MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"Evaluation and Improvement of Annotations of Virus Orthologous Groups"

verfasst von / submitted by

Sigrid Koizar

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik UG2002
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

“42.”

Douglas Adams, The Hitchhiker’s Guide To The Galaxy

UNIVERSITY OF VIENNA

Abstract

Centre for Microbiology and Environmental Systems Science, Universität Wien

Master of Science

Evaluation and Improvement of Annotations of Virus Orthologous Groups

by Sigrid KOIZAR

Viruses are among the most numerous biological entities on earth. Despite their wide abundance viral sequences are highly underrepresented in public databases. The Virus Orthologous Groups Database (VOGDB) clusters viral proteins obtained from NCBI RefSeq together, creating a database of remotely homologous proteins. The current consensus annotation of the groups is based on sequence similarities of VOG member proteins to the manually curated UniProtKB/Swiss-Prot database. The functional homogeneity of VOGs could be confirmed based on manually curated proteins assigned to the VOGs.

In this thesis, some potential issues with the current annotation, such as uninformative descriptions, transfer of functional information from cellular to viral proteins or annotations with proteins that are only partially covered by the alignments are highlighted. A new annotation approach applying the identification of manually curated proteins in the VOGs, remote homology search tools, as well as domain-based annotations is implemented. The validity of the new annotation pipeline could be verified by producing functional annotations matching those of literature-reviewed viral proteins.

As the VOGDB is heavily biased towards Caudovirales, potential marker gene VOGs for “Major Capsid Protein” and “Terminase, large subunit” were identified, however, universal viral marker genes could not be obtained due to the lack of diversity in the VOGDB.

UNIVERSITÄT WIEN

Abstrakt

Centre for Microbiology and Environmental Systems Science, Universität Wien

Master of Science

Evaluierung und Verbesserung der Annotierung Viraler Orthologer Gruppen

von Sigrid KOIZAR

Viren gehören zahlenmäßig zu den meist-repräsentierten biologischen Einheiten der Erde. Dennoch sind sie nur spärlich in öffentlichen Datenbanken anzutreffen. Die Datenbank viraler orthologer Gruppen (VOGDB) platziert entfernte homologe Virenproteine in Gruppen, die mithilfe von Sequenzähnlichkeitssuchen zu manuell kuratierten Proteinen in der UniProtKB/Swiss-Prot Datenbank annotiert werden. Die funktionale Homogenität wurde durch die Präsenz einheitlich annotierter Swiss-Prot Proteine in den VOGs bestätigt.

In dieser Arbeit werden potenzielle Problematiken der Annotierung untersucht. Diese beinhalten uninformative Funktionsbeschreibungen, Annotierungen mit nur partiell von der Alinierung abgedeckten Proteinen, sowie Funktionsübertragungen von Proteinen zellulärer Organismen. Ein neuer Ansatz zur Funktionsvorhersage, basierend auf der Identifikation von manuell kuratierten Proteinen in den Gruppen, Suchstrategien für entfernte Homologien, sowie einer Domän-basierten Annotierung wurde implementiert. Die Validität der neuen Methodik konnte durch virale Proteine, welche in Publikationen beschrieben sind, verifiziert werden.

Da die VOGDB in erster Linie aus Proteinen der Ordnung Caudovirales besteht, konnte sie nicht als Ressource für universale virale Markergene herangezogen werden. Als “Hauptkapsidprotein” oder “Terminase, große Untereinheit” beschriebene VOGs wurden als caudovirale Markergene identifiziert.

Acknowledgements

I would like to thank all of those who have helped and encouraged me during my work on this thesis. I am thankful to have been supervised by Univ.-Prof. Mag. Dr. Thomas Rattei, who gave me the chance to work on a project that was dear to me. Many thanks to Lovro Trgovec-Greif for always answering to my emails without the slightest delay and for some interesting discussions about the VOGDB. Special thanks to my dad for countless hours on the phone, educating me on runtime issues as well as giving me advice in other ways related to the technical aspects of this project. I am tremendously thankful for my mother, who continues to support me and for being my most frequent visitor, wherever I end up living during the year. I would also like to thank Raphael Bednarsky and Nikola Vinko with whom I have frequently worked on group projects throughout the past few years and who continue to be some of the best in discussing Bioinformatics-related topics.

Contents

List of Figures

List of Tables

List of Abbreviations

BC	Baltimore Classification
BLAST	Basic Local Alignment Search Tool
COG	Cluster of Orthologous Genes
eggNOG	evolutionary genealogy of genes: Non-supervised Orthologous Groups
HCV	Hepatitis C Virus
HGT	Horizontal Gene Transfer
HIV	Human Immunodificiency Virus
HMM	Hidden Markov Model
ICNV	International Committee on Nomenclature of Viruses
ICTV	International Committee on Taxonomy of Viruses
KEGG	Kyoto Encyclopedia of Genes and Genomes
LCA	Last Common Ancestor
LUCA	Last Universal Common Ancestor
MSA	Multiple Sequence Alignment
MSV	Multiple Segment Viterbi
NGS	Next Generation Sequencing
OMA	Orthologous Matrix
pVOGs	Prokaryotic Virus Orthologous Groups
PSSM	Position-specific Scoring Matrix
PSI-BLAST	Position-Specific Iterated Basic Local Alignment Search Tool
RefSeq	Reference Sequence Database
SP	Swiss-Prot
SPP1	Bacillus subtilis Bacteriophage SPP1
VHG	Viral Hallmark Gene
VOG	Virus Orthologous Group
VOGDB	Virus Orthologous Groups Database

Dedicated to my family, who has always been there for me.

Chapter 1

Introduction

1.1 Viruses

Viruses are some of the smallest, yet most numerous biological entities on Earth. Due to their dependence on host organisms to replicate, there is an ongoing debate if viruses are to be considered living or non-living entities. Upon infection of the host cell, the cell's machinery is modified to aid in the production of new viruses. Every cellular organism that has been studied so far has its own viruses or virus-like selfish genetic elements [1]. Viruses infecting bacteria (bacteriophages) are of high abundance on the planet, their number being estimated to 1x10^{31}, and they are responsible for removing 20–40% of all bacterial cells each day [2].

1.1.1 Impact of viruses on their surroundings

A recent reevaluation of the question if bacteria outnumber human cells in the body estimated that the ratio of bacterial cells to human cells is close to 1:1 [3]. Even though the number of bacterial cells harbored within our bodies is 10x less than the previously reported 10:1 ratio, they are of utmost importance to our existence, assisting in digestion and defense [4]. While prokaryotes are persistent companions of macro-organisms, they are in turn constantly infected by bacteriophages. The number of viruses in the human body is estimated to 380 trillion, hereby vastly outnumbering bacterial and human cells [5]. The majority of them are bacteriophages and do not infect the macro-organism's cells. By playing key roles in the biology of microbes, viruses indirectly impact the host, or, more generally speaking, the environment, when infecting bacteria not harbored by living organism but found in soils, water and other habitats [5]. The human virome has not yet been extensively studied, as its importance was not appreciated until about a decade ago, but there have already been advances in applying the ability of phage lysins to kill bacterial cells with antibiotic resistance [6].

While it is acknowledged that viruses influence their environment by killing their hosts, their impact on evolution is often overlooked. They force their hosts to evolve to escape their virulence, but also serve as vectors for horizontal gene transfer (HGT) between different species. Bacteria are known to benefit from HGT, but the importance of HGT in animals has not been studied in depth. In recent years Verster and colleagues showed that prokaryotic toxin genes have been found in insects and are most closely related to orthologs from a bacteriophage infecting a bacterium aiding aphids in their defense against parasitoid wasps [7]. With ongoing research into viral communities, both positive and detrimental effects of viruses on their surroundings will be studied in greater detail.

1.1.2 Diversity of Viruses

Viruses exhibit a high level of diversity, not only considering the wide range of hosts they infect, but also their internal make-up, from genome size and architecture over structural properties to their mode of replication. Structurally viruses can be classified as having helical morphology, where capsid proteins are wrapped around a helical filament of nucleic acid, or icosahedral morphology, having a more spherical shape [8, 9].

Genome sizes vary the most in DNA viruses, where the genome sizes range over four orders of magnitude. Their genomes can be of high complexity with sizes as large as several megabases. Most large viruses such as herpes- or poxviruses belong to the dsDNA viruses [10, 11]. Another example of a giant virus is the Mimivirus, the name stemming from "mimicking microbes" due to their large size as well as their Gram-staining properties [12, 13]. It was discovered in 2003 when researchers initially thought they had identified another bacterium infecting amoebea, however, the absence of ribosomes revealed that the large coccus-like structures were in fact virions [14, 15]. Since then more members of the nucleocytoplasmic large DNA virus (NCLDV) supergroup have been identified [12, 16–20].

RNA viruses show a smaller divergence of genome sizes; their genomes' size is generally limited by the tendency of single stranded RNA strands to break more easily and by the higher rate of mutations [10]. Some of the smallest viral genomes belong to bacteriophage MS2 (3569 nucleotides, ssRNA), or, as a representative of eukaryotic viruses, the Porcine circovirus (1726 base pairs, DNA). The bacteriophage MS2 genome encodes only four proteins: the maturation protein (A-protein), the lysis protein, the coat protein, and the replicase protein. The genes are overlapping, permitting the small genome, and protein expression is regulated by RNA secondary structure and translation [21]. The Porcine circovirus similarly only encodes one capsid protein and two replicase proteins, for which one of them is the truncated version of the other [22]. Viral genomes can be made of DNA or RNA, and they can be single-stranded or double-stranded. The different viral genome types are discussed in more detail in The Baltimore Classification of Viruses.

1.1.3 Viral Proteins

As viruses exhibit high diversity and genome sizes vary greatly even within classes, the number of encoded proteins also differs between viral species. Some dsDNA viruses encode thousands of proteins. Examples include the Mamavirus with 1023 predicted proteins, as well as other Mimivirus strains [23]. On the contrary, some small-genome RNA viruses only encode a few proteins. Examining the protein lengths of 16,331 viral proteins from the 5,152 publicly available viral reference proteomes in Swiss-Prot (SP) [24] showed that proteins from dsDNA viruses are mostly composed of 2000 amino acids or less, while the protein sizes in ssRNA viruses vary greatly. An offset towards large proteins with more than 6000 amino acids has been observed [25]. These large proteins likely encode polyproteins. Polyproteins are collections of various functional protein domains that are initially expressed as a single protein, before being cleaved by viral and/or cellular proteases into their distinct functional units. They occur in most single-stranded RNA viruses, some double-stranded RNA viruses and retroviruses with polycistronic genomes [26]. This strategy of genome organization allows for a condensation of the genome by using a single set of transcriptional and translational control elements to produce several proteins required for viral infection and it allows for regulation based on the cleavage state of the polyprotein [27, 28].

1.1.4 Viral Phylogenetics

An Introduction to Phylogeny

Phylogeny is the study of the evolutionary relationships of species. While in the early days phylogenies were depicted based on apparent similarity, newer methods quantify genome and protein sequence alignments to depict a tree [29, 30]. As mistakes are prone to happen during replication of the genome, nucleotides can be substituted, deleted or inserted, leading to mutations. Mutations can also be caused by environmental factors such as radiation or chemicals [31]. Such modifications can be neutral, harmful or advantageous. If mutations are advantageous, the lineage is more likely to succeed, and as mutations accumulate the lineages diverge over time. The depicted trees are hypotheses based on the data and only approximate the true phylogeny which remains unknown. When building phylogenetic trees it is crucial to distinguish between homologies, which are similarities as a result of common ancestry and analogies, or similarities caused by the organism responding to similar environmental conditions [29].

All cellular organisms can be placed into a phylogenetic tree as they share a last universal common ancestor (LUCA). This tree including the totality of cellular species is known as the Tree of Life [1]. Initial approaches of phylogeny used ubiquitous prokaryotic 16S RNA sequences and their eukaryotic 18S RNA counterparts (both are small subunit rRNAs, or ssu-rRNAs) to construct tree topologies [32–35]. This approach was later replaced with the comparison of whole genomes, as the evolutionary history of a single gene does not necessarily represent the evolution of a species (see Microbiome Analysis). Evolutionary tree reconstruction is further complicated by horizontal gene transfers between species. It is possible that trees reconstructed by comparing different orthologous genes are substantially different, proving that the concept of the Tree of Life is not as straight-forward as researchers originally assumed it to be [36].

Viruses and the Tree of Life

While cellular life forms have a LUCA, viruses have multiple ancestors, i.e. they are polyphyletic. Evidence is provided by the diversity of genome architecture types as described above. As recent studies demonstrated, there are several connections between viruses of different Baltimore classes [37]. Before the abundant availability of molecular data viruses could not be phylogenetically classified because of several reasons. There were no fossils of ancient viruses and their quick evolution driven by high mutation rates, lack of proofreading, short generations and pressure from co-evolution of their hosts led to quick divergences. Additionally, as viruses evolve alongside their hosts, there are frequent gene transfers over time [30, 36]. As new techniques enabled the fast and cheap collection of large amounts of sequencing data with next generation sequencing (NGS), viral phylogeny also became a more active area of research. Due to the lack of universal marker genes such as the ribosomal RNA [38, 39] or mitochondrial DNA [40] characterizing complete viral phylogeny has been an unsolved task up to date, even though some small-scale phylogenetic relationships have been determined for well-studied lineages with practical importance, e.g. epidemiology and diagnostics. Such well-studied viruses include, but are not limited to: influenza virus, human immunodeficiency virus (HIV), hepatitis C virus (HCV) and poliovirus [30].

