal of developing new gene products with high innovation value for applications in biotechnology, pharmaceutical, medical, and the life science sectors. Viral gene pool analysis is also essential to obtain fundamental insight into ecosystem dynamics and to investigate how viruses influence the evolution of microbes and multicellular organisms. The Virus-X Consortium, established in 2016, included experts from eight European countries. The unique approach based on high throughput bioinformatics technologies combined with structural and functional studies resulted in the development of a biodiscovery pipeline of significant capacity and scale. The activities within the Virus-X consortium cover the entire range from bioprospecting and methods development in bioinformatics to protein production and characterisation, with the final goal of translating our results into new products for the bioeconomy. The significant impact the consortium made in all of these areas was possible due to the successful cooperation between expert teams that worked together to solve a complex scientific problem using state-of-the-art technologies as well as developing novel tools to explore the virosphere, widely considered as the last great frontier of life.

Introduction

The virosphere comprises the largest reservoir of unknown genetic diversity in the whole biosphere and is considered to be the last great frontier of life. It is constituted by viruses, which are sophisticated acellular infectious entities that typically consist of genetic material (RNA or DNA) enclosed in a protein shell called a capsid. Infection of a host organism, either prokaryotic or eukaryotic, is necessary to activate their genetic blueprints. Viruses replicate by hijacking the host metabolic machinery, and the viral life cycle is completed when newly assembled viral particles are released into their environment, ready to infect new host organisms and repeat the cycle of infection and replication. The number of virus particles on Earth has been estimated to be in the order of 10^{31}, which makes them the most abundant entities in the whole biosphere (Hendrix *et al.* 1999; Mushegian 2020). By virtue of their biology, viruses are ubiquitous and can be found in all conceivable ecosystems, including those with extremes of temperature, pH, salinity, and pressure. Such ecosystems include hot springs in Iceland with temperatures close to the boiling point of water to ice-cold deep-sea habitats of the Arctic Ocean at water pressures of several hundred atmospheres (Romancer *et al.* 2007; Harrison *et al.* 2013). The abundance and diversity of viruses that flourish in extreme habitats is truly striking. Metagenomic analyses of water in hot springs, for example, reveal that these geochemical hot spots are rich in viruses with a reported titer ranging from 10^5 to 10^6 ml^{-1} and a production rate of approximately 10^9 viral particles per liter per day (Breitbart *et al.* 2004). While the overall diversity of the viral community is extremely high, only a small fraction of it (< 1%) has been explored so far (Forterre and Prangishvili 2009; Mokili *et al.* 2012; Krishnamurthy and Wang 2017). Therefore, the viral gene pool is critical to study functional dynamics of ecosystems, to investigate how viruses influence evolution of microbes and multicellular organisms (Rodriguez-Valera *et al.* 2009; Clokie *et al.* 2011; Sandaa and Bratbak 2018). As viruses seem to be a driving force in shaping biogeochemical cycles on a global scale (Middelboe *et al.* 1996; Gobler *et al.* 1997; Thingstad 2000; Danovaro *et al*. 2016), they may also provide the key to understanding biodiversity and biosphere functioning at the level of individual species and genomes (Thingstad 2000; Urich *et al.* 2014; Harrison and Brockhurst 2017; Hreggvidsson *et al.* 2017; Tuttle and Buchan 2020). Much remains to be investigated regarding virus-host interplay in nature and in an exploration of the relatively unexploited global virome as an immense genetic resource for innovation potential (Rinke *et al.* 2013; Paez-Espino *et al.* 2016). The virosphere thus offers enormous promise for the development of unique technologies for life science, industrial, medical and diagnostic applications.

The exploitation of the virosphere and its genetic diversity was a central objective of the Virus-X project (Viral Metagenomics for Innovation Value; http://virus-x.eu). This EU Horizon 2020 endeavour launched in 2016 and continued through 2020, engaged in a collaborative effort of four biotech companies representing Small and Medium-sized Enterprises (SME partners) and ten research groups from eight European countries. The project acronym Virus-X refers to the unknown virus and thus to the uncharted territory of viral genomics. The research activities of the consortium particularly focused on bacterial and archaeal viruses from extreme habitats, their impact on microbial dynamics, and viruses as a source of novel enzymes for biotech applications. The latter was crucial as, due to their highly unusual biology and lifecycles, viruses have developed specialized replication machineries employing proteins with unique properties often very different from cellular host enzymes (Kazlauskas *et al.* 2016; Zhu *et al*. 2017; Kazlauskas *et al.* 2018). Many milestones of modern molecular biology and biotechnology were accomplished using viral proteins as their genomes are packed with genes encoding unusual nucleic acid processing enzymes with remarkable properties (Ofir and Sorek 2018). Today, key tools in biotechnology originate from only very few bacterial viruses (e.g. λ, T4 or T7), mostly infecting the model organism *E. coli* (Murray and Gann 2007). Furthermore, many new applications are based on old enzymes put to new uses, albeit sometimes following structural/functional modification (Kaczorowski and Szybalski 1998; Wilson *et al.* 2013; Lu *et al.* 2020). Further technological progress in the field depends on the discovery and development of enzymes from novel sources (van den Burg 2003; Castelan-Sanchez *et al.* 2019; Jin *et al*. 2019). Until now, this reservoir has been near inaccessible since methodologies for infecting and propagating viruses in a laboratory environment are cumbersome for many microbes, and only a small fraction of microorganisms can be cultivated. The advent of metagenomics that exploits Next Generation Sequencing technologies (NGS) has finally made the virome genetic content accessible for exploration (Angly *et al*. 2006; Kristensen *et al*. 2010; Schoenfeld *et al*. 2010; Rosario and Breitbart 2011; Beerenwinkel *et al.* 2012). However, exploring the virosphere remains a challenging task of emerging metagenomic methodologies for a number of reasons. Sampling in extreme environments is difficult and grueling, the concentration of viral genetic material is often very low, and up to 70% of the genes in novel environmental viruses bear little or no similarity to known nucleotide sequences deposited in the GenBank non-redundant database (Breitbart *et al.* 2002; Edwards and Rowher 2005; Hjörleifsdottir *et al.* 2014; Gil *et al.* 2021). Thus, this enormous genetic resource, largely unexplored, can be considered as the dark matter of life. Consequently, the Virus-X project was set to analyse and exploit the virome of

extreme natural habitats with the ultimate goal of developing new gene products with high innovation value for applications in the bioeconomy. For this objective, enzyme discovery and development in the Virus-X project focused on specific non-structural viral proteins such as proteins participating in nucleic acids metabolism, e.g., polynucleotide kinase (Blöndal *et al.* 2005a) and RNA ligase (Blöndal *et al.* 2005b); viral recombination machinery (Stefanska *et al.* 2014, Stefanska *et al.* 2016); nucleotide metabolism; transcription; replication, including DNA polymerases (Hjorleifsdottir *et al.* 2014), DNA helicases, single-strand DNA binding proteins, and other DNA and RNA processing enzymes. Also of particular interest were enzymes involved in bacterial cell lysis as potential antimicrobials (Plotka *et al.* 2014; Plotka *et al.* 2015) and components of antiviral mechanisms, including the CRISPR system (Nordberg Karlsson *et al.* 2020). In addition, the consortium focussed on the *unknown* – gene products with not yet determined properties and the potential of completely new functionalities.

Project overview.

All activities within the project were divided into four platforms that cover metagenome retrieval (P1), bioinformatics (P2), protein characterisation (P3), and invention to innovation (P4). While each platform was led by one partner, all partners contributed to at least two platforms, creating a highly intertwined network of collaboration and ensuring the overall success of the project (Fig. 1).

One key objective of the Virus-X project was the development of new technologies for metagenomics ranging from bioprospecting of complex viromes in extreme environments to the analysis and annotation of large-scale metagenomic data sets. For this purpose, new algorithms were developed, which formed part of the Virus-X metagenomics toolbox, many of which are now implemented in freely available webservers or as free, open-source software (Table 1).

Virus-X workflow

Platform 1. Metagenome retrieval.

A wide range of natural habitats was selected as a source of viral metagenomes. The 50 sites within 12 hydrothermal regions in Iceland explored by scientists from Matis included intertidal geothermal coastal areas and terrestrial hot springs (16-96°C, pH 2.0-9.3) (Table S1, Supporting information). Another team from Institut Pasteur (France) specializing in viruses of Archaea collected 49 samples from hot springs in the western part of Georgia (50-95°C pH 7.0-8.0), Kuju and Beppu regions on Kyushu island in Japan (60-90°C, pH 2.0-7.0), as well as solfataric field in Pozzuoli (87-93°C, pH 1.5-2.0) and hot springs of the Campi Flegrei volcano (81-96°C, pH 1.0-7.0) in Italy (Table S2, Supporting information). Finally, the group from the University of Bergen (Norway) sampled 34 sites that included deep-sea hydrothermal vents located in the Arctic Mid-Ocean Ridge: The Ægir Vent Field - Central Mohn's Ridge and the Loki's Castle Vent Field (Northern Atlantic) and shallow to deep seawaters from the Norwegian sea (Jan Mayen Fracture Zone; -0.8°C - -3.9°C, pH 7.9), the eastern part of the Fram strait (areas around Svalbard; 1.9°C) (Table S3, Supporting information). Sampling of marine sites required access to research cruises equipped with a remote submersible operating robotic vehicle Ægir 6000 with up to 6000 m depth rating (Fig. 2).

In order to identify and characterise new viruses by electron microscopy, enrichment culture techniques based on simulating conditions that favour propagation of both bacterial and archaeal viruses were employed (Liu *et al.* 2019). It was hence possible to observe diverse virion shapes characteristic of viruses infecting hyperthermophilic members of the phylum Crenarchaota, of the domain Archaea (Fig. 3A). Their morphologies are unique, and such diversity has not been observed among viruses infecting Bacteria and Eukarya. This is matched by the low similarity of thermophilic viruses at the nucleotide sequence level, where the majority of genes show no homology to other viruses or cellular organisms (Krupovic *et al.* 2018; Wang *et al.* 2020a). Several new viruses that infect hyperthermophilic archaea and belong to families *Globuloviridae*, *Tristromaviridae*, *Lipothrixviridae*, *Rudiviridae,* and *Portogloboviridae* were isolated and characterized (Liu *et al.* 2017; Liu *et al.* 2019; Baquero *et al.* 2020). Portogloboviruses were also discovered to carry mini-CRISPR arrays containing spacers targeting each other as well as other viruses, exemplifying a novel mechanism

promoting interviral conflicts and superinfection exclusion in extreme environments (Medvedeva *et al.* 2019). A remarkable novel mode of DNA packaging was observed in the case of the *Sulfolobus* polyhedral viruses SPV1 and SPV2, the first members of the new family (*Portogloboviridae*) (Wang *et al.* 2019). In the icosahedral virions, the double-stranded DNA is wound up sinusoidally into a spherical coil filling the protein capsid (Fig. 3B and C). Electron cryo-microscopy enabled reconstruction at a near-atomic resolution of the virion structure of portoglobovirus SPV1, tristromavirus PFV2, lipothrixvirus SFV1 and rudivirus SSRV1 and revealed that, in all these cases, DNA is packaged in the form of nucleocapsid, where capsid proteins tightly wrap around the DNA and maintain it in A-form (Fig. 3B and C) (Liu *et al.* 2018; Wang *et al.* 2019a; Wang *et al.* 2020a; Wang *et al.* 2020b). This finding is striking as previous reports suggested that, in the case of bacterial viruses, genomic DNA predominantly adopt the B-form (Black and Thomas 2012). The more tightly packed A-form DNA could represent a general mechanism of adaptation to extreme thermal environments. Despite an increase in general knowledge on viral diversity, biogeography data is still scarce or almost missing from some specific geographic areas such as deep-sea vent systems (Le Moine Bauer *et al.* 2018) or arctic waters (Sandaa *et al.* 2018). Preliminary analysis of viromes from extreme marine habitats in Norwegian territories confirmed the distribution of dominant taxa in the ocean (order *Caudovirales* and family *Phycodnaviridae)* (Brum *et al.* 2015; Mihara *et al.* 2018; Endo *et al.* 2020), but due to our novel sampling and preparation protocol, we were also able to capture viruses with different genome types and large differences in capsid size in the same sample (Blanc-Mathieu *et al.* 2021).

Platform 2. High-throughput sequencing, assembly, and annotation.

Metagenomic samples collected in the first phase of the project were processed to isolate genetic material for sequencing. One of our approaches in Virus-X has been to capture the total viral diversity within marine samples from arctic and deep-sea vent systems, including viruses with all genome types and within all size ranges. The protocol developed was based on less rigorous filtration steps (only 0.45 µm) to capture even the largest virions, and the extraction of total nucleic acids (TNA) rather than double-stranded DNA only. The RNA fractions were enzymatically converted to cDNA prior to metagenomic library preparation for high-throughput sequencing. For the remaining samples, we used consecutive steps of microfiltration to remove cellular organisms, followed by concentrating the samples and DNA

extraction using standard procedures. Viral genomic DNA can be different from cellular DNA in many aspects. It often contains a high fraction of modified bases and complex genomic structures such as extremely long direct or inverted repeats and terminal redundancies (Li *et al.* 2014; Davison 2015; Weigele and Raleigh 2016). This leads to particular challenges both in the sequencing procedure and the downstream assembly with short sequence reads and uneven sequence depth (Klumpp *et al*. 2012; Beaulaurier *et al*. 2020). Sequencing efforts were guided by various parameters such as overall diversity of reads, estimated coverage based on the size distribution of contigs, rarefaction analysis, using several selected indicator genes, and the proportion of single reads. In the project, various NGS platforms (Illumina MiSeq, HiSeq; Oxford Nanopore) were used for metagenomic total DNA analysis as well as for assembly of individual viral genomes out of the complex viral metagenome sequence data.

In one exemplary study, metagenomic sequencing of polyhedral and filamentous viruses that infect archaea belonging to the order Sulfolobales yielded seven complete or near-complete genomes (Liu *et al.* 2019). They were assigned at the nucleotide sequence level to viruses belonging to the family of polyhedral *Portogloboviridae* (SPV1, SPV2), filamentous *Rudiviridae* (SBRV1), and *Lipothrixviridae* (SBFV3). In addition, two genomes of filamentous viruses (SBFV1 and SBFV2) could not be assigned to any family and hence form a new group of filamentous archaeal viruses. The same applies to the seventh viral genome (SBV1), which is likely to represent a new virus family. A striking feature was the observation that ca. 75% of genes contained in analysed genomes show no homology to genes in other viruses or cellular organisms, making archaeal viruses a valuable source of unknown genes. A similar conclusion was reached from the analysis of genome content of bacterial viruses from Iceland metagenomic samples. Out of 2015 genes from 22 phages that infect bacteria of genus *Thermus*, 129 genes (6%) were assigned as type-A (known function), 157 (8%) were type-B (putative function), and 1727 genes (86%) were classified as type-C (no known function).

In total, metagenomic sequencing within the Virus-X project reached a final output of 290 Gbases. The computational analysis started with assembly and binning, which allowed the preliminary assessment for extensive downstream analysis (Fig. 4). We implemented a modern workflow using the Common Workflow Language (CWL) to keep it maintainable and portable across computing environments (Amstutz *et al.* 2016). Briefly, the workflow consists of the following steps: (i) quality control to trim or even discard the raw reads based

on the quality score of their individual bases; (ii) assembly of input reads into longer contigs. The MEGAHIT assembler (Li *et al.* 2015) was chosen for this task based on its evaluation within the Critical Assessment of Metagenome Interpretation challenge (Sczyrba *et al.* 2017); (iii) mapping of reads to assembled contigs to determine sequencing coverage of each contig; (iv) gene prediction stage to identify potential genes within the assembled contigs with the use of Prodigal (Hyatt *et al.* 2010); (v) taxonomic and functional annotation of predicted genes using DIAMOND (Buchfink *et al.* 2015) and MEGAN (Huson *et al.* 2016) tools. The functional annotation also included a prediction of functional domains using the PFAM database (El-Gebali *et al.* 2019), pathways using KEGG (Kanehisa *et al.* 2017), as well as the metagenomic tools developed within the project (vi; see below). In step (vii), binning of assembled contigs into metagenome-assembled genomes (MAGs) was performed using MetaBAT 2 (Kang *et al.* 2019), followed (viii) by taxonomic classification of MAGs utilizing GTDB-Tk (Chaumeil *et al.* 2019). Freely accessible bioinformatic tools developed within the Virus-X project are summarized in Table 1.

Applying the workflow to all Virus-X datasets resulted in a total assembly size of 23 Gbases, 38,417,734 contigs, and 54,106,508 predicted genes. In addition, 3591 viral metagenome datasets publicly available from NCBI's Sequence Read Archive (SRA) were also processed and predicted genes included in the subsequent clustering step (S1 file, Supporting information).

From this extensive database, it was possible to extract 157,428,937 open reading frames (ORFs, genetic sequences that are potentially translated to proteins) with a minimum length of 60 amino acids. To deal with the high degree of redundancy in this data set, amino-acid sequences were clustered using MMseqs2 (Steinegger and Söding 2017) down to 30% of pairwise sequence identity and using a conservative 90% minimum coverage threshold, resulting in 56,790,072 clusters. To substantially increase the protein sequence recovery, the open-source de-novo protein-level assembler (Plass) was developed (Steinegger *et al.* 2019). Of vital importance to the project was gene annotation, as viral genes display extensive sequence divergence making homology detection exceedingly difficult. At the same time, virus genes are ideal targets for homology-based function prediction because the functions and structures of viral enzymes are usually very well conserved despite the high sequence divergence. As a consequence, enzymes from bacteria or viruses detected as homologous to a viral protein usually still have a similar molecular function. Therefore, heavy emphasis in Virus-X was placed on gathering extensive expertise in gene annotation to develop a new

methodology for assigning functions to gene products, both by in-depth bioinformatics and by an experimental pipeline for selected genes. The deep annotation of ORFs was performed using the Uniboost database and HHboost procedure based on reverse profile-sequence search to detect proteins that show remote homology at the amino-acid sequence level (Mirdita *et al.* 2017; Steinegger *et al.* 2019). To further increase sensitivity, a consensus sequence was computed among each of the clusters eliminating sequencing errors and microdiversity. This database of ORFs, the Uniboost matches, and the annotations of each Uniboost entry, as well as the outputs of the workflow described above, were integrated into the main Virus-X EMGB data browser at CeBiTec at Bielefeld University, Germany (Fig. 4). This data is browsable to select gene targets of potential interest for cloning and expression. To further help in their annotation, ORFs were compared to PHROGs (Prokaryotic Virus Remote Homologous Groups), a database of deeply clustered and well-annotated viral protein groups (https://phrogs.lmge.uca.fr/).

In parallel to the discovery and analysis of single genes, metagenomic data were used to increase our understanding of the composition and dynamics of viral communities in these under-explored ecosystems such as deep-sea vents or arctic waters. Using a metabarcoding approach on 42 arctic seawater samples, covering the water column from 0 to 1000 m, and both the polar day and night, it was possible to gain a first glimpse into the viral ecology in these arctic environments (Fig. 5). Unlike with cellular organisms, there is no universal marker gene that targets all viruses. Using two genes targeting common groups of marine viruses, namely *g23,* capturing T4-like bacteriophages (Filee *et al*. 2005) and the *mcp* gene of large dsDNA phytoplankton viruses (Larsen *et al.* 2008), we demonstrated that seasonality is a key factor shaping arctic viral communities. Viral diversity and virus-to-host ratios dropped substantially at the beginning of the spring, then increased during the season, with the highest rates observed during the winter (Sandaa *et al.* 2018). In order to extend this analysis to the whole viral community, 20 arctic metagenomes were generated. Tailed bacteriophages belonging to the *Caudovirales* were the most dominant group, followed by giant dsDNA viruses belonging to the nucleocytoplasmic large DNA viruses (NCLDV). Viruses with other genome forms, such as ssDNA, dsRNA and ssRNA, were also detected. Viral operational taxonomic units (vOTU) were generated and compared to marine virus populations from Tara expeditions (Gregory *et al.* 2019) and to the RefSeq database. We confirmed the influence of

seasonality over arctic viral communities, even if sampling depths and sampling location also seemed to have a substantial impact on the differentiation of these viruses.

Platform 3. Protein selection, production, and characterisation.

Access to an extensive genetic database necessitates a careful selection of gene products for further analysis. All putative genes in the Virus-X database were divided into three categories with respect to assigned function: (i) type-A: assigned function through significant sequence identity at the amino-acid sequence level; (ii) type-B: putative function based on weak sequence similarity, and (iii) type-C: genes with no assigned function. Genes from all three categories were selected for cloning and expression for subsequent functional and/or structure determination. Genes that fall into categories A or B were evaluated and ranked according to the interest of the consortium members. In this process, particular emphasis was given to the SME partners to select targets with biotechnological potential. The selection of type-C target genes was also based on the following considerations. Genes located in a cluster of structural proteins were excluded, and genes showing extended conservation among different viral genomes but still with unknown function were prioritized. The particular focus was on gene targets coding for proteins with biotech potential that meet the following criteria: (i) heat or salt tolerance, (ii) cold-active/heat lability, (iii) high substrate specificity, (iv) high affinity, (v) strand displacement activity, (vi) DNA/RNA modifying/synthesizing/interacting, (vii) high fidelity, (viii) processivity and robustness in complex reaction settings. Most features outlined above are impossible to determine solely from nucleotide sequences alone. However, our approach involved experimental characterisation of target proteins as well as structural analysis by X-ray crystallography (Fig. 6) to shed light on protein function and allow for further optimization of annotation protocols.