The Baltimore Classification of Viruses

In 1971 virologist David Baltimore first proposed the classification of viruses into six groups, based on their genome type. Later a seventh group was added to the system which now includes: double-stranded DNA, single-stranded DNA, double-stranded RNA, positive-sense RNA, negative-sense RNA, reverse-transcribing RNA and reverse-transcribing DNA. Baltimore based his classification on the virus's way of following the central dogma of molecular biology, describing the flow of information from the genetic material over mRNA to encoding the amino acid sequence of proteins. Some classes have to undergo additional steps such as reverse transcription (groups VI and VII) in order to synthesize mRNA (Figure 1).

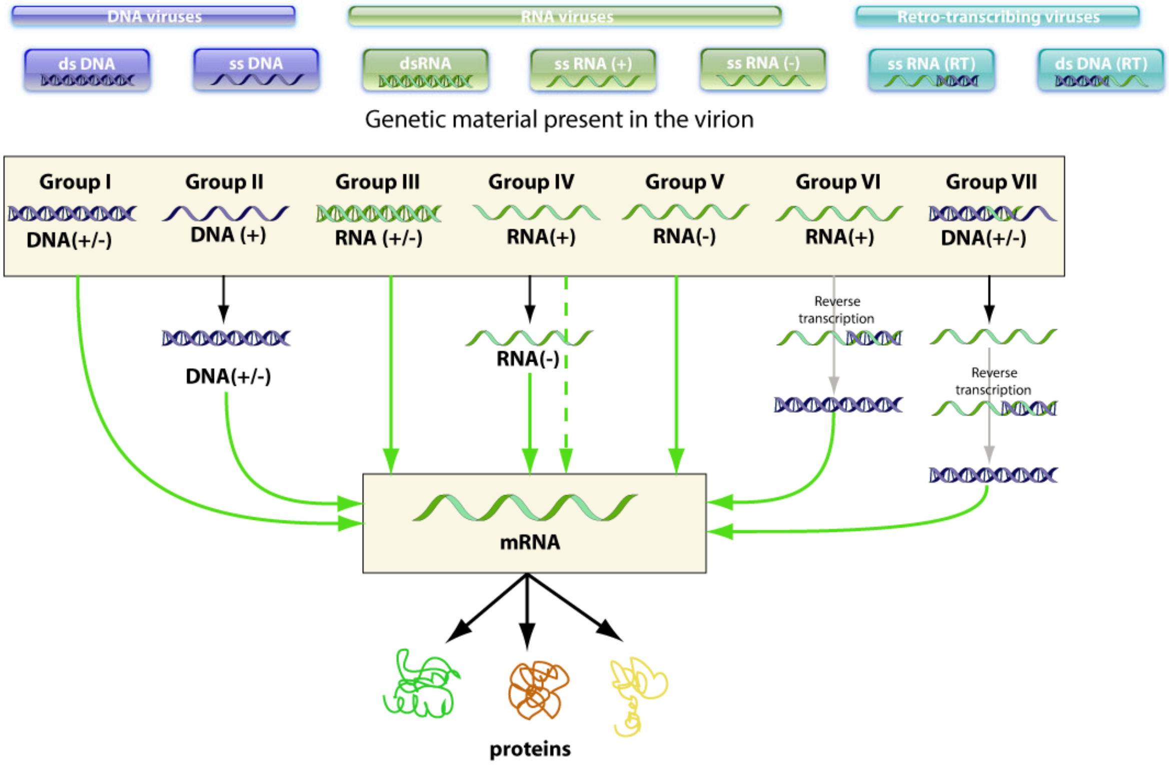


Figure 1: Baltimore Classification of Viruses. Viruses are divided into seven groups, based on their genome type and how mRNA is produced. Figure from [41].

Because of the lack of a LUCA for viruses, the Baltimore Classification (BC) can be looked at as seven distinct trees of life, even though the monophyleticity of the individual classes has been challenged by recent studies of metavirome data, mainly concerning RNA virus taxonomy [42–45]. Another potential pitfall of the BC is that evolutionary relationships are not considered. Nevertheless the BC system provides an uncomplicated and useful overview of viral classification and will likely continue to exist.

Virus Taxonomy

The increase in data from metagenomic sequencing has led to a tremendous increase in the number of newly discovered viruses. With this development the need for an official virus taxonomy emerged. In 1966 the International Committee on Nomenclature of Viruses (ICNV) was called into life to develop a system for naming taxa and classifying viruses. Later the ICNV became the International Committee on Taxonomy of Viruses (ICTV), which until today regularly publishes reports and updates. The first classification structure was published in 1971 and included five taxonomic ranks [46]. Viruses were classified based on biological properties such as in vitro properties, structure and antigenic relationships as well as on host factors including host range, pathogenicity and epidemiology [47]. With the discovery of more viral genomes through increased high-throughput sequencing, there was an increased need to extend the 5-rank-structure to accommodate newly found viruses with no experimental classification.

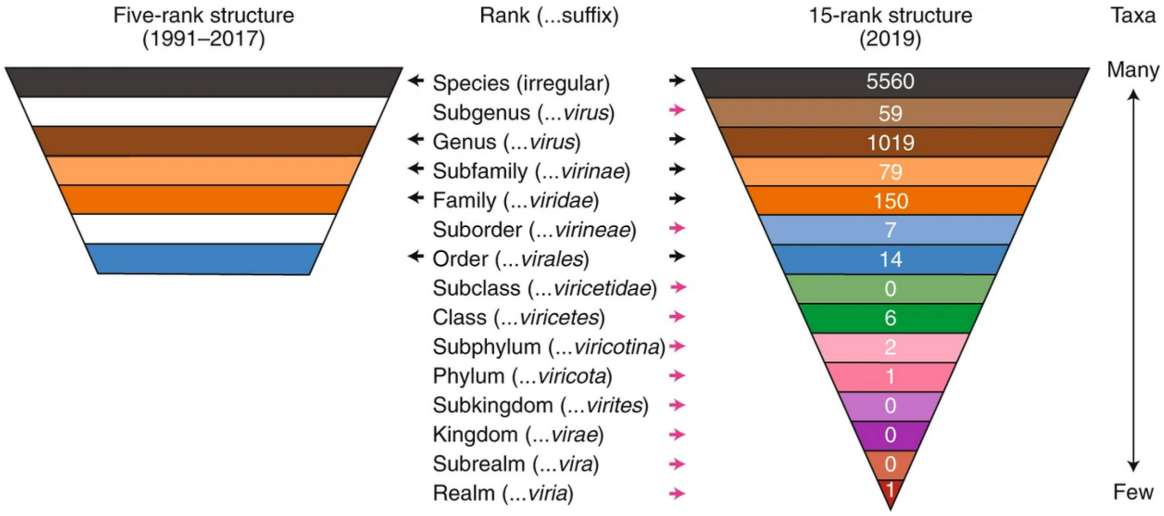


Figure 2: Viral Taxonomy. Taxonomic ranks are shown in relation to the distribution pattern of taxa. The number of taxa assigned to each rank (as recorded in the current ICTV Master Species List, release 2018b, MSL34) are shown in white font on the 15-rank structure. When the ranks are described as a hierarchy, the species rank is often referred to as the lowest rank and the realm rank as the highest rank. However, when the ranks are used as phylogenetic terms, the realm rank can be described as basal and the species rank as apical or terminal. Both conventions are used in this Consensus Statement. Black arrows, ranks common to the five- and 15-rank structure; pink arrows, ranks introduced in the 15-rank structure. Figure taken from [48].

In 2018 the classification system was extended to 15 ranks (Figure 2), including eight principal (or primary) ranks and seven derivative (or secondary) ranks [42]. The new system infers biological properties from phylogeny and homology detection, and in comparison to the old system it does not rely on biological data, but can be applied on metagenomic data [47]. With this adaptation viral taxa can now be accommodated at every level and supergroups and superfamilies that previously could not be placed into the system now populate a rank.

1.1.5 Viral Dark Matter

Most of the publicly available sequence databases lack adequate representation of viral sequences. Even though there has been an increase in viral sequences in databases, there is a large bias towards mammalian, plant and bacterial viruses [50]. Recent projects have investigated soil [51, 52] and ocean [53–57] viromes, expanding the known viral sequence space. Some metagenomic samples contain up to 90% of sequences with no homologs in public databases [58–61]. These sequences are termed "dark matter". Figure 3 provides an overview of metagenomic sequence classification; sequences that can be aligned to neither a nucleotide nor an amino acid database fall into the pool of "dark matter". In addition to the limited representation of viruses in reference sequence databases, other factors such as the divergence and length of virus sequences, as well as the limitations of alignment-based classification contribute to high amounts of viral dark matter [49].

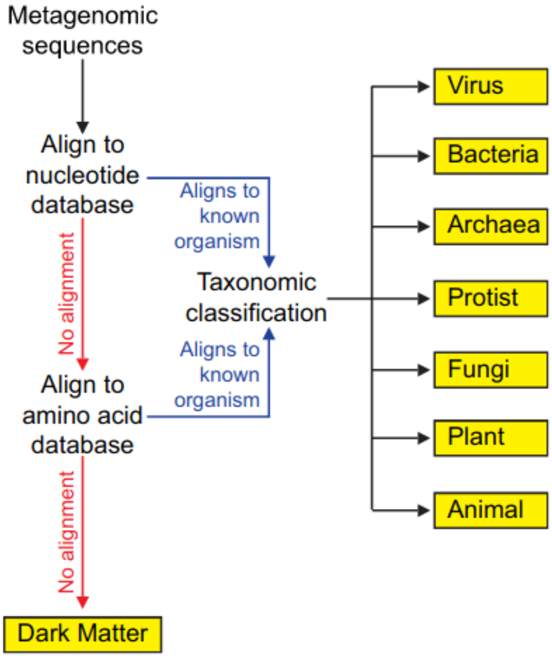


Figure 3: Viral Dark Matter. Generic overview of alignment based sequence identification. Image from [49].

1.2 Protein Annotation

The rapid increase in sequencing data called for the need to functionally describe gene products. Several pipelines have been developed to annotate entire genomes, such as Prokka [62], the NCBI's eukaryotic genome annotation pipeline [63] and prokaryotic genome annotation pipeline (PGAP) [64], as well as the Influenza Virus Sequence Annotation Tool [65]. Virus-specific annotation pipelines include Viral Annotation Pipeline and iDentification (VAPiD) [66], Vgas [67] and Viral Genome ORF Reader (VIGOR) [68], the latter being developed for gene prediction in influenza virus, rotavirus, rhinovirus and coronavirus subtypes. While these tools start from whole genomes and thus have to predict genes before annotating the proteins, others such as InterProScan [69] take a protein sequence as input. During the protein annotation step most approaches rely on BLAST (Basic Local Alignment Search Tool) [70], DIAMOND (double index alignment of next-generation sequencing data) [71], or HMMER [72] as search tools [73]. A hierarchical approach of annotation is taken by Prokka, which first tries to infer function by aligning the query sequence to a curated protein sequence database, which can be defined by the user. If no hits are returned, similarity searches are performed against Uniprot [24] and RefSeq [74] databases, and lastly against HMM databases including Pfam [75] and TIGRFAM [76].

1.3 Evolutionary Genomics

1.3.1 Evolutionary Genomics Definitions

In order to understand evolutionary genomics, the concepts of analogy, homology, orthology, and paralogy are crucial. Analogous genes perform a similar function due to convergent evolution, but they are not related by ancestry. Homologous genes derive from a common ancestor, but they do not necessarily perform the same function. Homologs can further be classified into orthologs and paralogs. Orthologous genes originate from an ancestral gene in the last common ancestor (LCA) of the compared genomes, and they often perform the same or similar functions. Paralogous genes are genes that are related via duplication events, and they can further be classified into in- and outparalogs, depending on the duplication occurring before (outparalogs) or after (inparalogs) a speciation event [78]. As the genes are now present in two copies, one copy can evolve to exhibit novel functions, without being detrimental to the organism. Figure 4 provides an overview of orthologs as well as in- and outparalogs. The blue lines connect the inparalogous genes A′ and A″ found in the same species, the green lines connect outparalogous genes in different species and the purple lines depict orthologous relationships. Remote homologs are homologous genes that often lack easily

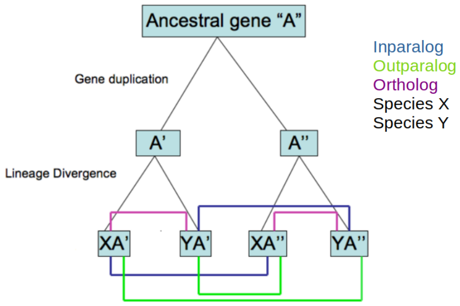


Figure 4: Diagram depicting evolutionary relationships between orthologs, inparalogs and outparalogs. Figure adapted from [77].

detectable sequence similarity, but nevertheless possess similar structures and functions. Table 1 summarizes important evolutionary genomics definitions.

Table 1: Evolutionary Genomics Definitions

Analogs	Unrelated genes with similar functions due to convergent evolution
Homologs	Genes sharing a common ancestor
Paralogs	Genes related via a duplication event
Inparalogs	Paralogous genes resulting from duplication(s) subsequent to a given speciation event
Outparalogs	Paralogous genes resulting from duplication(s) preceding a given speciation event
Orthologs	Genes origination from a single ancestral gene in the LCA
Remote Homologs	Homologs with low sequence identity

1.3.2 Applications of Orthology

There have been various applications of gene orthology. As genes derive from a common ancestor, they are likely to share the same function. If a novel gene is predicted to be orthologous to a known gene, the function can be transferred, aiding in the characterization of newly sequenced genomes. Another application of orthology is the identification of an appropriate model system for a given physiological problem, i.e. choosing a suitable model organism. Orthology also plays a role in evolutionary genomics, resolving species phylogenies and gene family-level evolution and adaption [79]. Recently the ortholog conjecture [80], i.e. the idea that orthologous genes share greater functional similarity than do paralogous genes, has been challenged by the absence of evidence of orthologs being more functionally similar than paralogs of equivalent levels of protein divergence [81]. Nevertheless, many algorithms aim to maximize the number of orthologous hits, while reducing the number of paralogous hits.