As structure-based multiple sequence alignments are more accurate than sequence-based alignments (Carpentier and Chomilier 2019), we employed the 3DM information systems developed by our partner Bio-Prodict that collectively cover the entire structural space for a given enzyme family (Kuipers *et al.* 2010). This repository was designed to allow the integration of target proteins in these 3DM systems, which allows the detailed analysis of new protein sequences in the context of all the information of complete protein families. This approach facilitated our target selection process and was proven to be essential to gain insights into the specific adaptations of the viral proteins in comparison to known cellular

proteins. The second internal database developed within the project was the protein Tiki Wiki that holds all experimental information from cloning, characterisation to structure determination of over 800 target proteins.

To facilitate speed in recombinant protein production, *Escherichia coli* was selected as a priority production system, based on its fast growth on cheap media, available cloning vectors, various promoters, and alternative solubility and purification tags, as well as co-plasmids with chaperonins to facilitate folding. Plasmid vectors used to express cloned target genes selected from the Virus-X database were mainly chosen to include tightly controlled inducible promoters like positively controlled L-rhamnose $rhaP_{BAD}$ (Wegerer *et al.* 2008) or T7 ϕ10 promoter (Studier and Moffatt 1986; Studier *et al.* 1990). In cases where the protein overproduction yielded mainly insoluble protein, the creation of fusion proteins by adding genes coding for the maltose-binding protein (MBP) improved the synthesis of soluble targets, as shown previously (Wang *et al.* 2013). Another successfully applied strategy was lowering the temperature at the induction step to 20-25°C. In other cases, a number of different *E. coli* expression strains, including those supplemented with a plasmid pRARE (Novagen) coding for rare-codon tRNAs, were tested, and the use of strains containing additional chaperone genes to aid protein folding (Nishihara *et al.* 1998; de Marco and de Marco 2004; de Marco *et al*. 2007) proved very beneficial for targets from deep-sea metagenomes.

In total, 659 genes were cloned into expression vectors and 478 tested in production trials. The majority of them, 65%, resulted in the production of recombinant proteins; however, only approximately 42% were produced in high yields in a soluble form (Fig. 6). These soluble target proteins were expressed in standard shaking incubators, usually at temperatures between 18 and 25°C and purified by a combination of affinity chromatography (immobilized metal affinity chromatography (IMAC) for His-tagged proteins or amylose affinity chromatography for MBP-fusion proteins) with size exclusion chromatography or ion exchange chromatography. After purification, activity testing and characterisation of recombinant MBP-fused proteins was, in the majority of cases, performed without MBP removal, while solubility tags were removed before recombinant protein was forwarded to crystallisation. Protein production was successfully upscaled for a majority of targeted constructs. In some cases, target genes were cloned, generating two or more variants that were subjected for protein production upscaling in parallel. For those type B and type C targets where structural solutions with *ab initio* methods were unsuccessful, selenomethionine substituted protein was

produced as described before (Turner *et al.* 2007; Russo *et al.* 2009) to enable structure determination by *multiple anomalous diffraction* (MAD) methods (Hendrickson 1990). Protein characterisation was divided into biophysical techniques that included Electron Spray Ionisation mass spectrometry (ESI-MS) to investigate protein integrity and covalent modifications, circular dichroism (CD) to confirm proper folding, and Thermal Shift Analysis (TSA) to study protein stability (Groftenhauge *et al.* 2015). TSA is beneficial to identify optimal storage and transport conditions as well as to guide subsequent crystallisation trials (Bruce *et al.* 2019) (Fig. 7A).

Protein structure determination through X-ray crystallography complemented bioinformatics for selected targets in all categories: i) to provide structural insight in atomic detail into gene products of established function but of particular interest to the consortium members; ii) to verify putative functional assignment, and iii) to provide important clues to the possible function of hypothetical genes that lack significant similarity to known sequences and have no assigned putative function. As a result, 24 crystal structures of target proteins have been determined, 7 in category A, 11 in B, and 6 in C, respectively. In addition, 16 targets are at the final stage of analysis, where we have already obtained crystals and collected diffraction data sets. Perfect examples of the collaborative effort are the functional and structural analysis of two enzymes with an antimicrobial activity where several Virus-X consortium members contributed: *Bacillus subtilis* prophage PBSX lytic cassette protein XepA (Fig. 7B) (Freitag-Pohl *et al.* 2019) and highly thermostable Ph2119 endolysin from extremophilic Ph2119 bacteriophage of *Thermus scotoductus* (Plotka *et al.* 2020) (Fig. 7C and D). The latter is N-acetylmuramoyl-L-alanine amidase (EC 3.5.1.28), with unique structural features among thermophilic phages lytic enzymes. It contains characteristic Zn^{2+} binding site and shows similarity to T3 and T7 phage lysozymes and also to eukaryotic peptidoglycan recognition proteins (PGRPs), which are engaged in innate immunity and are highly conserved from invertebrates to mammals (Dziarski and Gupta 2006a; Dziarski and Gupta 2006b; Wang *et al.* 2007). In the case of Ph2119 endolysin, the conserved motif is formed by three amino acid residues: His30-His132-Cys140 that coordinate Zn^{2+} (Fig 7D).

The Virus-X strategy to identify proteins with novel functions within the type B and C targets was based on two main elements. First, optimized sequence analysis followed by structure-based studies to unveil sequence-function data. Second, experimental characterisation based on the combination of biochemical and biophysical tests. This kind of approach allowed to thoroughly characterize 97 proteins from all categories, including novel thermostable DNA polymerases with strand displacement activity, robust lytic enzymes with high thermal

stability that can be considered as perspective antibacterial agents (Plotka *et al.* 2019a, Plotka *et al.* 2019b; Freitag-Pohl *et al.* 2019; Plotka *et al.* 2020), and single-strand DNA binding proteins that increase specificity either in PCR-based or isothermal amplification of DNA (Werbowy *et al.* 2020). Some of these proteins, including a set of novel thermostable heat-shock proteins that are capable of stabilizing proteins and supporting their folding process, have led to the patent application filed in 2020.

Platform 4. Invention to innovation.

Research projects are often invention-rich; however, the apparent gap between scientific research and commercialization is often difficult to bridge and has been called "The valley of Death" (Ehlers 1998). Virus-X partners have therefore developed an exploitation plan with clear commercialization goals recognizing the importance of turning inventions into innovations and exploitable intellectual property (IP). The project addressed real opportunity markets for the SMEs. The current molecular diagnostic methods face constraints such as DNA/RNA extraction from samples, low DNA content in a sample, inhibition of the polymerase chain reaction (PCR), mutation of amplified PCR products, and DNA contamination (Yang and Rothman 2004; Afzal 2020). Novel enzymes dedicated to molecular biology, such as robust thermostable strand-displacement DNA polymerases and DNA/RNA replication proteins, are in constant demand. New products, technologies, and knowledge that can lead to better diagnostics and research tools will, in the long run, save time and money and improve standards of living.

In addition, Virus-X put a strong emphasis on demonstration and dissemination activities for increasing visibility of commercially exploitable results to end-users, building on experience, established channels, and proven strategies in previous projects. Highlights of this work include the "Going to Extremes" session at the multidisciplinary Euroscience Open Forum 2018 in Toulouse, France, participation in the Europe-wide *Researcher's Night* in October 2019 in Reykjavik (http://virus-x.eu/uncategorized/virus-x-on-researchers-night/), a feature on the Futuris program of the international TV channel, Euronews (https://www.euronews.com/2020/09/14/virus-hunters-explore-iceland-s-geothermal-hot-springs-for-solutions) as well as a popular science publication (Ævarsson 2018).

COVID-19 task force

Since January 2020 the World has been struggling with the Coronavirus pandemic (COVID-19) caused by severe acute respiratory syndrome coronavirus SARS-Cov-2 (Zhu *et al.* 2020). In order to join the global efforts to combat the disease, the **Virus-X COVID-19 task force** was established. It aims to leverage our knowledge and expertise to support COVID-19 research by (i) participating in large-scale sequencing of virus isolates in different countries; (ii) tracking SARS-CoV-2 evolution; (iii) the development of quick, high-throughput diagnostic methods, and (iv) supporting structure-based drug discovery for specific antiviral treatments. The Virus-X research group at the University of Bielefeld has taken an early lead in complete genome SARS-CoV-2 sequencing and sequence analysis in Germany as part of the *Deutsche Covid-19 Omics consortium* (Schulte-Schrepping *et al.* 2020). The group at the University of Clermont, Auvergne, France, uses their bioinformatics programs to analyse hundreds of SARS-CoV-2 genomes published in open databases; phylogenetic analysis of local clusters provides essential information on the epidemiology of the disease. SARomics (Lund, Sweden) cofounded the LundaGUARD, a private-public consortium in Sweden developing a platform for the rapid scientific response to the COVID-19 crisis and future pandemic threats. Bio-Prodict (Netherlands) has released in open access their customized expert system 3DM for several proteins encoded by SARS-CoV-2, including viral enzymes, potential key targets for drugs, and the surface proteins fundamental for the development of vaccines. The Virus-X SME partner A&A Biotechnology (Poland) stepped up to the challenge and quickly increased the capacity of their RNA extraction kits used as the first step in the testing regime (Caruana *et al.* 2020). The company is also taking part in the development and implementation of coronavirus test kits. While the current gold standard tests are based mainly on RT-PCR, the consortium has also focused on new technologies, namely *reverse transcriptase Loop-mediated isothermal amplification* (RT-LAMP). As shown recently by other groups, this method offers a fast and reliable test for viral RNA in a single step at one temperature (Dao Thi *et al*. 2020; Ganguli *et al.* 2020; Huang *et al.* 2020). The Virus-X SME, ArcticZymes in Norway, uses the unique project resources to develop new tools for molecular diagnostics.

Conclusions

The Horizon 2020 research programme was set up by the European Union to support the goal to become the world's most competitive knowledge-based economy by coupling excellent research and innovation. The results obtained within the frame of the Virus-X project have expanded our knowledge and understanding in many areas: an exploration of new genetic territory and sequence diversity, development of new approaches and tools for viral metagenomics, understanding of microbial communities, identification of commercially valuable genes, and the corresponding impact on European biotech industry and companies involved in the project towards new marketable products and services. The project combined basic research questions with real market opportunities for the participating SME partners. The development of novel enzymes for molecular biology allows for the design and development of new tools for genetic engineering. Our investigations of thermophilic enzymes such as strand-displacement DNA polymerases and DNA/RNA replication proteins could significantly improve their current use as diagnostic tools that are of vital importance, particularly in times of a new pandemic. Current molecular diagnostic methods for COVID-19 face significant challenges, including DNA/RNA extraction from a range of sources, low DNA content in a sample, inhibition of the polymerase chain reaction, mutation of amplified PCR products, and DNA contamination. New products, technologies, and knowledge that can lead to better diagnostics and research tools will save time and money and improve standards of living. It is evident that the legacy of the Virus-X project in terms of vast sequences, cloned genes, and produced target proteins, as well as the bioinformatics software that was developed, will continue to be explored and exploited for years to come.

Acknowledgments

We would like to thank the undergraduate students involved at several Universities for their hard work and dedication. We are also very grateful to our technical staff for their support during the project.

Funding

Funding was provided by the Europan Union's Horizon 2020 Research and Innovation Programme Virus-X project: Viral Metagenomics for Innovation Value (grant no. 685778). This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B).

Conflict of interest: None declared.

References

Ævarsson A. Les virus a la source. *La Recherche* 2018;**537**:92-93. English version at http://virus-x.eu/uncategorized/virus-x-popular-science/.

Afzal A. Molecular diagnostic technologies for COVID-19: limitations and challenges. *J Adv Res* 2020;**23**:149-59

Amstutz P, Crusoe MR, Tijanić N *et al.* Common Workflow Language, v1.0. figshare. Dataset 2016, https://doi.org/10.6084/m9.figshare.3115156.v2.

Angly FE, Felts B, Breitbart M *et al*. The marine viromes of four oceanic regions. *Plos Biol* 2006;**4**:e368.

Baquero DP, Contursi P, Piochi M *et al.* New virus isolates from Italian hydrothermal environments underscore the biogeographic pattern in archaeal virus communities. *ISME J* 2020;**14**:1821-33.

Beaulaurier J, Luo E, Eppley JM *et al.* Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. *Genome Res* 2020;**30**:437-46.

Beerenwinkel N, Gunthard HF, Roth V *et al.* Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol* 2012;**3**:329.

Black LW, Thoms JA. Condensed genome structure. *Adv Exp Med Biol* 2012;**725**:469-87.

Blanc-Mathieu R, Dahle H, Hofgaard A *et al.* A persistent giant algal virus, with a unique morphology, encodes an unprecedented number of genes involved in energy metabolism. *J Virol* 2021; doi:10.1128/JVI.02446-20.

Blondal T, Hjorleifsdottir S, Aevarsson A *et al.* Characterization of a 5'-polynucleotide kinase/3'-phosphatase from bacteriophage RM378. *J Biol Chem* 2005a;**280**:5188-94.

Blondal T, Thorisdottir A, Unnsteinsdottir U *et al.* Isolation and characterization of a thermostable RNA ligase 1 from a *Thermus scotoductus* bacteriophage Ts2126 with good single-stranded DNA ligation properties. *Nucl Acids Res* 2005b;**33**:135-42.

Breitbart M, Salamon P, Andresen B, *et al.* Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 2002;**99**:15250-5.

Breitbart M, Wegley L, Leeds S *et al.* Phage community dynamics in hot springs. *Appl Environ Microbiol* 2004;**70**:1633-40.

Bruce D, Cardew E, Freitag-Pohl S *et al*. How to stabilize protein: stability screens for thermal shift assays and nano differential scanning fluorimetry in the Virus-X project. *J Vis Exp* 2019;**144**:e58666.

Brum JR, Ignacio-Espinoza JC, Roux S *et al.* Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* 2015;**348**:1261498.

Buchfink B, Xie C, Huson D. Fast and sensitive protein alignment using DIAMOND. *Nat Meth* 2015;**12**:59-60.

Carpentier M, Chomilier J. Protein multiple alignments: sequence-based versus structure-based programs. *Bioinformatics* 2019;**35**:3970-80.

Caruana G, Croxatto A, Coste AT *et al*. Diagnostic strategies for SARS-CoV-2 infection and interpretation of microbiological results. *Clin Microbiol Infect* 2020;**26**:1178-82.

Castelan-Sanchez HG, Lopes-Rosas I, Garcia-Suastegui WA *et al.* Extremophile deep-sea viral communities from hydrothermal vents: structural and functional analysis. *Marine Genomics* 2019;**46**:16-28.

Chaumeil PA, Mussing AJ, Hugenholtz *et al.* GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 2020;**36**:1925-7.

Christendat D, Yee A, Dharamsi A *et al.* Structural proteomics of an archaeon. *Nat Struct Biol* 2000;**7**:903-9.

Clokie MRJ, Millard AD, Letarov AV *et al.* Phages in nature. *Bacteriophage* 2011;**1**:31-45.

Dahle H, Økland I, Thorseth IH *et al.* Energy landscapes shape microbial communities in hydrothermal systems on the Arctic mid-Ocean Ridge. *ISME J* 2015;**9**:1593-606.

Dao Thi VL, Herbst K, Boerner K *et al*. A colorimetric RT-LAMP assay and LAMP-sequencing for detecting SARS-CoV-2 RNA in clinical samples. *Sci Transl Med* 2020;**12**:eabc7075.

Danovaro R, Dell'Anno A, Corinaldesi C *et al.* Virus-mediated archaeal hecatomb in the deep seafloor. *Sci Adv* 2016;**2**:e1600492.

Davison J. Pre-early functions of bacteriophage T5 and its relatives. *Bacteriophage* 2015;**5**:e1086500.

De Marco A, de Marco V. Bacteria c-transformed with recombinant proteins and chaperones cloned in independent plasmids are suitable for expression tuning. *J Biotechnol* 2004;**109**:45-52.

De Marco A, Deuerling E, Mogk A *et al*. Chaperone-based procedure to increase yields of soluble recombinant proteins produced in *E. coli. BMC Biotechnol* 2007;**7**:32.

Dziarski R, Gupta D. The peptidoglycan recognition proteins (PGRPs). *Genome Biol* 2006;**7**:232.

Dziarski R, Gupta D. Mammalian PGRPs: novel antibacterial proteins. *Cell Microbiol* 2006;**8**:1059-69.

Edwards RA, Rowher F. Viral metagenomics. *Nat Rev Microbiol* 2005;**3**:504-10.

Ehlers VJ. Unlocking our future: toward a new National Science Policy. A report to Congress by the House Committee on Science. Washington D.C., GPO, 1998. https://www.aaas.org/sites/default/files/s3fs-public/GPO-CPRT-105hprt105-b.pdf.

El-Gebali S, Mistry J, Bateman A *et al.* The Pfam protein families database in 2019. *Nucl Acids Res* 2019;**47**:D427-D432.

Endo H, Blanc-Mathieu R, Li Y *et al.* Biogeography of marine giant viruses reveals their interplay with eukaryotes and ecological functions. *Nat Ecol Evol* 2020;**4**:1639-49. https://doi.org/10.1038/s41559-020-01288-w.

Filee J, Tetart F, Suttle CA *et al*. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc Natl Acad Sci USA* 2005;**102**:12471-6.

Freitag-Pohl S, Jasilionis A, Håkansson M *et al.* Crystal structures of the *Bacillus subtilis* prophage lytic cassette proteins XepA and YomS. *Acta Cryst D* 2019:**75**:1028-39.

Ganguli A, Mostafa A, Berger J *et al*. Rapid isothermal amplification and portable detection system for SARS-CoV-2. *Proc Natl Acad Sci USA* 2020;**117**:22727-35.

Gil JF, Mesa V, Estrada-Ortiz N *et al*. Viruses in extreme environments, current overview, and biotechnological potential. *Viruses* 2021;**13**:81.

Gobler CJ, Hurchins DA, Fisher NS *et al.* Release and bioavailability of C, N, P, Se and Fe following viral lysis of a marine Chrysophyte. *Limnol Oceanogr* 1997;**42**:1492-504.

Gregory AC, Zayed AA, Conceicao-Neto N *et al.* Marine DNA viral macro- and microdiversity from Pole to Pole. *Cell* 2019;**177**:1109-23.

Groftenhauge MK, Hajizadeh NR, Swann MJ *et al.* Protein-ligand interactions investigated by thermal shift assays (TSA) and dual polarization interferometry (DPI). *Acta Cryst D* 2015;**71**:36-44.

Harrison E, Brockhurst MA. Ecological and evolutionary benefits of temperate phage: what does or doesn't kill you makes you stronger. *Bioessays* 2017;**39**:1700112.

Harrison JP, Gheeraert N, Tsigelnitskiy D *et al.* The limits for life under multiple extremes. *Trends Microbiol* 2013;**21**:204-12.

Hendrickson WA. Determination of macromolecular structures from anomalous diffrection of synchrotron radiation. *Science* 1991;**254**:51-8.

Hendrix RW, Smith MC, Burns RN *et al.* Evolutionary relationships among diverse bacteriophages and prophages: all the wolrd's phage. *Proc Natl Acad Sci USA* 1999;**96**:2192-7.

Henke C, Sczyrba A. EMGB – A fast and versatile metagenome annotation browser for the web. *Manuscript in preparation.*

Hjorleifsdottir S, Aevarsson A, Hreggvidsson OH *et al.* Isolation, growth and genome of the *Rhodothermus* RM378 thermophilic bacteriophage. *Extremophiles* 2014;**18**:261-70.

Hreggvidsson GO, Petursdottir SK, Stefansson SK *et al.* Divergence of species in the geothermal environment. In Stan-Lotter H, Fedrihan S (eds) Adaptation of Microbial Life to Environmental Extremes. Springer Verlag, 2017, pp. 41-74.

Huang C, Wang Y, Li X *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;**395**:497-506.

Huang WE, Lim B, Hsu C.C. et al. RT-LAMP for rapid diagnosis of coronavirus SARS-CoV-2. *Microb Biotechnol* 2020;**13**:950-61.

Huson DH, Beier S, Flade I *et al.* MEGAN community edition- interactive exploration and analysis of large-scale microbiome sequencing data. *PLOS Comput Biol* 2016;**12**:e1004957.

Hyatt D, Chen GL, LoCascio PF *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform* 2010;**11**:119.

Jin M, Gai Y, Guo X *et al.* Properties and applications of extremozymes from deep-sea extremophilic microorganisms: a mini review. *Mar Drugs* 2019;**17**:656.

Kaczorowski T, Szybalski W. Genomic DNA sequencing by SPEL-6 primer walking using hexamer ligation. *Gene* 1998;**223**:83-91.

Kanehisa M, Furumichi M, Tanabe M *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucl Acids Res* 2017;**45**:D353-D361.

Kang DD, Li F, Kirton E *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *Peer J* 2019;**7**:e7359.

Kazlauskas D, Krupovic M, Venclovas C. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucl Acids Res* 2016;**44**:4551-64.

Kazlauskas D, Sezonov G, Charpin N *et al.* Novel families of archaeo-eukaryotic primases associated with mobile genetic elements of bacteria and archaea. *J Mol Biol* 2018;**430**:737-50.