1.3.3 Viral Orthology

When talking about viruses the lines between the above terms become somewhat blurred, as there is no universal ancestor for all viruses. To avoid possible misclassification of paralogous or orthologous relationships, the superordinate term “homolog” can be used. However, throughout the rest of this work, the word “orthologs” will be used in accordance with the terminology of the Virus Orthologous Groups Database (VOGDB).

1.3.4 Databases of Orthologous Groups

The first well-known database of orthologous groups of proteins was the Clusters of Orthologous Genes (COG) Database, which was released in 1997 with the aim to aid in the study of protein function and evolution. Since the last update it contains 1187 bacterial and 122 archaeal genomes encoding 3,213,196 proteins that are grouped into 4,877 COGs [82, 83]. The algorithm used to create COGs is based on the assumption that orthologs have a higher sequence similarity to each other than to other proteins, and in an all-against-all comparison of all proteins in a pool they will be each other's best hit. They are thus said to be bidirectional best hits or symmetrical best hits (SymBets). To account for in-paralogs, SymBets from proteins of the same genome are clustered together. Subsequently a graph is constructed with proteins forming the nodes, and SymBets forming the edges. Nodes that create a triangle are considered a minimal COG, and these minimal COGs are clustered together with the EdgeSearch algorithm if their subgraphs share a common edge [25]. Due to the lack of viral genomes in the COG database, an additional database containing prokaryotic viruses was created. pVOGs (Prokaryotic Virus Orthologous Groups) are constructed with the COG framework. Up until now the pVOG database only contains viruses infecting bacteria and archaea, with eukaryotic viruses likely being added in the future [84]. Since then other databases of orthologous groups have been published, including the Orthologous Matrix (OMA) Database [85], the OrthoDB [86], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [87], and the evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) Database [88]. While some of the above mentioned databases do not include viral proteins at all (COG, OMA), others contain only phages (pVOG), or a mix of proteins stemming from all forms of life (eggNOG, OrthoDB, KEGG). The Virus Orthologous Group Database (VOGDB) contains only viral proteins. Table 2 provides an overview of several databases and information about the numbers of species from different domains of life and viruses included in the databases, as well as the number of orthologous groups and total genes.

Table 2: Databases of Orthologous Groups with summary statistics

Database	Host	Eukaryotes	Prokaryotes	Viruses	Orthologous Groups	Total Genes
EggNOG (v5.0)	EMBL	447	4643	2502	4.4M	n.d.
OrthoDB (v10.1)	Swiss Institute of Bioinformatics	1271	6013	6488	8.5M (29063 with viral proteins)	37M
COG (2021)	NCBI	"KOGs"	1309	–	4877	3.2M
OMA (2021/04)	ETH Zurich	2424		–	1.04M	17M
KEGG Orthology (2021/12)	Kyoto University	678	7033	355	24839	39M
pVOG (2017)	University of Iowa	–	–	3000 (phages)	9518	0.3M
VOGDB (v208)	CUBE, University of Vienna	–	–	10046	28386	0.439M

1.4 An Introduction to Homology Search

With the rapid explosion of available sequences in ever-growing DNA and protein databases, there has been an increased need to design methods that are both fast and sensitive when searching for sequences similar to a query sequence. In this chapter, various homology search programs and approaches will be described. Figure 5 displays how confidently homologs can be inferred based on sequence length and sequence identity. As viral sequences are short and sequence identities are low, few viral homologs fall into the safe zone. Over the years more sensitive homology detection methods have been developed. The following provides an overview of common tools.

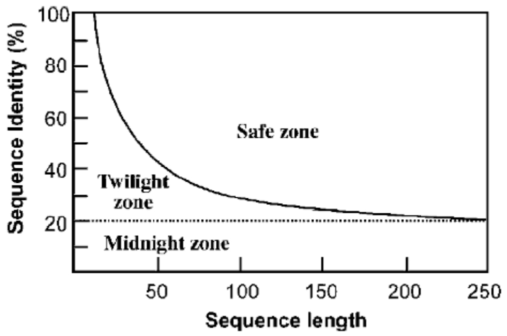


Figure 5: The three zones of protein sequence alignments. Protein sequences can be considered homologs if the percentage of sequence identity falls in the safe zone. Sequence identities above 20% but not in the safe zone are said to be in the twilight zone, where homologous relationships are less certain. Below 20% sequence identity, the homologous relationships are not reliable. Figure from [89].

1.4.1 BLAST

Basic Local Alignment Search Tool (BLAST) [90] is a widely used sequence similarity search tool. It was first introduced by the NCBI in 1989, and has undergone several updates since then, increasing speed and flexibility [91]. A major advantage in speed when compared to the Smith-Waterman algorithm for local sequence alignment stems from BLAST using initial seed alignments with small word sizes (default word size for BLASTP is 3). The local alignment is then extended from the region of a word hit in the query and target sequence. Alignment scores are calculated on the go and the extension of the alignment is stopped once the score falls below a threshold in low similarity regions. The heuristic approach of utilizing seed alignments increases the speed of the alignment algorithm at the cost of sensitivity, when compared to the exhaustive Smith-Waterman approach, in which the optimal alignment is produced. However, with the great abundance of data (>500.000 sequences in the UniProtKB/Swiss-Prot DB [24]), users rely on heuristic methods such as BLAST in the practical application of sequence similarity searches.

Even though pairwise alignments produce reliable results in terms of finding significant homologs, the main shortcoming is that a lower number of homologs is found compared with other methods based on models of protein families such as PSI-BLAST or HMM-based methods [92].

1.4.2 PSI-BLAST

Position-Specific Iterated BLAST (PSI-BLAST) [93] is more sensitive than a regular BLAST similarity search. From a single query protein input sequence, similar sequences are searched with BLASTP and aligned to produce a multiple sequence alignment (MSA). From this MSA a position-specific scoring matrix (PSSM) is calculated. After every iteration, newly found sequences are added to the MSA and the PSSM is adjusted. This is an improvement to BLAST, where a predefined scoring matrix (default: BLOSUM62) is used. The PSI-BLAST search is said to have converged when no new sequences are added to the MSA. One potential pitfall of PSI-BLAST is that once a non-homologous domain has been included in the MSA, the PSSM could become skewed and the probability of finding more proteins of the non-homologous family increases. The effects of an inaccurate first alignment can be reduced by supplying a MSA to build the first PSSM.

1.4.3 Hidden Markov Models

While PSSMs only represent substitutions, Hidden Markov Models (HMMs) also take insertions and deletions into account. HMMs, just like PSSMs, are built from MSAs of homologous proteins. They contain information about the amino acid frequencies, as well as the frequencies of insertions and deletions in each column of the MSA . The added information increases the sensitivity of homology searches and allows for the detection of remote homologs with low sequence identity [94]. Initially HMM-based homology searches were 100x slower than BLAST, which rendered them rather useless for wider applications. The introduction of software packages based on HMMs including HMMER [95] developed by the Eddy lab and HH-suite [94] from the Soeding lab has led to a more time-efficient application of HMM-based searches. Acceleration heuristics such as the multiple segment Viterbi (MSV) algorithm, in which an optimal sum of multiple ungapped local alignment segments using a striped vector-parallel approach is used, as well as the favoring of high-scoring MSV hits, are applied [72]. The hmmsearch tool from the HMMER packages allows searching of sequence databases with an HMM query. HH-suite is based on HMM-profile to HMM-profile comparisons, and does not allow the search in a sequence database. The hhblits tool from the HH-suite package is faster than the equivalent hmmscan tool from the HMMER package.

VOG HMMs: MSAs are created for each VOG with Clustal Omega [96]. To reduce the number of aligned sequences, sequences with identities of >90% are clustered with cdhit [97] prior to the sequence alignment. Subsequently, hmmbuild [72] builds an HMM profile from the MSAs. Both the MSAs and HMMs are accessible in the VOGDB.

1.5 Domain-based Annotations - InterProScan

InterProScan [69] is a widely used protein function prediction software package, combining several different member databases to classify proteins based on their family membership and domain architecture. It is used extensively in genome sequencing projects as well as by the UniProt Knowledgebase [24] to lay out a draft for protein function prediction. The integrated databases include Pfam [75], TIGRFAMs [76], SMART [98], PIRSF [99], PANTHER [100], HAMAP [101], Prosite [102], ProDom [103], PRINTS [104], CATH-Gene3D [105], and SUPERFAMILY [106]. InterPro is regularly updated, the latest version being released in

November 2021. InterPro release 87.0 contains 40,037 entries representing homologous superfamilies, families, domains, repeats and sites. A total of 34,917 GO terms have been mapped to InterPro entries [107].

InterProScan uses a variety of algorithms including BLAST [90] and HMMER [72] to search a sequence against multiple models. Hmmscan from the HMMER package helps in identifying protein domains by comparing the sequence against the HMMs built from family members of e.g. one Pfam family. Once domains have been identified, the internal Database is queried to find matching InterPro [108] entries and optionally Gene Ontology [109] (GO) terms can be added to the results. InterProScan automatically parses the output for overlapping regions or same clan matches. While some of the integrated Databases mainly focus on cellular genomes (TIGRFAM, SMART, ProDom) and model organisms (PANTHER), others such as the SUPERFAMILY, Prosite or Pfam also contain viral protein domains.

1.6 Microbiome Analysis

The community of microorganisms in a habitat is called the microbiome. Microbiomes can be analyzed with two main approaches. Upon sample selection and collection, the genetic material is isolated, and then either amplified with specific primers (marker-based approach), or sequenced completely (metagenomics approach) [110, 111].

1.6.1 Whole-genome-shotgun (WGS) metagenomics

In metagenomics analysis the totality of DNA from all organisms in a sample is sequenced, providing a large number of short reads [110, 111]. Genomes can then be assembled de novo or by being mapped to reference genomes. To facilitate de novo assembly, reads can be placed into bins based on coverage [112] or k-mer frequencies [113, 114]. Mapping viral reads to reference genomes is nearly impossible due to the large amount of sequences without homologs in databases (see Viral Dark Matter). Therefore species in the sample that are not well-represented in the reference-database are likely to be missed [115], and with the current, rather poor representation of viral species in public databases, the mapping approach will remain unfeasible. De novo assembly of viral genomes is also troublesome because of the high sequence diversity and the difficulty to place short contigs into bins, however some virus-specific assembly tools such as IVA [116] for RNA viruses or VICUNA [117] have been developed. The assembly of viral genomes, which are relatively small compared to their cellular counterparts, will be facilitated by third generation sequencing methods that yield longer reads than second generation methods. To correct for the higher error rates, the combination of second and third generation reads in hybrid assembly can be a solution. An additional step in the processing of viral datasets is the removal of host sequences. This can be achieved by mapping the reads against the host genome and removing reads that can be confidently mapped [118]. Proviruses integrated into the host genome still pose a challenge. Approaches addressing this issue include masking them in the host genome or integrating information from other databases, such as HMM-profiles that are either virus- or host-specific, to remove flanking host sequences from proviruses. The latter approach is implemented in CheckV [119] and proved to be more sensitive than other provirus detection approaches used by different tools such as VirSorter [120], PhiSpy [121] and Phigaro [122].

1.6.2 Marker Gene based Analysis

In marker gene based analysis of microbiomes, particular genes from all organisms in a sample are amplified. Suitable marker genes are essential in function, ubiquitous, evolutionary mimicking a molecular clock [123] and only present in a single-copy. Sequences are combined based on their similarity to reference sequences (such as 16S rRNA gene sequence) or to operational taxonomic units (OTUs) [110, 124]. The success of marker gene based analysis depends on the completeness of the reference database, with novel sequences belonging to previously unidentified taxonomic lineages being impossible to be classified [125, 126]. Marker gene based analysis is complicated by marker genes being present in multiple copies and by spacer regions of uneven lengths [127]. Nevertheless many microbial communities have been characterized with this approach. As marker genes are expected to occur only once within a genome, comparing the number of single-copy marker genes found within a draft genome to the number of expected marker genes provides an estimation of completeness, while additional copies of a marker gene can be used as an indicator of contamination [128]. Another advantage of using marker genes is that the resulting read coverages can be used to estimate species abundance without having to normalize by genome size or copy number [115].

Prokaryotic Marker Genes

The 16S rRNA gene has first been used in phylogenetics by Woese and Fox in the 1970s and it has since become one of the most commonly used molecular markers in microbial ecology [32–34]. While some of the marker gene properties are met (essential function, ubiquity), one pitfall of using 16S rRNA is that it is often present in multiple copies and with variable sequences in the same genome or in closely related taxa [38, 129–132]. While possible solutions include the analysis of information on 16S rRNA copy numbers and genome sizes of genome-sequenced bacteria to estimate the relative abundance of individual taxa in a sample [132], other approaches investigate the use of alternative molecular markers [35, 133]. Such novel markers include, but are not limited to, the recombinase A gene family or the RNA polymerase beta subunit (RpoB) gene [134–136]. Compared to rRNA genes, protein-coding genes may have less (or at least different) nucleotide compositional bias than small subunit rRNAs. In 2013 Wu et al produced an extended set of marker genes for bacteria and archaea containing 40 genes [35]. 30 of them are ribosomal protein subunit genes, one is a translation initial factor, one is a translation elongation factor, and three markers are rRNA synthesis related genes. The rest of them are involved in protein metabolism including peptide degradation and exporting, RNA degradation, heme biosynthesis and purine nucleotide synthesis [35]. Pipelines for phylogenomic analysis include AMPHORA [137], the updated version AMPHORA2 [138], which now includes a greatly expanded phylogenetic marker database and can analyze both bacterial and archaeal sequences, as well as TIPP2 [115], a marker gene-based abundance profiling method, which combines phylogenetic placement with statistical techniques to control classification precision and recall. One of the most widely used tools to assess genome completeness and quality of prokaryotic genomes is CheckM. It uses an initial set of universal single-copy marker genes to identify the clade of a genome and subsequently uses clade-specific sets to estimate the quality [128, 139].

Eukaryotic Marker Genes

Some of the universal markers in eukaryotes include rRNAs (small subunit rRNAs) and nuclear protein-coding genes (e.g. EF-1 alpha, alpha-tubulin, beta-tubulin, Actin, RPB1, HSP 70, HSP90, RNA polymerase, myosin), as well as a number of conserved genes encoded in the chloroplast and mitochondrial genomes [140]. These initial marker genes sets were not able to resolve deep relationships within eukaryotes [141]. Ren and colleagues identified a set of 943 low-copy eukaryotic marker genes, that allowed the resolution of more difficult eukaryotic phylogenies [142], however not all of them are universal for all eukaryotes. Taxon-specific marker genes are often used, such as the ribosomal internal transcribed spacer (ITS) region as a marker for Fungi [143, 144]. EukCC, a tool for estimating the quality of eukaryotic genomes based on the automated dynamic selection of single copy marker gene sets, determines the LCA node and estimates the completeness and contamination based on the chosen set [128]. EukCC shows improvements compared to other tools such as CEGMA [145] or BUSCO [146] with regards to automatically choosing a marker gene set.

Viral Marker Genes

The absence of a universally conserved marker gene in viruses poses a major challenge in applying marker gene based analysis to viromes. The fact that viruses are polyphyletic (see Viruses and the Tree of Life) further complicates the definition of universal markers. Even genes that are conserved across various viral groups do not necessarily stem from a common ancestor, but could be the product of gene acquisition events [37]. Koonin introduced the term "Viral hallmark genes" (VHGs) to describe genes shared by many diverse groups of viruses, with only distant homologs in cellular organisms, and with strong indications of monophyly of all viral members of the respective gene families [1]. Since then, marker datasets for some taxonomic groups have been created. Kristensen and colleagues identified a set of bacteriophage markers for different clades based on phage orthologous gene clusters [147]. These signature genes include structural proteins (major capsid protein, tail protein), as well as replication-associated proteins (resolvase, integrase, polymerase, transcriptional regulator). In recent years Koonin and colleagues have built on the idea of VHGs, introducing super-VGHs, that are present in enormously diverse viruses and span two or even three Baltimore Classes [37]. These "super-VHGs" encode the following:

1. RNA-directed RNA polymerases (RdRps) that form an apparently monophyletic group of palm domain-containing polymerases and unite the three BCs (III, IV, and V) of RNA viruses

2. RNA-directed DNA polymerases, or reverse transcriptases (RTs), that unify the two BCs (VI and VII) of reverse-transcribing viruses, along with the related MGEs, and belong to the same branch of the palm domain polymerases as the RdRps

3. Superfamily 3 helicases (S3Hs) that are encoded almost exclusively by MGEs, including (+)RNA (BC IV), most of the ssDNA (BC II), and diverse groups of dsDNA (BC I) viruses

4. Single-jelly-roll capsid proteins (SJR-CPs), which are the most common form of CPs among (+)RNA (BC IV) and ssDNA (BC II) viruses

5. Double-jelly-roll capsid proteins (DJR-CPs), widespread among dsDNA viruses and also found in some ssDNA viruses

6. Rolling-circle replication initiation endonucleases (RCREs) that are encoded by the great majority of ssDNA viruses but are also present in some dsDNA viruses.

Different types of polymerases are found in most viruses, but they are often composed of subunits, making it harder to identify unique genes. Other replication-associated proteins are the rolling-circle replication proteins which are present in viruses with DNA genomes.

Helicases are motor proteins that use energy derived from ATP hydrolysis to separate nucleic acid strands. In dsDNA and dsRNA viruses, the double helix must be separated for copying. In viruses with single-stranded genomes, the duplexes that form after the replication of the genome must be separated. Helicases are also required for transcription of viral mRNAs, translation, disruption of RNA-protein complexes, and packaging of nucleic acids into virions [148, 149].

Single-jelly-roll and double-jelly-roll capsid proteins can be combined as “Major capsid proteins”, that span almost the complete viral world.

1.7 The Virus Orthologous Groups Database (VOGDB)

The VOGDB is the first database of orthologous groups that specifically focuses on viral genomes. While the eggNOG database also includes viral genomes, a major improvement of the VOGDB is that it also considers remote homologs, which are a typical feature of viral proteins due to the fast replication and high mutation rate. The VOGDB is automatically updated when the NCBI Refseq Database [74] is updated, about every two months. Prior to the formation of clusters a quality filter is applied, removing proteins with no annotation and entries containing no sequences. An emphasis of the VOGDB is to provide functional annotation of the proteins in a cluster. Currently the VOGDB (version 208) contains 438,852 proteins from 10,046 genomes that have been placed into 28,386 VOGs (Table 2).

1.7.1 Generation of VOGs

Viral genomes are obtained from the NCBI RefSeq database [74], providing a high-quality dataset with low redundancy (Figure 6). Viral genomes are initially split up into phage and non-phage genomes, due to their different genome architecture and absence of recent common evolutionary origin [25]. Records have to pass a quality filter in order to enter the VOG

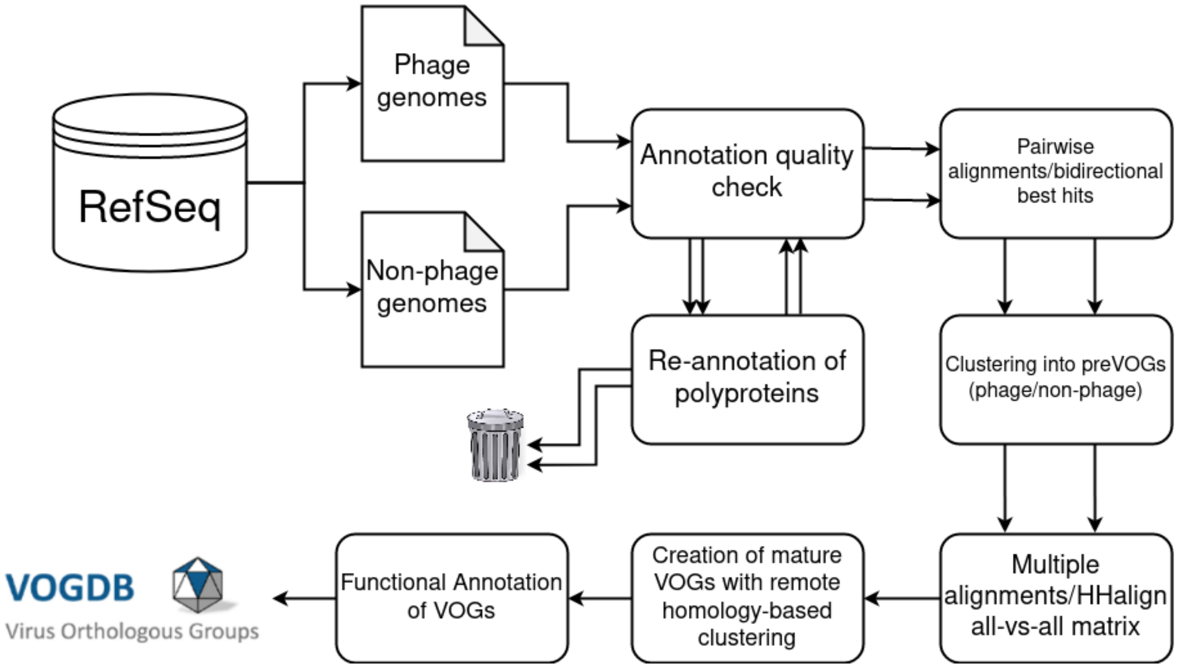

RefSeq
Phage genomes
Non-phage genomes
Annotation quality check
Pairwise alignments/bidirectional best hits
Re-annotation of polyproteins
Clustering into preVOGs (phage/non-phage)
Multiple alignments/HHalign all-vs-all matrix
Creation of mature VOGs with remote homology-based clustering
Functional Annotation of VOGs
VOGDB
Virus Orthologous Groups

Figure 6: Overview of the VOG workflow. Phage and non-phage genomes are separated initially. Poorly annotated records and unannotated polyproteins are removed. Bidirectional best hits (SymBets) are identified and clustered into preVOGs with NCBI COGsoft. HMM profiles are created and used for remote homology-based clustering of preVOGs into mature VOGs. Here phages and non-phages are reunited. Functional annotation and virus specificity (not shown) is determined for each VOG. Parallel arrows indicate where phages and non-phages are treated as two separate groups. Figure adapted from [25].

construction pipeline. Hereby records lacking annotations or sequences are removed, and polyproteins are attempted to be re-annotated. If the annotation of polyproteins is successful, they are re-integrated into the pipeline, otherwise they are discarded. Then the COGsoft algorithm [150] is applied to produce clusters of proteins in the phage and non-phage bins. These clusters are referred to as preVOGs. Up to this point no remote homology tools have been used. In the next step of the pipeline, HHalign from the HHsuite package [94] clusters the preVOGs into larger clusters based on remote homology. Each VOG is then annotated with a consensus function. The next subchapter will explain the current VOG annotation process in more detail.

1.7.2 Current VOG annotation

The current annotation of VOGs is based on simple homology searches of the amino acid sequences with BLAST [90] against the manually curated UniProtKB/Swiss-Prot database [24]. Using a manually curated database as the main source of annotation is implemented in other annotation pipelines as well, such as Prokka, which additionally derives annotations from proteins with high transcript evidence, or performs domain-based annotations based on HMM databases if the annotations based on homologous proteins in databases was not successful [62]. In the VOGDB annotation, the sequence similarity search is performed for every protein in the VOG with BLASTP [90], and the VOG is then annotated with the most frequent match.

To reduce the number of false positives, only results with an e-value <1e-10 are considered, and the minimum query coverage is set to 90% to avoid partial matches. If there are no BLAST hits, the VOG is annotated with the most frequent protein description from RefSeq. Figure 7 provides a schematic overview of the current VOG annotation. As mentioned above, viral genomes evolve rapidly, which is why sequences of viruses sharing a common ancestor are soon not recognized as homologs by a simple homology search using BLAST, which calls for the need to evaluate and adapt the current annotation approach.

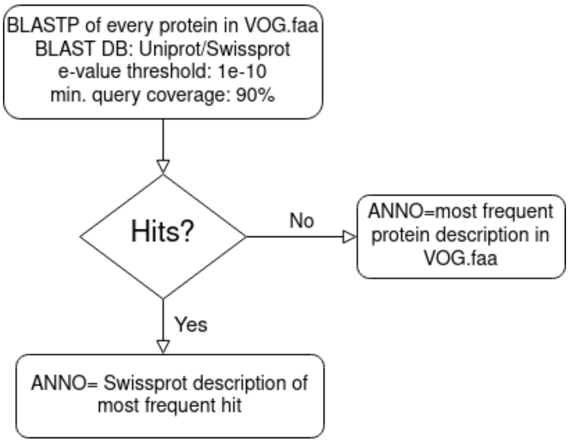

BLASTP of every protein in VOG.faa
BLAST DB: Uniprot/Swissprot
e-value threshold: 1e-10
min. query coverage: 90%
Hits?
No
ANNO=most frequent protein description in VOG.faa
Yes
ANNO= Swissprot description of most frequent hit

Figure 7: Current VOG annotation. A BLASTP search is performed for every protein in the VOG. The VOG is then annotated with the description of the most frequent hit, or, if no hits are available, with the most frequent description from RefSeq.

1.8 The UniProtKB/Swiss-Prot Database

The Universal Protein Resource [151] consortium is an initiative of the Swiss Institute of Bioinformatics (SIB), the European Bioinformatics Institute (EBI) and the Protein Information Resource (PIR) to provide the scientific community with a central resource for protein sequences and functional information. The UniProt consortium maintains the UniProt KnowledgeBase (UniProtKB), updated every 4 weeks, and several supplementary databases including the UniProt Reference Clusters (UniRef) and the UniProt Archive (UniParc).The SP section of the UniProt KnowledgeBase (UniProtKB/Swiss-Prot) was established in 1986 and it contains publicly available protein sequences obtained from a broad spectrum of organisms that are manually annotated by experts [24, 152]. The purpose of the UniProtKB/Swiss-Prot database is to provide a high level of annotation and integration with other databases, while minimizing redundancy [153].

1.8.1 Manual Curation of Swiss-Prot Entries

Manual curation is done by experts in the field, adhering to a standard operating procedure (SOP). The six main steps in the process are: (1) sequence curation, (2) sequence analysis, (3) literature curation, (4) family-based curation, (5) evidence attribution, (6) quality assurance and integration of completed entries [154].

1.8.2 Taxonomic composition of viral UniProtKB/Swiss-Prot Entries

The current release (2021_04) of the UniProtKB/Swiss-Prot contains more than 500.000 manually curated entries, most of which are sequences of bacterial origin (59.28%), followed by sequences from well-studied eukaryotes (34.24%). Viral sequences merely represent 3.01% of the database. The number of archaeal sequences is equally low, also only accounting for 3.48% of the total number of sequences. The most represented species are Human, Mouse and Mouse-ear cress, the best-represented non-eukaryotic species are Escherichia coli (strain K12) (8th most frequent) and Bacillus subtilis (strain 168) (10th most frequent). Acanthamoeba polyphaga mimivirus (APMV) is the only viral species in the top 250 best-represented species (43rd most frequent) [155].