Klumpp J, Fouts DE, Sozhamannan S. Next generation sequencing technologies and the changing landscape of phage genomics. *Bacteriophage* 2012;**2**:190-9.

Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res 2017;**239**:136-142.

Kristensen DM, Mushegian AR, Dolja VV *et al*. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* 2010;**18**:11-9.

Krupovic M, Cvirkaite-Krupovic V, Iranzo J *et al.* Viruses of archaea: structural, functional, environmental and evolutionary genomics. *Virus Res* 2018;**244**:181-93.

Kuipers RK, Joosten HJ, van Berkel WJH *et al.* 3DM: systematic analysis of heterogenous superfamily data to discover protein functionalilties. *Proteins* 2010;**78**:2101-13.

Larsen JB, Larsen A, Bratbak G et al. Phylogenetic analysis of members of the Phycodnaviridae virus family using amplified fragments of the major capsid protein gene. *Appl Environ. Microbiol* 2008;**74**:3048-57.

Le Romancer M, Gaillard M, Geslin C *et al.* Viruses in extreme environments. *Rev Environ Sci Biotechnol* 2007;**6**:17-31.

Le Moine Bauer S, Stensland A, Daae FL *et al.* Water masses and depth structure prokaryotic and T4-like viral communities around hydrothermal systems of the Nordic Seas. *Front Microbiol* 2018;**9**:1002.

Levy Karin EL, Mirdita M, Söding J. MetaEuk – sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* 2020;**8**:48.

Li D, Liu CM, Luo R *et al.* MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674-6.

Li S, Fan H, An X *et al*. Scrutinizing virus genome termini by high-throughput sequencing. *PLOS One* 2014;**9**:e85806.

Liu Y, Ishino S, Ishino Y, Pehau-Arnaudet G *et al.* A novel type of polyhedral viruses infecting hyperthermophilic archaea. *J Virol* 2017;**91**:e00589-17.

Liu Y, Osinski T, Wang F *et al.* Structural conservation in a membrane-enveloped filamentous virus infecting a hyperthermophilic acidophile. *Nat Commun* 2018;**9**:3360.

Liu Y, Brandt D, Ishino S *et al.* New archaeal viruses discovered by metagenomic analysis of viral communities in enrichment cultures. *Environ Microbiol* 2019;**21**:2002-14.

Lu B, Dong L, Yi D *et al.* Transposase-assisted tagmentation of RNA/DNA hybrid duplexes. *eLife* 2020;**9**:e54919.

Medvedeva S, Liu Y, Koonin EV *et al.* Virus-borne mini-CRISPR arrays are involved in interviral conflicts. *Nat Commun* 2019;**10**:5205.

Nordberg Karlsson E, Sardari RRR, Ron EYC *et al.* Metabolic engineering of thermophilic bacteria for production of biotechnologically interesting compounds. (In N Lee, ed.) *Biotechnological applications of extremophilic microorganisms.* 2020*,* pp. 73-96*,* deGruyter, eBook (PDF), ISBN 978-3-11-042433-1.

Middelboe M, Jorgensen NOG, Kroer N *et al.* Effect of viruses on nutrient turnover and growth efficiency of noninfected marine bacterioplankton. *Appl Environ Microbiol* 1996;**62**:1991-7.

Mihara T, Koyano H, Hingamp P *et al.* Taxon richness of "megaviridae" exceeds those of bacteria and archaea in the ocean. *Microbes Environ* 2018;**33**:162-171.

Mirdita M, von den Driesch, Galiez C *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucl Acids Res* 2017;**45**:D170-D176.

Mirdita M, Steinegger M, Söding J, MMseqs2 desktop and local web server app for fast, interactive sequence searches, *Bioinformatics* 2019;**35**:2856-83.

Mirdita M, Steinegger M, Breitwieser F *et al.* Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* 2021;**37**:btab184. https://doi.org/10.1093/bioinformatics/btab184.

Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. *Curr Opinion Virol* 2012;**2**:63-77.

Motejadded H, Altenbuchner J. Construction of a dual-tag system for gene expression, protein affinity purification and fusion protein processing. *Biotechnol Lett* 2009;**31**:543-9.

Murray NE, Gann G. What has phage lambda ever done for us. *Current Biology* 2007;**17**:R305-12.

Mushegian AR. Are there 10^{31} virus particles on Earth, or more, or fewer? *J Bacteriol* 2020:**202**:e00052-20.

Nishihara N, Kanemori M, Kitagawa M *et al.* Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in *Escherichia coli*. *Appl Environ Microbiol* 1998;**64**:1694-9.

Ofir G, Sorek R. Contemporary phage biology: from classic models to new insights. *Cell* 2018;**172**:1260-70.

Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA *et al.* Uncovering Earth's virome. *Nature* 2016:**536**:425-30.

Plotka M, Kaczorowska AK, Stefanska A *et al.* Novel highly thermostable endolysin from *Thermus scotoductus* MAT2119 bacteriophage Ph2119 with amino acid sequence similarity to eukaryotic peptidoglycan recognition proteins. *Appl Environ Microbiol* 2014;**80**:886-95.

Plotka M, Kaczorowska AK, Morzywolek A *et al.* Biochemical characterization and validation of a catalytic site of a highly thermostable Ts2631 endolysin from *Thermus scotoductus* phage vB_Tsc2631. *Plos One* 2015;**10**:e0137374.

Plotka M, Sancho-Vaello E, Dorawa S *et al.* Structure and function of the Ts2631 endolysin of *Thermus scotoductus* phage vB_Tsc2631 with unique N-terminal extension used for peptidoglycan binding. *Sci Rep* 2019a;**9**:1261.

Plotka M, Kapusta M, Dorawa S *et al.* Ts2631 endolysin from the extremophilic *Thermus scotoductus* bacteriophage vB_2631 as an antimicrobial agent against Gram-negative multidrug-resistant bacteria. *Viruses* 2019b;**11**:657.

Plotka M, Szadkowska M, Håkansson M *et al.* Molecular characterization of a novel lytic enzyme LysC from *Clostridium intestinale* URNW and its antibacterial activity mediated by positively charged N-terminal extension. *Int J Mol Sci* 2020;**21**:4894.

Rinke C, Schwientek P, Sczyrba A *et al*. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013;**499**:431-7.

Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B *et al.* Explaining microbial population genomics through phage predation. *Nat Rev Microbiol* 2009;**7**:823-36.

Rosario K, Breitbart M. Exploring the viral world through metagenomics. *Curr Opin Virol* 2011;**1**:289-97.

Russo S, Schweitzer JE, Polen T *et al.* Crystal structure of the caseinolytic protease gene regulator, a transcriptional activator in *Actinomycetes*. *J Biol Chem* 2009;**284**:5208-16.

Sandaa RA, Bratbak G. Is the virus important? And some other questions. *Viruses* 2018;**10**:442.

Sandaa RA, Storesund JE, Olesin E *et al.* Seasonality drives microbial community structure, shaping both eukaryotic and prokaryotic host-viral relationships in an Arctic marine ecosystem. *Viruses* 2018;**10**:715.

Schoenfeld T, Liles M, Wommack KE *et al*. Functional viral metagenomics and the next generation of molecular tools. *Trends Microbiol* 2010;**18**:20-9.

Schulte-Schrepping J, Reusch N, Paclik D *et al.* Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* 2020;**182**:1419-40.

Sczyrba A, Hofmann P, Belmann P *et al.* Critical Assessment of metagenome Interpretation – a benchmark of metagenomics software. *Nat Methods* 2017;**11**:1063-71.

Steen IH, Dahle H, Stokke R *et al.* Novel barite chimneys at the Loki's Castle Vent Field shed light on key factors shaping microbial communities and functions in hydrothermal systems. *Front Microbiol* 2016;**9**:1510.

Stefanska A, Kaczorowska AK, Plotka M *et al.* Discovery and characterization of RecA protein of thermophilic bacterium *Thermus thermophilus* MAT72 phage Tt72 that increases specificity of a PCR-based DNA amplification. *J Biotechnol* 2014;**182-183**:1-10.

Stefanska A, Gaffke L, Kaczorowska AK *et al.* Highly thermostable RadA protein from the archaeon *Pyrococcus woesei* enhances specificity of simplex and multiplex PCR assays. *J Appl. Genetics* 2016;**57**:239-49.

Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;**35**:1026-8.

Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. *Nat Commun* 20018;**9**:2542.

Steinegger M, Meier M, Mirdita M *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 2019;**20**:473.

Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* 2019;**16**:603-6.

Studier FW, Moffatt BA. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol* 1986;**189**:113-30.

Studier FW, Rosenberg AH, Dunn JJ *et al*. Use of T7 RNA polymerase to direct expression of cloned genes *Methods Enzymol* 1990;**185**:60-89.

Terzian P, Olo Ndela E, Galiez C *et al.* PHROG: families of prokaryotic virus proteins clustered using remote homology. Submitted (2021).

Thingstad TF. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol Oceanogr* 2000;**45**:1320-8.

Turner P, Pramhed A, Kanders E *et al.* Expression, purification, crystallization and preliminary X-ray diffraction analysis of *Thermotoga neapolitana* beta-glucosidase B. *Acta Cryst.* 2007;**F63**:802-6.

Tuttle MJ, Buchan A. Lysogeny in the oceans: lessons from cultivated model systems and a reanalysis of its prevalence. *Environ Microbiol* 2020;**22**:4919-33.

Uchiyama T, Miyazaki K. Functional metagenomics for enzyme discovery: challenges to efficient screening. *Curr Opin Biotechnol* 2009;**20**:616-22.

Urich T, Lanzen A, Stokke R *et al.* Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environ Microbiol* 2014;**16**:2699-710.

Van den Burg B. Extremophiles as a source for novel enzymes. *Curr Opin Microbiol* 2003;**6**:213-8.

Wang M, Liu LH, Wang S *et al*. Human peptidoglycan recognition proteins require zinc to kill both gram-positive and gram-negative bacteria and are synergistic with antibacterial peptides. *J Immunol* 2007;**178**:3116-25.

Wang L, Watzlawick H, Fridjonsson OH, *et al.* Improved soluble expression of the gene encoding amylolytic enzyme Amo45 by fusion with the mobile-loop-region of co-chaperonin GroES in *Escherichia coli*. *Biocatal Biotransform* 2013;**31**:335-42.

Wang F, Liu Y, Su Z *et al.* A packing for A-form DNA in an icosahedral virus. *Proc Natl Acad Sci USA* 2019a;**116**:22591-7.

Wang F, Baquero DP, Su Z *et al.* Structure of a filamentous virus uncovers familial ties within the archaeal virosphere. *Virus Evol* 2020a;**6**:veaa23.

Wang F, Baquero DP, Beltran LC *et al.* Structures of filamentous viruses infecting hyperthermophilic archaea explain DNA stabilization in extreme environments. *Proc Natl Acad Sci USA* 2020b;**117**:19643-52.