To further investigate the representation of different viral orders and families in the SP database, all reviewed viral records were downloaded and the number of entries for each represented order/family was determined. The viral order with the most entries is Herpesvirales, followed by Articulavirales and Caudovirales. Articulavirales infect invertebrates and vertebrates, while Herpesvirales only infect animals, and Caudovirales primarily infect bacterial cells. Chitovirales, Ortervirales, Imitervirales, Mononegavirales, Asfuvirales, and Reovirales infect only eukaryotes [156]. Thus the SP database contains a bias towards protein entries from viruses infecting eukaryotic cells (Table 3).

Table 3: Distribution of viral orders in the Swiss-Prot database. SP-DB=Swiss-Prot Database.

Viral Order	Number of Species	% of viral SP-DB
Herpesvirales	2273	13.70%
Articulavirales	1588	9.57%
Caudovirales	1585	9.55%
Chitovirales	1376	8.29%
Ortervirales	1122	6.76%
Imitervirales	909	5.48%
Mononegavirales	828	4.99%
Asfuvirales	807	4.86%
Reovirales	806	4.86%
Zurhausenvirales	581	3.50%
Nidovirales	544	3.28%
Pimascovirales	483	2.91%
Rowavirales	418	2.52%
Lefavirales	407	2.45%
Martellivirales	344	2.07%
others	2522	15.20%

1.9 Project Overview

The goal of this project is to analyze the current annotation of VOGs and to investigate methods to aid in the improvement thereof. The functional homogeneity of proteins assigned to a VOG will be examined based on SP-proteins assigned to the VOG. Additionally, the current annotation method using BLASTP to identify homologous proteins will be examined in terms of target sequence coverage and the number of VOG member proteins identifying the protein as a significant hit. The taxonomic composition of the cellular organisms used for VOG annotation is explored and the quality of the functional descriptions is evaluated based on alignments and the presence of SP-proteins in the VOGs.

A new annotation pipeline including remote-homology-based as well as domain-based annotations will be implemented. The information content in the annotations will be increased by adding the most frequent Keywords and GO terms from SP-proteins in the VOG and the method by which the annotation was derived will be stated. Throughout this thesis "current" refers to the VOG annotation approach used to annotate version 208 of the VOGDB, while "new" refers to the newly implemented annotation approach presented in this work.

Another goal is to query the VOGDB to identify marker genes for viruses of the order Caudovirales. With the rapid increase in sequencing data, being able to map novel protein sequences to groups of orthologs can provide insight into the protein's function. Functionally annotated VOGs can be a valuable resource in annotating newly sequenced genomes.

Chapter 2

Methods and Data

2.1 VODGB Data

Flatfiles of the VOGDB v208 were downloaded from vogdb.org. All analyses and experiments throughout this thesis are performed on this version of the VOGDB.

2.2 Taxonomic Composition of the VOGDB

The number of proteins for each species in the file containing all proteins used for VOG construction and in VOG members was determined. Then the ratio was calculated, and lineage information was added. Lineage information was extracted using python and the ETE toolkit [157]. Subsequently, species not represented in the VOGDB were analyzed in more depth, based on lineage and number of proteins. The presence of SP-proteins from the lineages in the VOGDB and in the file containing all proteins from the genomes used for VOG construction was determined as described below, and coverages were calculated. Additionally, species were categorized based on which Baltimore Class they are assigned to.

2.3 Representation of Viral Swiss-Prot Proteins in the VOGDB

2.3.1 Download of Viral Swiss-Prot Proteins

The UniProtKB (version 2021_04) [24] was queried to identify viral manually curated entries, not including polyproteins (search parameters: Taxonomy: Viruses, reviewed:yes, NOT name: polyprotein). A total of 15852 entries were downloaded in fasta format.

2.3.2 Construction of BLAST Databases

A BLAST database was constructed for all proteins used for VOG construction (found in the file “vog.proteins.all.faa”) using the NCBI BLAST+ toolkit [70]. A second BLAST database was created for all VOG member proteins. All VOG.faa files were concatenated, and the VOG number was added to each description line to be able to identify proteins that are placed in several VOGs. Duplicate entries were removed using SeqKit [158] prior to creating the BLAST database.

2.3.3 Identification of Viral Swiss-Prot Proteins in VOGs

For every downloaded viral SP-protein sequence, a BLASTP search was performed against the VOG-member database with the query coverage and the percent identity set to 95%, with an e-value of 1e-10 and a word size of 3. The remaining parameters were the default parameters for BLASTP version 2.11.0+ [70]. For proteins with no hits in the VOG-member database, an additional BLASTP search was performed with the same parameters against the BLAST database containing all proteins used for VOG construction. Each SP-protein is placed in one of three categories, depending on it being assigned to a VOG, unassigned, or excluded from VOG construction.

2.3.4 Analysis of Functional Homogeneity of Swiss-Prot Proteins assigned to the same VOG

For VOGs containing more than one SP-protein, the names of these proteins were manually checked for similarity. The results are classified as “agreeing”, “not agreeing” or “undetermined” and summarized for VOG00001-VOG00150. As proteins with descriptions such as “Gene (. . .) protein” or “uncharacterized protein” could not be evaluated in that way, their SP entries were analyzed to retrieve domain and family information and to check if this information matches that of other SP VOG-members. VOG06194, VOG06147, VOG00011 and VOG00153 were evaluated in this way. These VOGs were chosen from the “Top 10 most wanted VOGs”. MSAs were created with Clustal Omega (version 1.2.4) [159] with the default parameters.

2.4 Analysis of the current VOG Annotation

2.4.1 Overview of the current Annotations

Determination of uninformative protein descriptions

SP-protein descriptions containing the patterns "uncharacterized (...) protein", "gene (...) protein", "Protein (...)", "gene product", "putative protein (...)", were considered to be uninformative. The same approach was taken for Refseq annotations by searching for the descriptions containing "orf", "gp" and "hypothetical protein" in the VOG annotations file.

Determination of the Number of VOGs and Proteins belonging to each Category

A bash script was written to extract the number of SP based annotations containing one of the above patterns. The amount of informative SP annotations was calculated as the difference between the number of annotations containing "sp|" and the number of annotations containing both patterns. The approach was repeated to determine the number of RefSeq annotations with and without the patterns determined to indicate uninformative RefSeq descriptions. The number of proteins in the different categories was calculated as the sum of the proteins assigned to the VOGs belonging to each of the four annotation groups.

2.4.2 Analysis of Annotations based on Homologies to Proteins in the Swiss-Prot Database

For every VOG annotated based on homology to a protein in the SP database, the lineage information for that protein was retrieved from the SP database (release 2021_04) [24]. Subsequently the numbers of VOGs annotated with Proteins of viral, bacterial, eukaryotic and archaeal origin was determined. For the different origins of annotation, a more in-depth analysis was performed by comparing function descriptions with the descriptions of viral SP-proteins in the respective VOG and by manually inspecting SP entries if any doubt remained.

Analysis of identified BLAST Hits

The parameters for the BLASTP search are kept the same as in the current VOG annotation approach, but the output format was modified slightly to include the alignment coordinates on the target sequence. The SP protein used in the current annotation is identified, and the target sequence coverage is calculated by dividing the target sequence length by the

length of the alignment onto the target sequence. For VOGs with more than one protein identifying the same SP-protein as a significant hit, the average target sequence coverage is calculated. Average target sequence coverages of the proteins used for annotation are analyzed for (i) VOGs annotated based on SP-proteins that are assigned to the VOG, (ii) VOGs containing SP-proteins annotated with a protein not in the VOG, (iii) VOGs not containing any SP-proteins that are annotated with a SP-protein, as well as VOGs annotated with proteins from (iv) Bacteria, (v) Archaea, and (vi) Eukaryotes.
Visualization: For each of the six categories, VOGs were divided into five bins based on the average target sequence coverage of BLAST hits to the protein used for annotation. Proteins used for annotation that had low target sequence coverages in the BLAST alignment were inspected on the UniProt website and then viewed in InterPro to analyze domain architecture. Coordinates form the BLAST hits were retrieved from previously run BLAST searches, and the presence of functional domains in that region was determined.

The number of VOG proteins per target is counted and VOGs are binned based on the percentage of VOG proteins identifying the SP-protein as a target. Relevant information was extracted from the BLAST output table with bash scripts.

2.5 Test Data Sets based on viral Proteins described in Literature

While the SP database can be used to check the quality of VOGs based on the presence of manually curated proteins present in the VOG, it does not take into consideration that the annotations themselves depend on homologies to the SP database. A set of literature-reviewed proteins is created to provide a SP independent quality check of VOG annotations. NCBI Pubmed [160] was queried for recent publications on functional descriptions of viral genes. In “The Revisited Genome of Bacillus subtilis Bacteriophage SPP1” Godinho et al [161] describe an updated annotation of the SPP1 genome. The authors revisited the sequence and organization of the bacteriophage SPP1 genome and provide an updated annotation based on available experimental data and bioinformatics. A revised version of the Genbank entry is in the process of submission, but not yet publicly available. The 38 functionally described SPP1 genes were looked up in the existing GenBank entry for SPP1 (AC: X97918.2), and their associated UniProt Accessions (UniProtKB, release 2021_04) [24] and RefSeq Accsessions (release 209) [64] were identified. Based on the RefSeq Accession, the VOG(s) (VOGDB version 208) the protein is assigned to was determined. Subsequently the VOG annotation was compared to the annotation of the protein to check if there is reasonable agreement between the functional descriptions. SP entries of the proteins used for VOG annotation, that did not show agreement with the SPP1 protein, were analyzed to determine if the disagreement stems from the annotation being too specific (over-annotation), too vague (under-annotation), completely contradicting, or a partial agreement. As some of the proteins from the SPP1 genome are already deposited in the SP database, a second data set was constructed by querying UniProt [24] with the following parameters: database:(type:refseq) taxonomy:"Viruses [10239]" length:[80 TO 2000] NOT name:polyprotein existence:"Evidence at protein level [1]" AND reviewed:no. Ten proteins with functional descriptions from literature, but not yet in the manually curated SP database, were added to a panel, and as with the SPP1 proteins the function of the VOGs they are assigned to were compared to the functional description of the protein.

2.6 Program Structure of the new Annotation Pipeline

The new annotation pipeline consists of several scripts. The initial script ("A0allvoganno.sh" in the diagram) is a preparatory script, which builds a Database from the SP database text file, from which information will be extracted during the annotation process. Building the Database allows for fast information retrieval in later steps. Other preparatory steps are the extraction of all viral SP-protein sequences to a separate file and building a BLAST database from it. Then the annotation script (A1voganno.py) is called. Here a connection to the database built in the previous step is established, and all VOG members are scanned for the presence of domains from different families. All VOG files are iterated over and annotation is done for each of them. There are four levels of annotation, the first one being the annotation with SP-proteins present in the VOG. If an annotation approach is not successful, the next approach is called. For VOGs without SP-proteins, homology-based annotations (A2bhmmsearch.py) are attempted next, before annotating with domains found by InterProScan (A2cinterproscan.sh) or RefSeq descriptions (A2drefseqanno.sh). Annotations are returned to the subordinate script A1voganno.py and written to a file. Upon finishing the annotation process, the database connection is closed. The individual annotation approaches are described in more detail in the following section.