Wegerer A, Sun T, Altenbuchner J. Optimization of an *E. coli* L-rhamnose-inducible expression vector test of various genetic module combinations. *BMC Biotechnol* 2008;**8**:2.

Weigele P, Raleigh EA. Biosynthesis and function of modified bases in bacteria and their viruses. *Chem Rev* 2016;**116**:12655-87.

Werbowy O, Stefanska-Kazmierczak, Jurczak-Kurek A *et al.* The characteristics of new SSB proteins from metagenomic libraries and their use in biotech applications. *Proceedings* 2020;**50**:135.

Wilson RH, Morton SK, Deiderick H *et al.* Engineered DNA ligases with improved activities *in vitro*. *Protein Eng Des Sel* 2013;**26**:471-8.

Yang S, Rothman RE. PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *Lancet Infect Dis* 2004;**4**:337-48.

Zhang R, Mirdita M, Levy Karin E *et al.* SpacePHARER: sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics* 2021;**37**:btab222. https://doi.org/10.1093/bioinformatics/btab222

Zhou B, Wang L, Mitsunobo H *et al*. Deep-sea vent phage DNA polymerase specifically initiates DNA synthesis in the absence of primers. *Proc Natl Acad Sci USA* 2017;**114**:E2310-E2318.

Zhu N, Zhang D, Wang W *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* 2020;**382**:727-33.

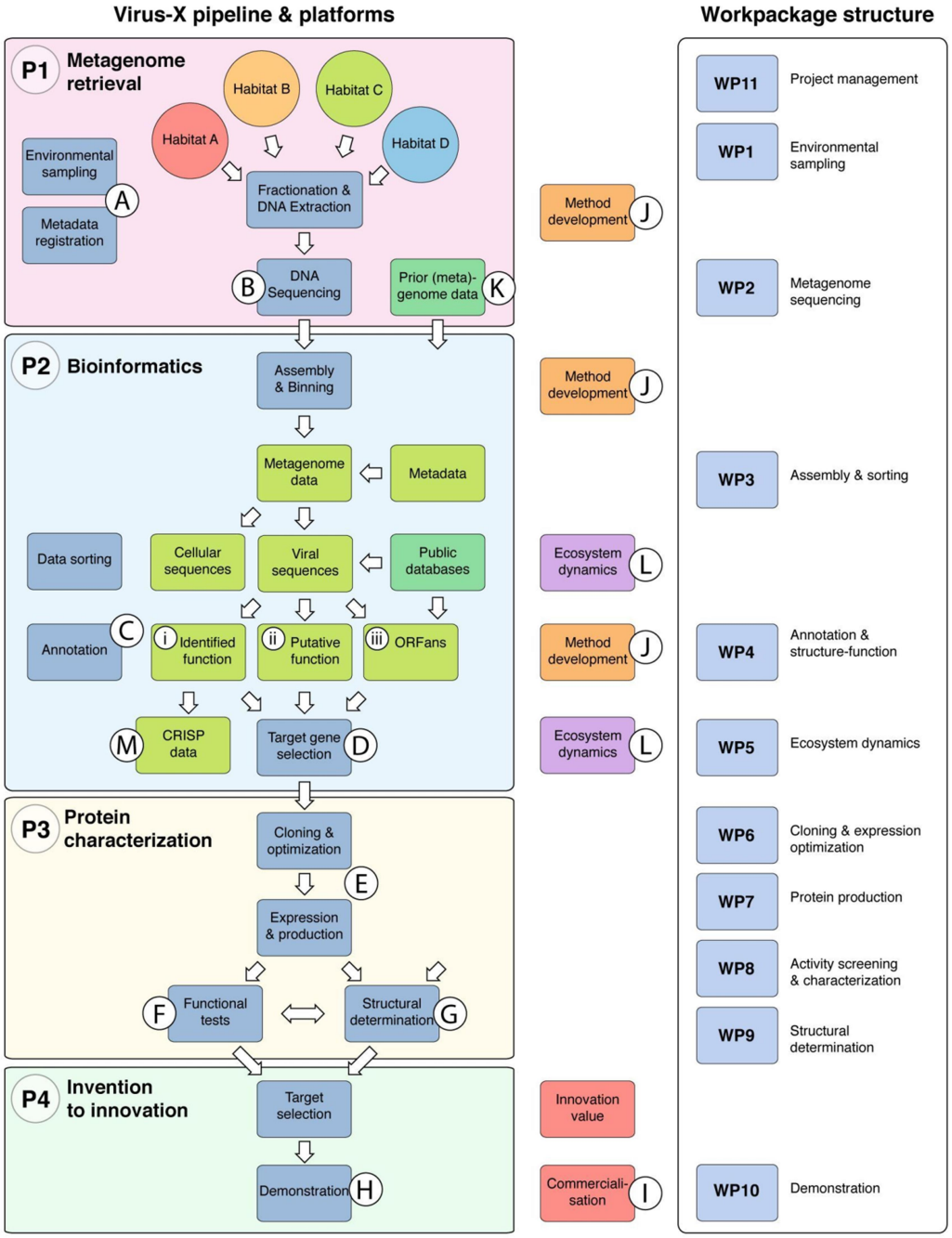

Fig. 1. The Virus-X discovery pipeline was designed to bring viral gene products from environmental sampling (**A**) via sequencing (**B**), bioinformatics and annotation (**C**), selection of gene targets (**D**), cloning, expression and production (**E**), functional (**F**) and structural characterisation (**G**), to demonstration (**H**) and ultimately to commercialization (**I**). Although industrially driven, the project was designed to be a vehicle for method development (**J**) in the

field of metagenomics as well as to address some specific challenges that set viral metagenomics apart from metagenomics of cellular genomes. The viral metagenome approach brought the need for improvements and innovative approaches in the functional assignment of genes as the focus was on exploring the outer realm of sequence diversity where many gene sequences lack significant similarity to known sequences (Uchiyama and Miyazaki 2009). Therefore, particular emphasis was put on bioinformatics approaches to gene annotation. Data from previous bacteriophage metagenome studies and genome data were used as reference sequences and to initiate bioinformatics and the downstream workflow in the project at an early stage in the project timeline (**K**). *In silico* annotation was complemented with functional and structural studies to establish the function of targeted gene products. Functional studies (**F**) required cloning and expression of selected genes and functional assays in a high-throughput set-up. Structure determination by X-ray crystallography (**G**) complemented bioinformatics and biochemical characterisation for selected targets in all categories. In addition to the research and development of selected gene products, the work program involved the study of specific aspects of the functional dynamics of microbial communities (**L**) and the interplay between prokaryotic cellular population and the associated viral population, such as by the study of viral diversity, bacterial immunity and the CRISPR system (**M**). Sequence data were thus also acquired from the cellular members of the sampled communities.

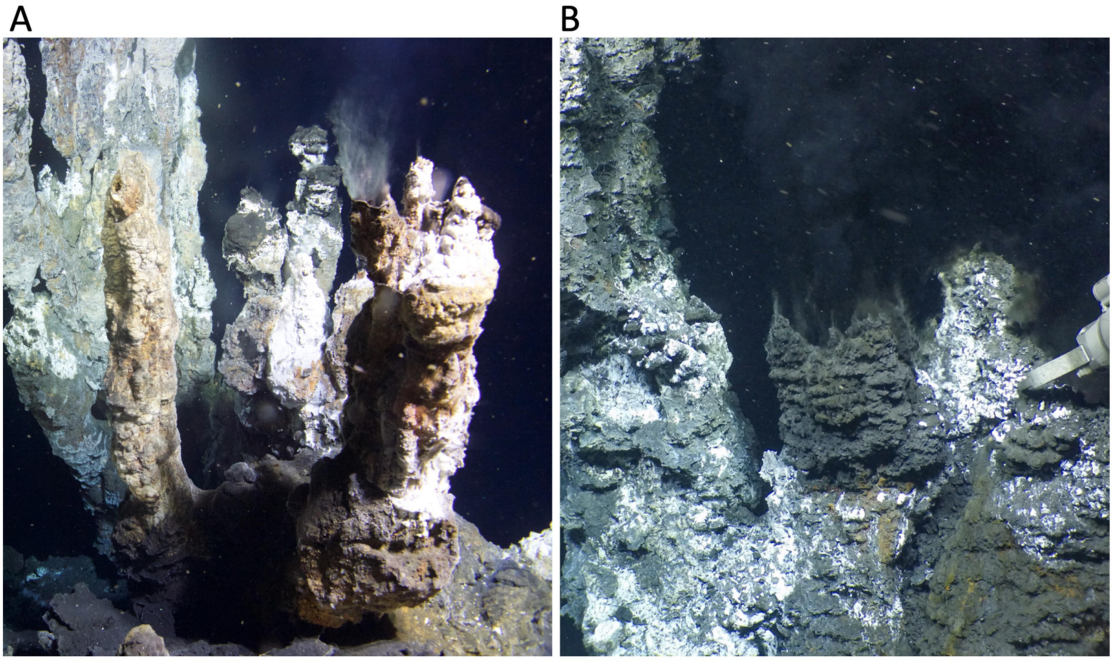

Fig. 2. Deep-sea hydrothermal vents located north of Iceland on the Arctic Mid-Ocean Ridge, in the Norwegian-Greenland Sea. White smokers at the Soria Moria vent field (A) and black smokers at the Loki´s Castle vent field (B) host diverse chemosynthetic microbial communities that derive their energy from the chemical disequilibria that form when reduced hot hydrothermal fluids, rich in potential electron donors (e.g., H_2, CH_4, H_2S, mix with cold seawater rich in potential electron acceptors (Dahle *et al.* 2015; Steen *et al.* 2016). Photos: ÆGIR team, NORMAR.