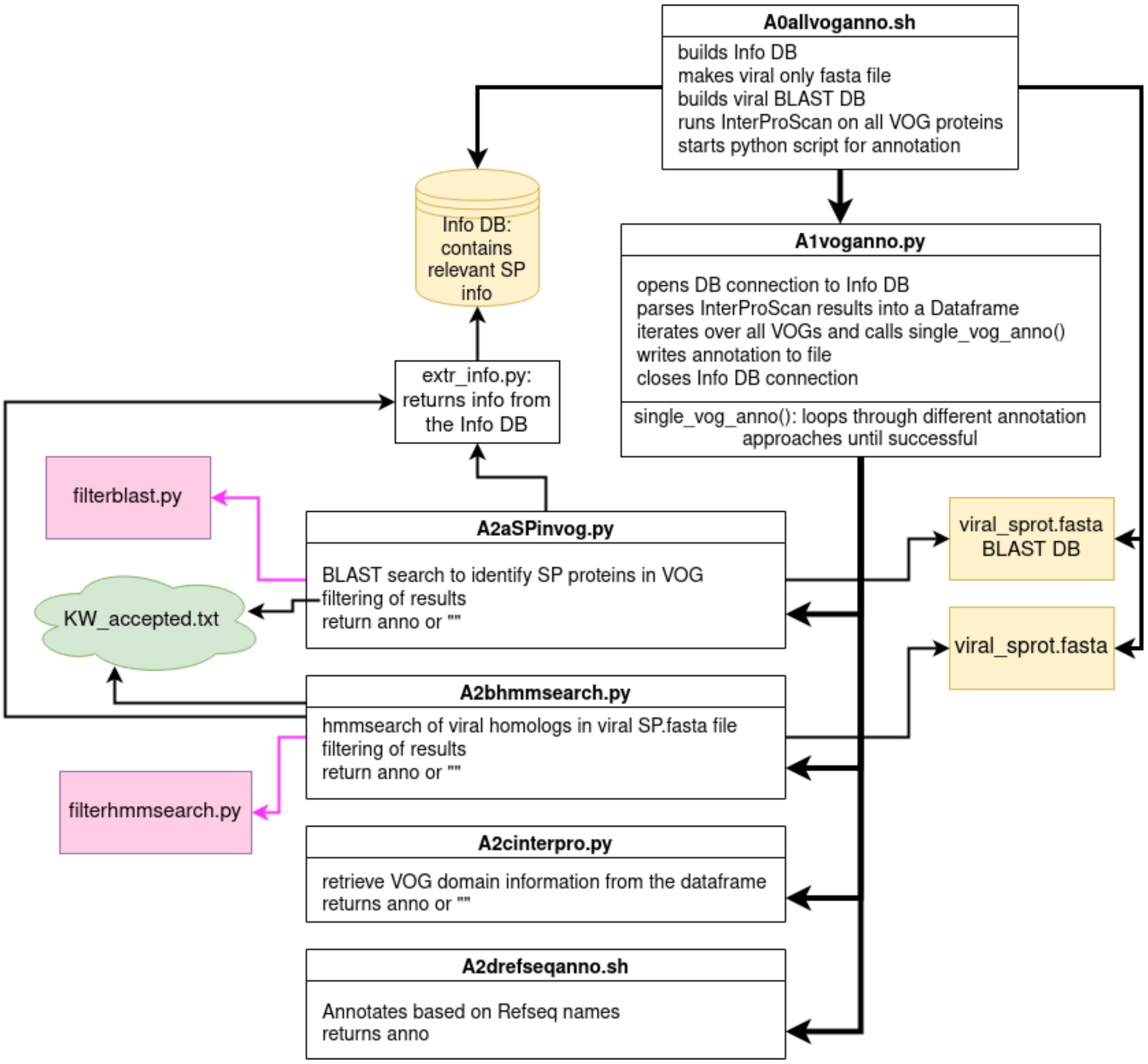

A0allvoganno.sh
builds Info DB
makes viral only fasta file
builds viral BLAST DB
runs InterProScan on all VOG proteins
starts python script for annotation
Info DB: contains relevant SP info
A1voganno.py
opens DB connection to Info DB
parses InterProScan results into a Dataframe
iterates over all VOGs and calls single_vog_anno()
writes annotation to file
closes Info DB connection
single_vog_anno(): loops through different annotation approaches until successful
extr_info.py: returns info from the Info DB
filterblast.py
A2aSPinvog.py
BLAST search to identify SP proteins in VOG
filtering of results
return anno or ""
viral_sprot.fasta BLAST DB
KW_accepted.txt
viral_sprot.fasta
A2bhmmsearch.py
hmmsearch of viral homologs in viral SP.fasta file
filtering of results
return anno or ""
filterhmmsearch.py
A2cinterpro.py
retrieve VOG domain information from the dataframe
returns anno or ""
A2drefseqanno.sh
Annotates based on Refseq names
returns anno

Figure 8: Program Structure of the new VOG Annotation Pipeline. The initial script (A0) builds a database from which Swiss-Prot information can later be extracted efficiently. A file containing only viral SP proteins is created, and from it a BLAST database is built. The actual annotation process is started in a python script, where the connection to the just created Database is established, and subsequently VOGs are annotated one at a time. First SP proteins in VOGs are identified via BLAST (script A2a), and annotation is performed based on those proteins. If no annotation can be obtained, the next script (A2b) is called, attempting to annotate the VOG based on viral homologs in the SP database. Both annotations rely on filtering of the search outputs, and access the DB to retrieve Keyword and GO-info which are included in the annotation. If no homologs are identified, InterProScan identifies domains present in VOG proteins, and if no domains are found, annotation is performed based on the RefSeq names of the proteins. SP=Swiss-Prot

2.6.1 Annotations based on Swiss-Prot proteins present in the VOG

SP-proteins in VOGs are identified with a BLASTP sequence similarity search [70], with an e-value of 1e-10, word size of 4, percent identity of 95%, query coverage of 95%, and sequence coverage of 95% to avoid alignments to polyproteins. Uninformative protein names such as “Gene (...)”, “ORF”, or “uncharacterized protein” are removed from the hit list, and the hits are sorted by most frequent names and protein evidence. Sorting by protein evidence ensures that large numbers of homology-inferred proteins are not preferred over single high quality proteins in the annotation process. The VOG is then annotated with the SP-protein(s) with the most frequent description and highest level protein evidence. The accession numbers of all SP-proteins assigned to the VOG are provided along with the most frequent GO terms and keywords associated with all SP-proteins assigned to the VOG.

2.6.2 Homology-based Annotations

In order to find remote homologs in the SP DB, hmmsearch from the HMMER package (v.3.3) [72] is applied. The HMM profiles from the VOGDB flat files in the “hmm/” directory are used as queries. Only hits with evalues < 1e-04 and a mean posterior probability of aligned residues in the MEA alignment > 0.7 are kept to ensure that hits are of high alignment quality and having a low probability of being false positives. Additionally a query and target coverage of 80% is applied to avoid hits to just single domains. Subsequently proteins with uninformative names are filtered out. The hits are then grouped based on identical descriptions and level of protein evidence. Again proteins with uninformative names are removed as above. The most frequent description with the highest protein evidence is chosen as the annotations, and a SP accession number is included in the annotation as a reference.

2.6.3 Domain-based Annotations: InterProScan

Analysis is performed with models from the Pfam [75], PrositeProfiles [102] and SUPERFAMILY [106] databases, as these are known to contain viral genomes and to be updated regularly. They are also well integrated into InterPro, with integrated signatures of 95.3%, 92.6% and 80.1% for Pfam, PrositeProfiles and SUPERFAMILY, respectively. The number of proteins containing each domain is reported in the output, and InterPro domains as well as GO information is added for each domain by switching on the “goterms” flag. The remaining parameters are the default parameters for interproscan-5.53-87.0 [69]. Prior to running

InterProScan stop codons denoted as "*" in the sequence files are replaced by "X", as InterProScan does not accept stop codons. For each domain the number of VOG proteins containing it is specified. No e-value cut-offs are applied for the domains reported by InterProScan, as e-values are specific to the domain databases, and therefore not comparable. InterProScan assumes all reported hits to be true hits.

2.6.4 Annotations based on RefSeq Descriptions

First, duplicates are removed from the faa files, and then names are sorted by number, excluding uninformative names such as "hypothetical protein", "orf" and "gp". VOGs are then annotated with the most frequent description and the number of proteins sharing that name is specified.

2.7 Identification of Marker Gene VOGs

2.7.1 Major Capsid Protein

First all VOGs with annotations containing "major capsid" were extracted from the new annotations file, then low scoring VOGs (RefSeq-based annotations with just one protein with that description, or just one protein with that domain) were removed. Upon inspection of species represented in more than 1 VOG, additional VOGs were removed, if they always coincided with another VOG (e.g. VOG00427 was removed, because species in this VOG are also represented in VOG00711). VOG26441 was also removed, since it codes for a capsid protein, but VOG26440 codes for the other unit of the capsid, and they always occurred together. Then VOGs with descriptions, keywords or GO terms containing "helical capsid protein", and "g8p" (Inoviridae) capsid proteins, as well as for "C:T=", which describes the shape of the capsid, were selected, as these MCPs are specific to some viruses.

2.7.2 Terminase, large subunit

All VOGs with annotations containing both the words "terminase" and "large" were extracted from the VOG annotations file created with the new annotation pipeline described above. The same approach for selecting terminase, large subunit VOGs was applied using the current annotation file.

Chapter 3

Results and Discussion

3.1 Taxonomic Composition of the VOGDB

Out of 10046 species used for VOG construction, only 4776 are represented in VOGs. The majority of viral species represented in the VOGDB has dsDNA genomes (94.07%), followed by 3.37% of ssDNA viruses, while RNA viruses are not well-covered, comprising just 0.27%, 0.54% and 1.47% for dsRNA, negative-strand RNA and positive-strand RNA viruses, respectively. Only 7 viruses in the VOGDB are reversely transcribing. Two species belong to each alphasatellites, betasatellites and unclassified archaeal viruses (Table 4).

Table 4: Species Representation in the VOGDB (v208) by Baltimore Class

Baltimore Class	Number of Species in the VOGDB	% of all Species in the VOGDB
dsDNA	4493	94.07%
ssDNA	161	3.37%
dsRNA	13	0.27%
ss(-)RNA	26	0.54%
ss(+)RNA	70	1.47%
rev.transcr.	7	0.15%
alphasatellites	2	0.04%
betasatellites	2	0.04%
unclassified archaeal v.	2	0.04%

Caudovirales are the most abundant order in the VOGDB, accounting for 80.17% of all proteins used for VOG construction that were obtained from viral RefSeq genomes and passed the quality filter (Table 5). The next frequent orders are Imitervirales (5.59%), Lefavirales (2.33%), Algavirales (2.15%) and Chitovirales (1.83%).

The lineage distribution at the order level changes after VOG construction, as 97.28% of all proteins assigned to VOGs are from Caudovirales, followed by Chitovirales (0.51%), Imitervirales (0.39%) and Herpesvirales (0.37%) (Table 5). No lineage information could

Table 5: Taxonomic composition of the VOGDB. Viral orders with the number of proteins used for VOG construction (allprot) and the number of proteins per order assigned to VOGs. Columns 3 and 5 show the fraction of total proteins and VOG member proteins belonging to that order. Numbers and coverages of manually curated Swiss-Prot (SP) proteins are shown, with Caudovirales being the dominant order with all SP proteins belonging to that order having been assigned to VOGs.

Order	# prot in all-prot	% of all proteins	# prot in VOG-DB	% of VOG proteins	coverage of prot in the VOG-DB	# SP in all-prot	# SP in VOG	coverage of SP-proteins in the VOGDB
Caudovirales	461532	80.17 %	426898	97.28 %	92.50 %	1439	1439	100.00 %
Imitervirales	32171	5.59 %	1717	0.39 %	5.34 %	909	502	55.23 %
Lefavirales	13439	2.33 %	875	0.20 %	6.51 %	0	0	–
Algavirales	12397	2.15 %	53	0.01 %	0.43 %	6	6	100.00 %
Chitovirales	10520	1.83 %	2257	0.51 %	21.45 %	1264	1126	89.08 %
Herpesvirales	10302	1.79 %	1627	0.37 %	15.79 %	2025	910	44.94 %
Pimascovirales	10243	1.78 %	169	0.04 %	1.65 %	452	63	13.94 %
Ligamenvirales	1417	0.25 %	1223	0.28 %	86.31 %	208	145	69.71 %
other	23706	4.12 %	4033	0.92 %	17.01 %			
total:	575727		438852		76.23 %			

be retrieved for 15 species, accounting for 819 proteins in the file containing all proteins, and 767 proteins in the file containing VOG member proteins. Manual inspection revealed that most of these species belong to the order of Caudovirales. 33 additional species are unclassified on the taxonomic order level, accounting for a total of 836 proteins.

Analysis of the unrepresented species showed that 2076 of the them come from an unclassified lineage, and the most dominant orders in unrepresented species are Geplafuvirales (645 species), Mononegavirales (331 species) and Cirlivirales (220 species). These orders are poorly covered in the VOGDB, with average coverages per order of 4%, 5% and 1% for Geplafuvirales, Mononegavirales and Cirlivirales, respectively. The average number of proteins in the uncovered species is only 22.9, with 4386 species having less than 10 proteins. Contrarily, the average protein number in VOG-covered species is 100.2. Thus the majority of species not covered in the VOGDB are from badly covered lineages and contain small numbers of proteins.

While the taxonomy of VOGs is skewed towards Caudovirales due to the wide abundance of genomes from Caudovirales in the RefSeq database, the SP database, which is the basis for VOG annotation does not show such a vast difference between the numbers of proteins from different orders (Taxonomic composition of viral UniProtKB/Swiss-Prot Entries).

Low SP-protein coverages were observed for Herpes-, Imiter- and Pimascovirales, all of which have low protein coverages in the VOGDB (Table 5). Even though the order Chitovirales is poorly covered in the VOGDB, 89.09% of SP-proteins belonging to that order are covered. Therefore the coverage of SP-proteins in VOGs does not necessarily depend on the overall coverage of proteins from that lineage. It is, however, an indicator that SP-proteins from some lineages are not placed in VOGs due to the general poor coverage for that lineage. Nevertheless, these manually annotated proteins can provide valuable information for the functional annotation of VOGs based on homologies.

3.2 Representation of Viral Swiss-Prot Proteins in the VOGDB

3.2.1 Number of Swiss-Prot proteins in VOGs

5057 of the obtained viral SP-proteins are assigned to VOGs, 6481 are used for VOG construction, but remain unassigned to VOGs, and 4314 were found in neither the database created from VOG members nor the database created of all proteins used for VOG generation. The latter proteins are thus not used in the VOGDB construction (Table 6). Reasons for this could be that they belong to blacklisted genomes that are excluded due to the applied quality filters or due to a lacking cross-reference to RefSeq [74].

Table 6: Number of viral Swiss-Prot proteins assigned to VOGs, unassigned, or not included in VOG construction.

Assigned to VOGs	5057
Not assigned to VOGs	6481
Excluded from VOG-generation	4314

3.2.2 Distribution of Swiss-Prot proteins within VOGs

A total of 1841 VOGs contain viral SP-proteins, however, most of these VOGs only contain a single SP-protein. The distribution of SP-proteins within VOGs in shown in Table 7 and Figure 9.

Table 7: Distribution of viral Swiss-Prot proteins in VOGs. SP=Swiss-Prot.

# SP in VOG	# VOGs
1	1092
2	263
3	113
4	86
5	70
6-10	143
11-20	54
21-30	12
31-40	7
> 40	1
total:	1841

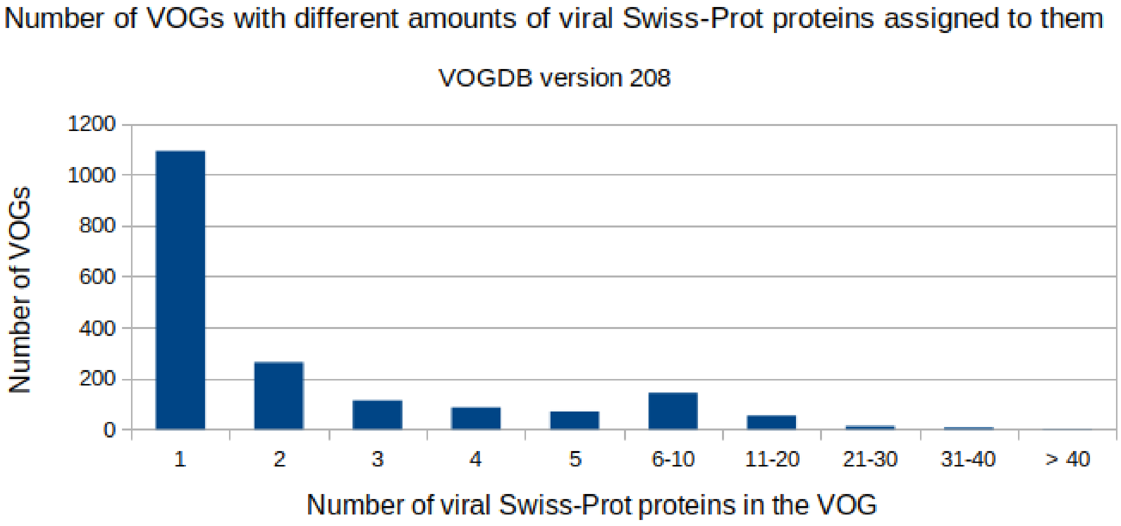


Figure 9: Viral Swiss-Prot Protein Distribution within VOGs. In the majority of VOGs containing Swiss-Prot proteins, only one Swiss-Prot protein is present.

3.2.3 Functional Description of VOGs based on Swiss-Prot Proteins assigned to them

675 out of 1092 VOGs containing just one SP-protein have uninformative functional descriptions (e.g. “uncharacterized”, “gene product”) based on these proteins. The SP-proteins in the remaining 417 VOGs have names that provide more information about the function (Figure 10).

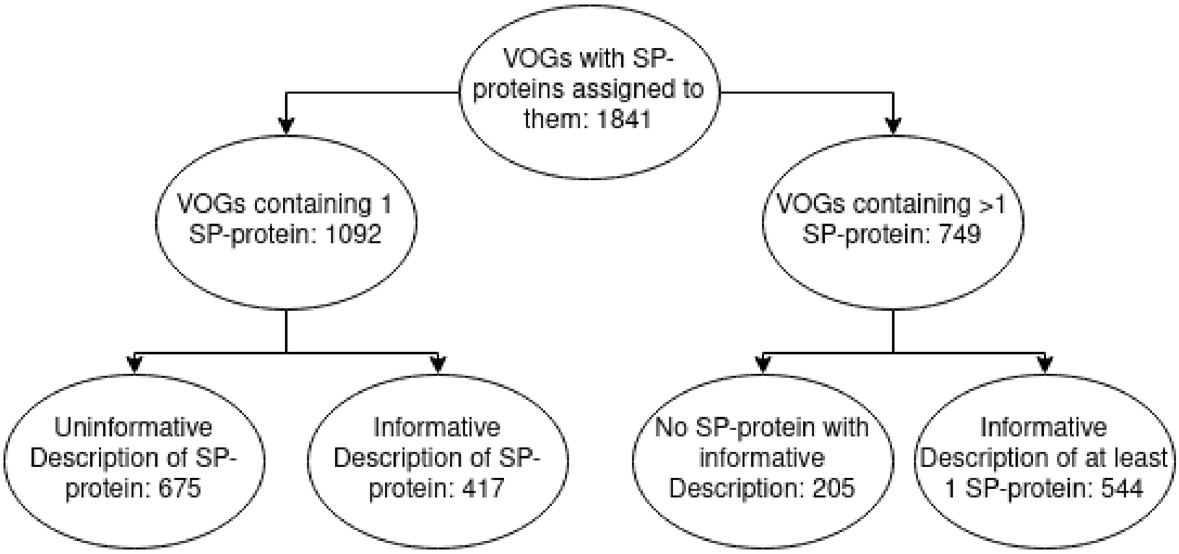

VOGs with SP-proteins assigned to them: 1841
VOGs containing 1 SP-protein: 1092
VOGs containing >1 SP-protein: 749
Uninformative Description of SP-protein: 675
Informative Description of SP-protein: 417
No SP-protein with informative Description: 205
Informative Description of at least 1 SP-protein: 544

Figure 10: Functional Descriptions of VOGs with viral Swiss-Prot proteins assigned to them. Classification of VOGs containing viral Swiss-Prot proteins into VOGs containing 1 or more than 1 Swiss-Prot proteins, and further classification based on the information content of the descriptions. SP=Swiss-Prot

3.2.4 Analysis of Functional Homogeneity of VOGs containing more than one Swiss-Prot Protein

Table 8: VOGs00001-00150 that contain more than 1 Swiss-Prot protein. "Undefined" refers to Swiss-Prot descriptions that neither agreed nor obviously disagreed.

Agreement	No Agreement	Undetermined
31	1	6

For most of the examined VOGs containing several SP-proteins the functional descriptions were agreeing between them. Out of the first 150 VOGs, which were observed in greater detail, 31 VOGs containing more than one SP-protein had matching functional descriptions for the SP-members. For 6 VOGs it was unclear if the functional SP descriptions were in agreement due to protein names such as "Tegument Protein" (VOG00036) that provide little information (Table 8). A special case occurred in VOG00001. This large VOG contains a total of 1097 proteins, some of which have manually curated annotations. Some proteins in the VOG are annotated as "repressors" or "negative regulators" (e.g. SP-accession: Q05286: Repressor-like immunity protein (Gp71), SP-accession: P06020: "Negative regulator of transcription (Ner, Gene product 2, gp2)", SP-accession: P06903: "Negative regulator of transcription (Ner)"), while others are "Transcriptional activators" (SP-accession: P03042: "Transcriptional activator II", SP-accession: P03041: "Transcriptional activator protein C1").

Transcriptional regulators are known to diverge to have different functions, and although placed in the same VOG, there is no single consensus function. Manually VOG00001 would be annotated as "Transcriptional Regulator", which includes activators as well as repressors.

Functional Homogeneity of Swiss-Prot Proteins assigned to 4 selected VOGs

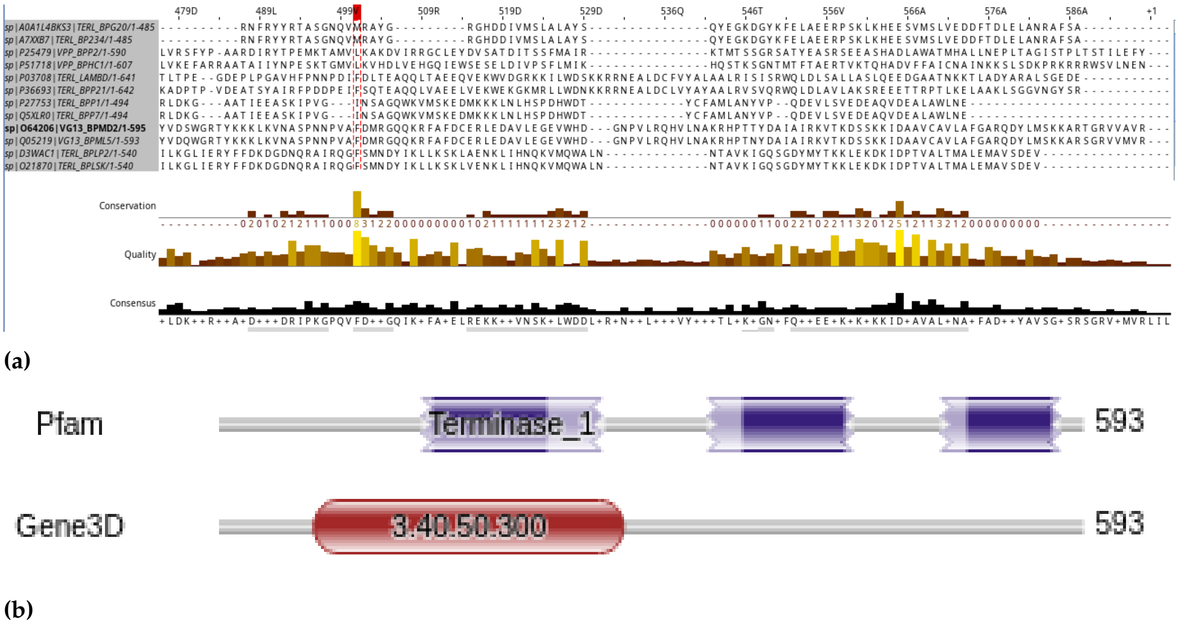
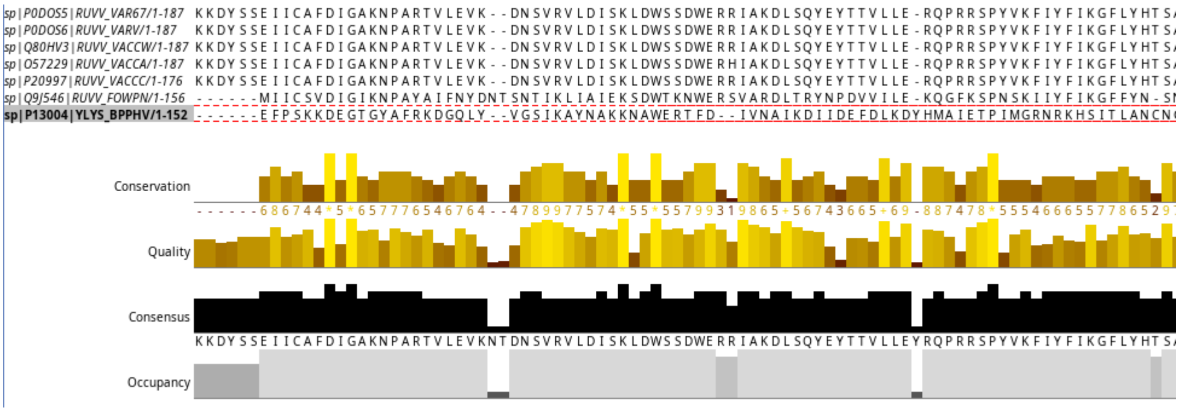


Figure 11: MSA and Domain Inspection of VOG06194. a) Part of the MSA of all SP-proteins in VOG06194. The protein in bold letters is the "Gene 13 protein" used in the current annotation, and it is set as a reference. The red column indicates the coordinates of the reference sequence where the start of the "Terminase" Pfam domain is located. **b)** Identified Pfam and Gene3D domains in the "Gene 13 protein".

VOG06194: VOG06194 is currently annotated as "Gene 13 protein" (SP-accession: Q05219). In the SP entry the "Gene 13 protein" is indicated to belong to the phage terminase family. All other SP-proteins assigned to VOG06194 are annotated as "(probable) Terminase, large subunit". Hmmscan [72] identified the "Terminase, large subunit" Pfam domain (PF03354) with an e-value of 1e-09 at the coordinates 498-577 in the Q05219 protein. This region is conserved in the MSA of all SP-proteins present in the VOG (Figure 11a), with a highly conserved Proline residue at position 501 in the MSA highlighted as the red column. Another hit was found in the Gene3D database (e-value: 1e-08), this domain belonging to the "P-loop containing nucleotide triphosphate hydrolases" superfamily, and the "Terminase DNA packaging enzyme large subunit" Functional Family (Figure 11b).

VOG06147: VOG06147 is currently annotated as "Gene 4 protein" (SP-accession: O64200). This protein contains a Pfam domain for "endonuclease". VOG06147 has 37 SP proteins assigned to it. The MSA does not show high conservation and is not informative in terms of identifying similarities. 19 of the assigned SP proteins are endonucleases, the remaining proteins are "uncharacterized", "Protein (...)", or "Gene (...) Protein". Manual observation showed that these proteins contain Pfam domains associated with endonuclease activities (PF01844, PF02945, PF00149, PF07460).

VOG00011: This VOG is currently annotated as "uncharacterized protein near lysin gene (Fragment)" (SP-accession: P13004). VOG00011 contains 7 SP-proteins, all but one are annotated as Resolvase proteins. The "uncharacterized protein near lysin gene (Fragment)" comes from a Caudovirales species, while the other SP proteins in the VOG stem from Varidnaviria. As the VOGDB mainly contains Caudovirales proteins, it is not surprising that the protein stemming from a Caudovirales genome is most often identified as a significant BLAST hit for VOG members. While the alignment shows high conservation, this is mostly due to the Resolvase proteins, the "Uncharacterized protein P13004" indicated in bold letters does not align well to the other proteins (Figure 12). Pairwise alignments between the P13004 protein and the other SP proteins in the VOG show short alignment lengths with low identities. P13004 contains a "Ribonuclease H-like" superfamily domain (IPR012337), which is described to be also present in resolvases, providing a link between the functional descriptions of the SP-proteins assigned to VOG00011.



```
sp|P0DOS5|RUVV_VAR67/1-187   KKDYSSEIICAFDIGAKNPARTVLEVK--DNSVRVLDISKLDWSSDWERRIAKDLSQYEYTTVLLE-RQPRRSPYVKFIYFIKGFLYHTS
sp|P0DOS6|RUVV_VARV/1-187    KKDYSSEIICAFDIGAKNPARTVLEVK--DNSVRVLDISKLDWSSDWERRIAKDLSQYEYTTVLLE-RQPRRSPYVKFIYFIKGFLYHTS
sp|Q80HV3|RUVV_VACCW/1-187   KKDYSSEIICAFDIGAKNPARTVLEVK--DNSVRVLDISKLDWSSDWERRIAKDLSQYEYTTVLLE-RQPRRSPYVKFIYFIKGFLYHTS
sp|O57229|RUVV_VACCA/1-187   KKDYSSEIICAFDIGAKNPARTVLEVK--DNSVRVLDISKLDWSSDWERHIAKDLSQYEYTTVLLE-RQPRRSPYVKFIYFIKGFLYHTS
sp|P20997|RUVV_VACCC/1-176   KKDYSSEIICAFDIGAKNPARTVLEVK--DNSVRVLDISKLDWSSDWERRIAKDLSQYEYTTVLLE-RQPRRSPYVKFIYFIKGFLYHTS
sp|Q9J546|RUVV_FOWPN/1-156   ------MIICSVDIGIKNPAYAIFNYDNTSNTIKLIAIEKSDWTKNWERSVARDLTRYNPDVVILE-KQGFKSPNSKIIYFIKGFFYN-S
sp|P13004|YLYS_BPPHV/1-152   ------EFPSKKDEGTGYAFRKDGQLY--VGSIKAYNAKKNAWERTFD--IVNAIKDIIDEFDLKDYHMAIETPIMGRNRKHSITLANCN

Conservation
                             ------686744*5*657776546764--4789977574*55*557993198 65+56743665+69-887478*5554666557786529
Quality

Consensus
                             KKDYSSEIICAFDIGAKNPARTVLEVKNTDNSVRVLDISKLDWSSDWERRIAKDLSQYEYTTVLLEYRQPRRSPYVKFIYFIKGFLYHTS
Occupancy
```

Figure 12: MSA of Swiss-Prot proteins assigned to VOG00011. Part of the alignment of VOG00011 Swiss-Prot proteins. The protein currently used to annotate the VOG (SP-accession: P13004) is shown in bold letters and it is set as the reference sequence.