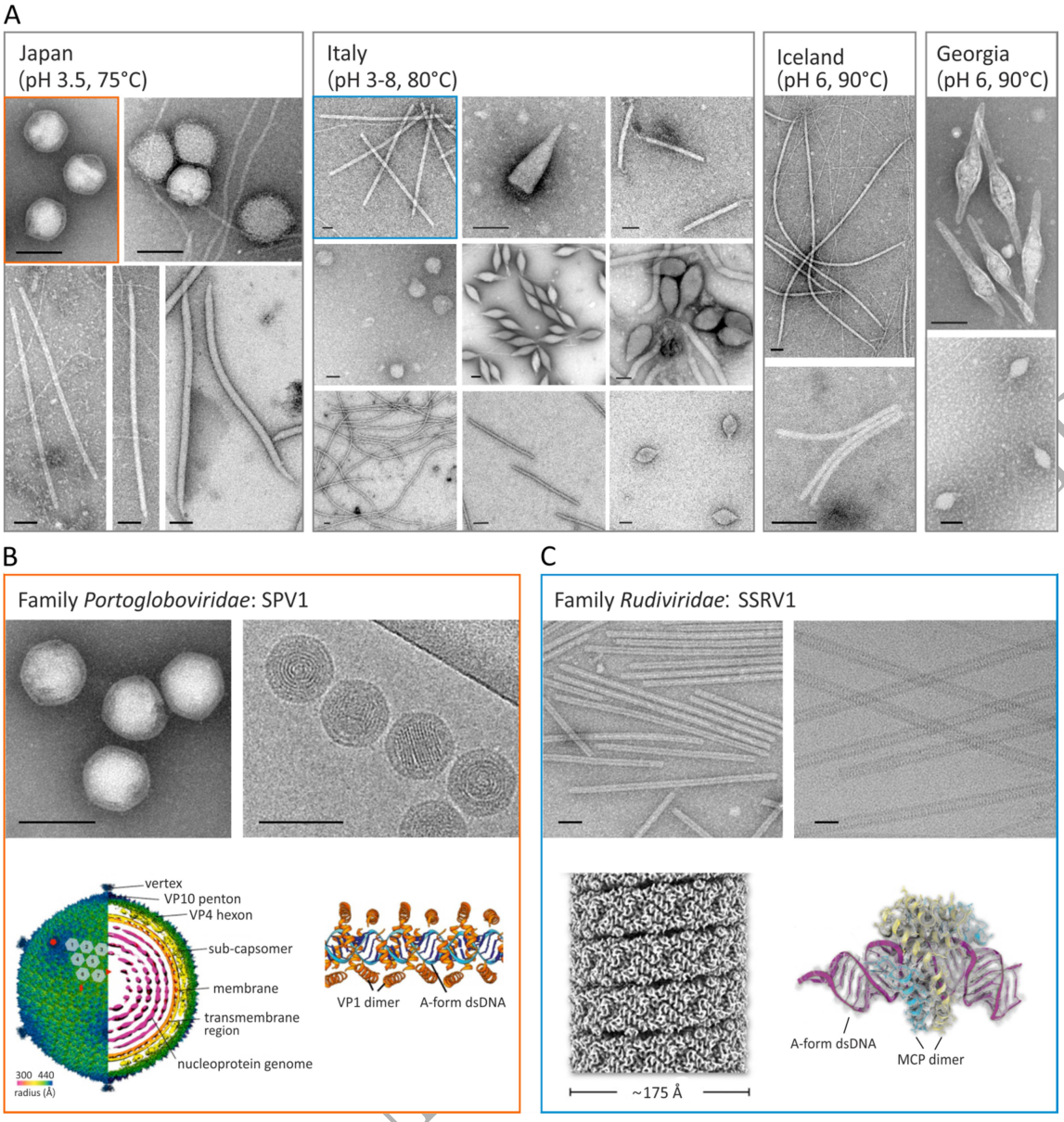

Fig.3. Diversity of virion morphotypes observed by electron microscopy from samples collected from hydrothermal environments at indicated locations. (A) Electron micrographs of virions in established enrichment cultures with denoted pH values and temperatures similar to that of their original environments. Pictures of the samples originated from Japan and Italy are adapted from Liu *et al.* (Liu *et al.* 2019) and Baquero *et al.* (Baquero *et al.* 2020), respectively. (B) and (C) The cryo-electron microscopy reconstruction of the new viruses isolated from enrichment cultures (framed in the same colors as in (A). SPV1 pictures are adapted from Liu *et al.* (Liu *et al.* 2017) and Wang *et al.* (Wang *et al.* 2019a). Virion components, including structural protein complex, inner membrane, and nucleoprotein genome consisting of A-form dsDNA and virion protein (VP) 1 dimers, are indicated. SSRV1

images are adapted from Wang *et al.* (Wang *et al.* 2020b) and Baquero *et al.* (Baquero *et al.* 2020). The helical structure of the virion that is composed of A-form dsDNA and the major capsid protein (MCP) homodimers are shown. Bars in electron micrographs, 100 nm.

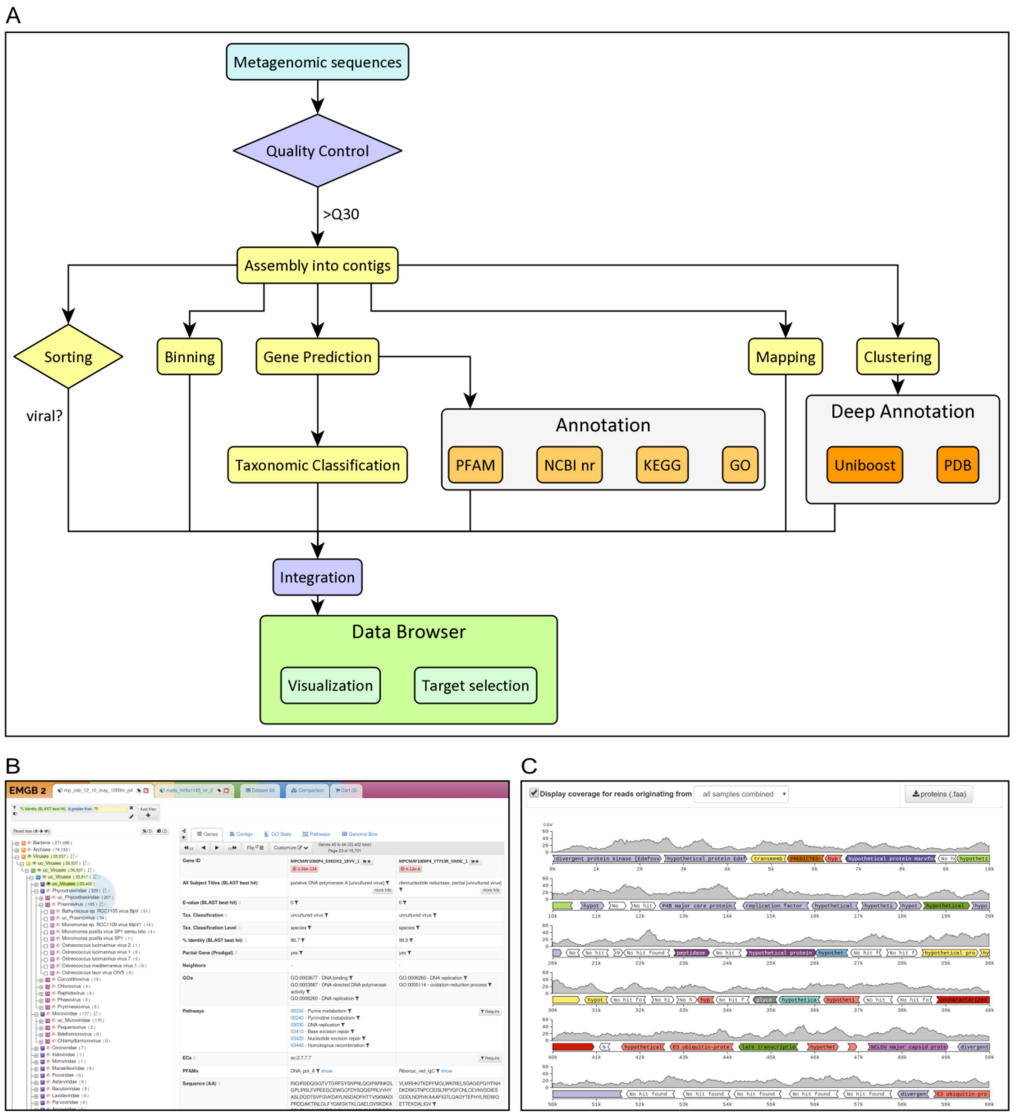

Fig. 4. Core elements of the bioinformatics workflow (A). The web-based project data browser EMGB (B). It integrates all workflow results and features tailored filtering and inspection tools for screening large metagenomic datasets. One particular important inspection tool is the contig viewer (C), which merges gene annotations and taxonomic assignments with reading coverage graphs, thus enabling a quick inspection of neighbouring genes and a comparison of coverage distributions across samples of a combined assembly.

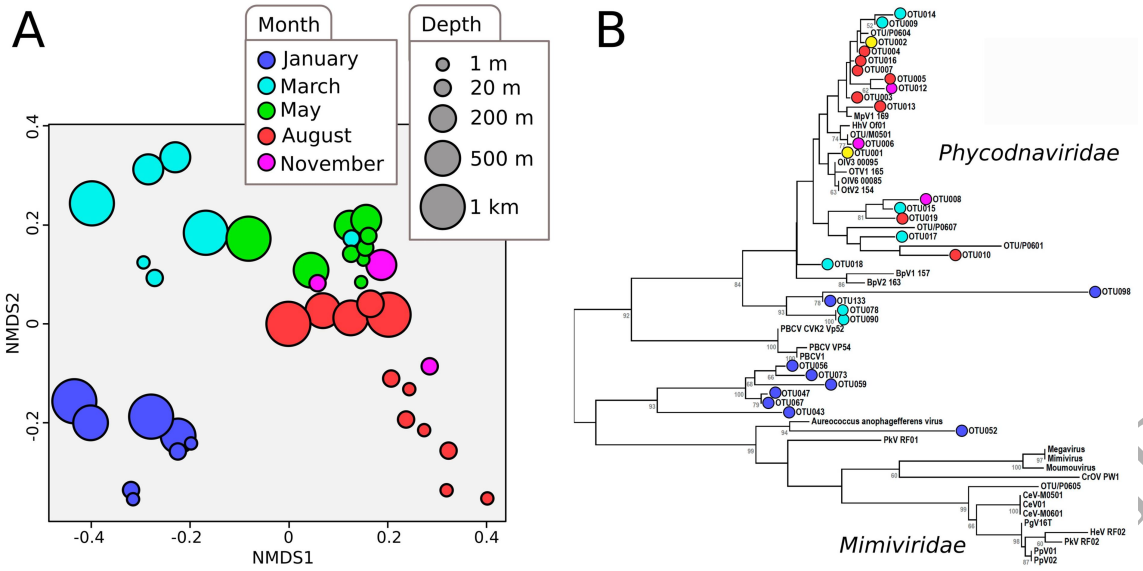

Fig. 5. Nonmetric multidimensional scaling (NMDS) analysis of OTU diversity for major capsid protein (gp23) of T4-like myoviruses (A). Sampling depths are represented by circle sizes and sampling month by colors. Bray–Curtis dissimilarity was used to compare OTU composition between samples. Maximum likelihood tree constructed from the 29 most abundant major capsid protein OTUs of giant algal dsDNA viruses (B). Each OTU is identified by a circle, colored by the sampling month in which the OTU was found as dominant. The two yellow circles represent cosmopolitan OTU, significantly present in samples from more than three different months.

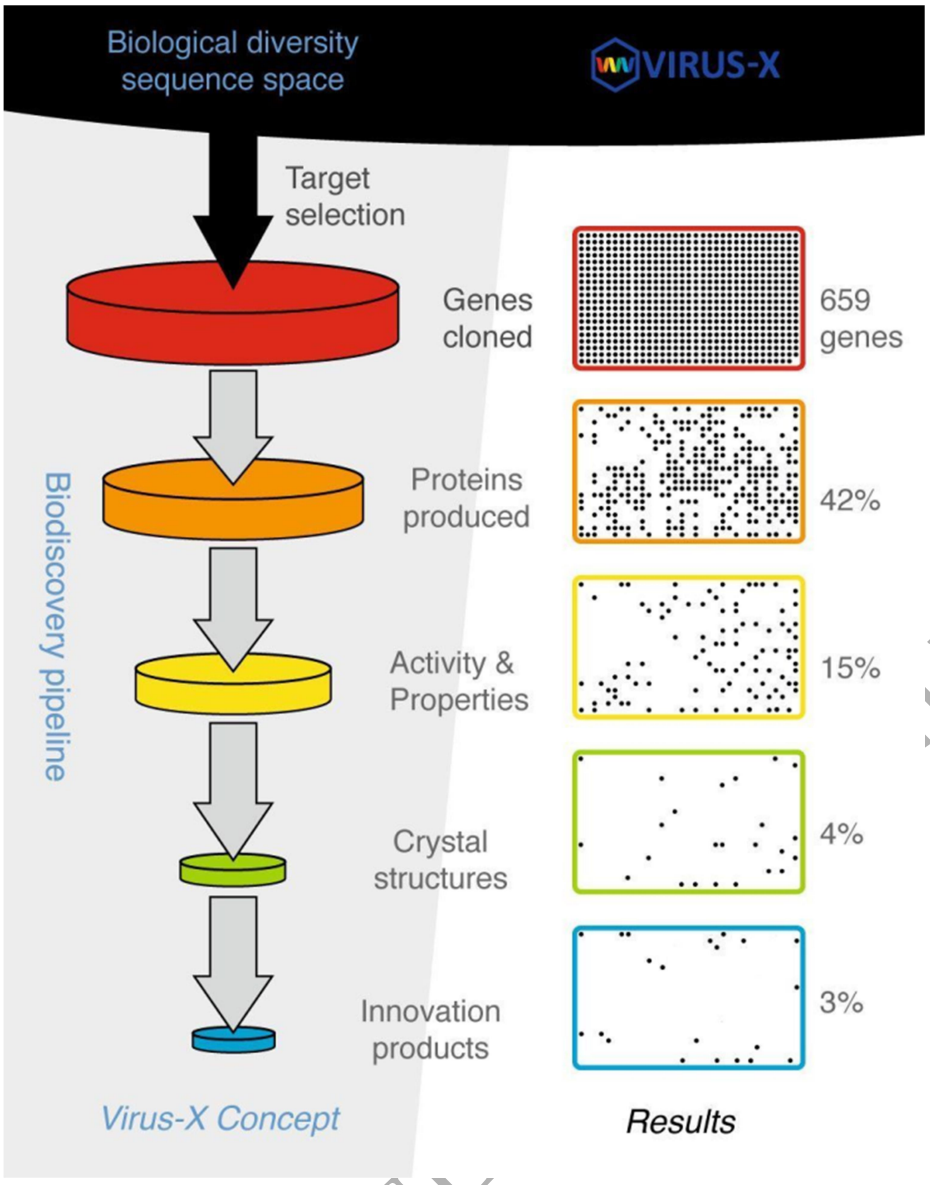

Fig. 6. The Virus-X project's conceptual framework is the biodiscovery pipeline with consecutive steps from biological diversity sequence space to innovations as seen on the left. The available exploration of sequences is provided by the metagenomic approach as well as isolated genomes of viruses. The concept is basically a funnel-shaped pipeline based on expected fall-off in numbers of targets as we move from one step to the next using, as a guideline, previously reported cases of comparable implementations such as seen in structural genomics projects (Christendat *et al.* 2000). The progress of the project and results, as seen on the right, reflects this expected outcome, and having large enough input into the pipeline ensures control of the final output and provides certain flexibility in terms of difficulties in moving a particular target from one step to another. An example was a persistent problem in

getting soluble proteins in expression experiments, apparently reflecting some inherent difficulties in expression of our viral protein targets. However, the funnel-shaped approach provided optimized use of efforts as challenging targets could be omitted in the process unless additional measures were justified for targets of particular value, such as a strong potential for becoming an innovation product, e.g., commercial enzyme.

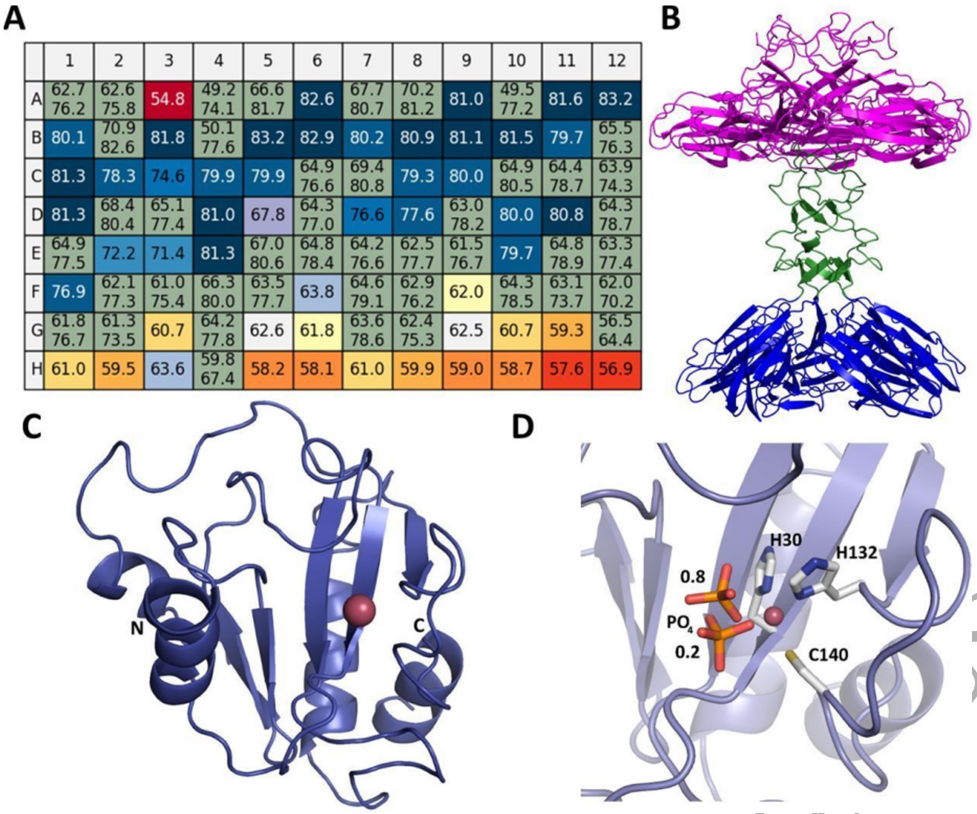

Fig. 7. Protein characterisation and crystal structure analysis. (A) Thermal shift analysis using the Durham pH screen to identify the optimal pH range and buffer type for protein storage of an exemplary single-stranded DNA binding protein (SSB). The melting temperature in the water of 76°C (A1 and A2) is increased at low pH to over 81°C (colour-coded by a change to darker shades to blue) and decreased to over 58°C at high pH over 9 (colour-coded from yellow to red). Data analysis was performed with NAMI (Groftehauge *et al.* 2014). (B) Ribbon diagram of the crystal structure of XepA determined at 2.1 Å resolution (Freitag-Pohl *et al.* 2019). The five N-terminal domains shown in magenta are presumably responsible for the antibacterial activity and are connected via a linker region (shown in green) to the C-terminal domains that have been suggested to connect to the corners of the viral capsids to aid its escape. (C) The crystal structure of Ph2119 endolysin of *Thermus scotoductus* MAT2119 bacteriophage Ph2119 determined to 1.2 Å resolution (PDB entry: 6SU5) with Zn^{2+} (pink sphere) in a conserved peptidoglycan binding site (Plotka *et al.* 2020). This enzyme represents the first thermostable endolysin with amidase activity and shows a similar architecture to eukaryotic peptidoglycan recognition proteins (PGRPs) and T7 lysozyme. (D) Close view of the Zn^{2+} binding site with coordinating side chains (H30, H132, and C140) and a phosphate

group (in sticks representation) shown in two positions with occupancies of 0.8 and 0.2, as indicated by the numbers in the figure.

Table 1. The Virus-X metagenomics toolbox includes freely available databases except for EMGB, free, open-source software, and a user-friendly webserver**.**

The Virus-X metagenomics toolbox	
PHROGs	database of deeply clustered and well-annotated viral protein groups (Terzian *et al.* submitted) https://phrogs.lmge.uca.fr
Uniclust	database providing functionally homogenous clusters of sequences at clustering depths of 90%, 50%, and 30% sequence identity (Mirdita *et al.* 2016); https://uniclust.mmseqs.com
Soil | Marine Reference Catalogues	catalogues of 2 billion proteins from soil metagenomes (SRC) and 300 million proteins from marine metatranscriptomes assembled (MERC) using Plass http://wwwuser.gwdg.de/~compbiol/plass
BFD	database of over 65 million protein profiles based on combined and clustered 2.5 billion protein sequences from SRC, MERC, Metaclust, and Uniprot https://bfd.mmseqs.com
WIsH	prediction of the prokaryotic host of viral contigs as short as3 kb that runs several hundred times faster than standard bioinformatic implementations (Galiez *et al.* 2017); https://github.com/soedinglab/wish
Linclust	first protein clustering algorithm that scales linearly with the input, clusters billion of protein sequences within a single day on a single server (Steinegger and Söding, 2018); https://github.com/soedinglab/mmseqs2
HH-suite3	software suite for fast remote homology detection and deep protein annotation (Steinegger *et al.* 2019); https://github.com/soedinglab/hh-suite
Plass	assembles short-read sequencing data on a protein level. Plass is able to extract ten times more proteins from complex metagenomes (Steinegger *et al.* 2019); https://github.com/soedinglab/plass
MetaEuk	enables large-scale eukaryotic metagenomics through reference-based, sensitive taxonomic and functional annotation (Levy Karin *et al.* 2020); https://metaeuk.soedinglab.org

SpacePHARER	sensitive identification of phages from CRISPR spacers in prokaryotic hosts (Zhang *et al.* 2021); https://spacepharer.soedinglab.org
MMseqs2 software suite	software suite to search and cluster huge protein and nucleotide sequence sets (Steinegger and Söding, 2017), sensitive taxonomy assignment (Mirdita *et al.* 2021), and user-friendly webserver (Mirdita *et al.* 2019); https://github.com/soedinglab/mmseqs2 https://search.mmseqs.com
EMGB	web-based metagenome annotation browser for large datasets (Henke *et al.* manuscript in preparation) https://github.com/metagenomics/EMGB2