VOG00153: VOG00153 contains 28 SP proteins, 22 of which are associated with transcription, while 5 are helicases and the protein annotated as "Putative protein p41" contains Helicase (IPR014001) and transcription regulation (IPR000330) InterPro domains.

Additional observation of Polyproteins

The presence of polyproteins was determined separately, with only 15 polyproteins being placed in VOGs. This is little surprising as polyproteins are filtered prior to VOG construction if they do not pass a quality filter. In the observations done in this chapter polyproteins are not included.

According to the above analysis, SP-proteins that lack detailed functional descriptions frequently contain domains or other fields in the SP entry matching the functions of other SP proteins assigned to the VOG, and in the annotation process these other, more informative SP description could be favored in describing VOG function. In total 880 out of 1841 (47.8%) VOGs containing SP-proteins do not contain proteins with descriptions providing information about the function. Therefore, VOG annotation based on SP-proteins assigned to VOGs is only applicable to the remaining 961 VOGs. These annotations, however, would be of high quality as these proteins are assigned to the VOGs directly. For the VOGs not annotated with SP-proteins assigned to them, the SP-proteins that have not been assigned to VOGs could be a valuable source of function prediction by determining homologies between VOG member proteins and the manually curated proteins not present in the VOGDB.

3.3 Overview of the current VOG Annotation

Out of 28386 VOGs (VOGDB v208), only 2435 VOGs are annotated based on SP-proteins, 1102 (3.88%) of which have uninformative descriptions, while 1333 (4.70%) have more informative functional descriptions. 25951 VOGs have annotations based on protein descriptions from the NCBI RefSeq database [74]. 24269 (85.5%) of those VOGs are annotated with descriptions lacking information about the proteins function, while 1682 (5.93%) have descriptions providing more information (Figure 13, Table 9). A closer observation of the VOGs revealed that the 1333 VOGs with informative SP annotations contain a total of 125158 proteins – on average these VOGs contain 93.9 proteins. 28.52% of all VOG-proteins are placed in these VOGs. While more than 80% of VOGs are annotated with uninformative RefSeq descriptions, only 52.61% of VOG-proteins are contained in them. The VOGs annotated with uninformative descriptions from SP and with informative RefSeq names contain 14.31% and 4.55% of all VOG-proteins, respectively. The VOGs annotated with “REFSEQ hypothetical protein” only contain an average of 9.5 proteins, thus the average VOG size of SP-annotated VOGs is about 10fold higher.

Table 9: Overview of VOGs annotated with different approaches, showing numbers of VOGs and proteins assigned to them, as well as percentages of total VOGs and proteins in the VOGDB annotated with the specific approach.

Annotation	# VOGs	# proteins	% VOGs	% proteins
SP Annotation informative	1333	125158	4.70%	28.52%
SP uninformative	1102	62816	3.88%	14.31%
RefSeq Annotation informative	1682	19977	5.93%	4.55%
RefSeq uninformative	24269	230901	85.5%	52.61%

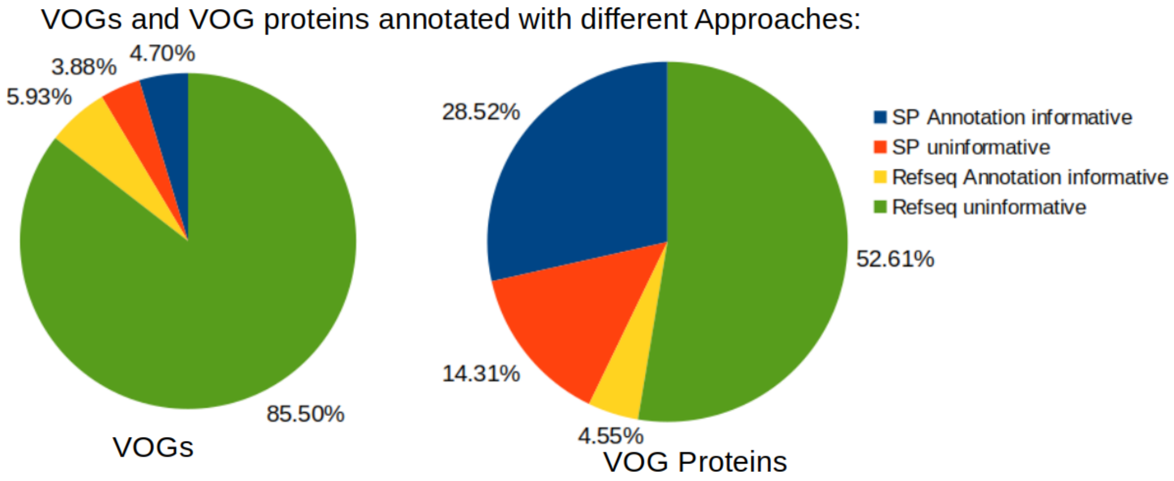


Figure 13: VOGs classified by Annotation Approach. 85.5% of all VOGs are annotated as "hypothetical protein", "orf" or "gp" based on protein descriptions from RefSeq containing 52.61% of all VOG proteins (green). 5.93% of VOGs are annotated based on other RefSeq descriptions - 4.55% of proteins are assigned to them (yellow) . Only a fraction of VOGs (4.70%) is annotated based on proteins from the UniProtKB/Swiss-Prot database that do not contain the patterns specified as uninformative in the description, totaling for 28.52% of VOG proteins (blue). 3.88% of VOGs have SP descriptions that do not provide information about the function and include 14.31% of VOG proteins. SP=Swiss-Prot.

3.3.1 Top 10 Most wanted Annotations

Table 10 shows the ten VOGs with the largest number of species that lack a functional annotation. All ten VOGs are annotated based on homologies to proteins in the SP database, but the descriptions do not shed light on the proteins' function. The overview of the current annotation shows that many annotations in the VOGDB do not provide information about the VOG function (Figure 13, Table 9). While proteins in the SP database are manually curated, their descriptions do not always represent the full information content present about that protein in the SP entry. Other fields of the entry such as Keywords or text descriptions frequently contain functional information.

Table 10: 10 VOGs with the largest number of species lacking an informative functional description in the VOGDB. # Prot=number of VOG proteins. # Spec=number of VOG species. Fct-cat=functional category.

VOG	# Prot	# Spec	Fct-cat	Consensus Functional Description
VOG06194	3200	3120	Xu	sp|Q05219|VG13_BPML5 Gene 13 protein
VOG06147	7764	3092	Xu	sp|O64200|VG04_BPMD2 Gene 4 protein
VOG02932	953	952	Xu	sp|Q05240|VG31_BPML5 Gene 31 protein
VOG00061	840	839	Xu	sp|O64214|VG20_BPMD2 Gene 20 protein
VOG06673	1093	762	Xu	sp|Q05241|VG32_BPML5 Gene 32 protein
VOG00147	712	711	Xu	sp|O64216|VG22_BPMD2 Gene 22 protein
VOG00934	710	710	Xu	sp|O64262|VG69_BPMD2 Gene 69 protein
VOG00011	666	666	Xu	sp|P13004|YLYS_BPPHV Uncharacterized protein near lysin gene (Fragment)
VOG00437	651	650	Xu	sp|O64208|VG15_BPMD2 Gene 15 protein
VOG00153	664	641	Xu	sp|Q9T1Q7|VP41_BPAPS Putative protein p41

3.4 Analysis of Annotations based on Homologies to Proteins in the Swiss-Prot Database

Out of 2435 SP-annotated VOGs, 2033 are annotated based on homologies to viral SP-proteins. 341, 49 and 12 VOGs are annotated based on homologies to a bacterial, eukaryotic and archaeal proteins, respectively. The number of uninformative VOG descriptions is high when using a viral SP-protein to annotate the VOG, 992 out of 2033 virus-based annotation do not provide functional information. For annotations based on bacterial, eukaryotic and archaeal proteins, the number of uninformative descriptions is 105, 4 and 1, respectively (Table 11).

Table 11: Origin of Swiss-Prot annotations. VOGs are grouped based on the proteins used for annotation stemming from viruses, bacteria, eukaryotes or archaea. A count of the VOGs with uninformative descriptions in each category is shown in the last column.

Origin of Swiss-Prot-protein used for Annotation	# of Swiss-Prot annotated VOGs	% of Swiss-Prot annotated VOGs	# uninformative descriptions
Viruses	2033	83.49%	992
Bacteria	341	14.00%	105
Eukaryota	49	2.01%	4
Archaea	12	0.49%	1
total:	2435	100%	1102

3.4.1 Analysis of VOGs annotated with archaeal proteins

A more in-depth inspection of the 12 VOGs annotated with an archaeal SP-protein did not reveal protein function descriptions that might only be specific to archaea. Common function descriptions include the words “transferase”, “helicase”, “dehydrogenase” and “synthase”. These function are also frequently performed by viral proteins. VOG00379 and VOG03416 each contain a viral manually annotated SP-protein with matching annotations: “(putative) Peptidyl-tRNA hydrolase” for VOG00379, and “(probable) DNA polymerase sliding clamp 1” for VOG03416, even though the hosts of the viruses are eukaryotic. Both VOGs are annotated with proteins that are important across all domains of life.

3.4.2 Analysis of VOGs annotated with eukaryotic proteins

Out of the 49 VOGs annotated based on homologies to eukaryotic proteins, 22 also contain a viral SP-protein. For 15 VOGs the descriptions from the eukaryotic and the viral protein matched, for 5 VOGs the descriptions of the viral proteins were uninformative, i.e. “uncharacterized” or “domain-containing”, and 2 eukaryotic descriptions were specific to mitochondria. “Mitochondrial DNA mismatch repair protein mutS homolog” (VOG00128) has endonuclease activity and would be better described with the less specific description from the viral SP-protein “Probable HNH endonuclease”. “Mitochondrial chaperone BCS1” (VOG0074) is ATP-binding and the viral description “Putative AAA family ATPase L572” might also be more accurate here. The viral hosts are different from the organisms the annotations were derived from. As the SP database shows a certain bias towards the manual curation of entries from eukaryotic organisms (see Taxonomic Composition of the UniProtKB/Swiss-Prot Database), these entries do not always provide a good description of the consensus function of VOG proteins. Additionally, as viral proteins are often of shorter length than eukaryotic proteins, the alignments possibly cover the viral sequences well, but only partially align to the eukaryotic protein sequence (see Target Sequence Coverages of BLAST Alignments).

3.4.3 Analysis of VOGs annotated with bacterial proteins

341 VOGs are annotated with functional descriptions of bacterial proteins. This is little surprising, as the majority of proteins (97.5%, see Taxonomic Composition of the VOGDB) in the VOGDB belong to the order Caudovirales (tailed bacteriophages). As these viruses infect bacteria, the transfer of genetic material and thus the presence of similar sequences in

virus and host genomes is likely. 54 of the 341 VOGs contain at least one viral SP-protein. 37 VOGs showed a good consensus between the bacterial and viral annotations. Out of the 17 remaining VOGs, 9 have bacterial annotations containing the word "prophage". For these VOGs the functional descriptions from the viral SP-proteins are more informative. 4 bacterial annotations contain the word "uncharacterized", here the viral annotation would also be more accurate. Only 1 VOG had an "uncharacterized" viral annotation, with a more descriptive bacterial annotation. For the three remaining VOGs the SP entries were examined because the presence or absence of functional equivalence could not be derived from the names only. For two of these VOGs, the viral annotations were more descriptive than the low-quality bacterial annotations. In the third case, the bacterial protein description "Response regulator inhibitor for tor operon" is specific to Escherichia Coli. This protein also acts as an excisiokinase, and the viral description "Probable excisionase hkaC" would be more appropriate to describe the consensus function. 120 out of 132 proteins in the VOG come from Mycobacterium phages, and only two of the VOG proteins stem from Enterobacteria phages, for which the description as "inhibitor for tor operon" could be accurate. Overall, annotations based on bacterial SP-proteins are accurate, but annotations from viral proteins are more descriptive for viral functions. This is especially true if the bacterial proteins used are described as "prophage-derived uncharacterized protein".

A more general overview of the functional descriptions of bacterial SP annotations revealed that a large fraction (103/341) of the descriptions contain the words "uncharacterized" or "UPF". 11 additional proteins contain the word "prophage" and are associated with a poor functional description. The most frequent informative descriptions are transferase, nuclease, synthase/synthetase, toxin, reductase, transposase, hydrolase, kinase, ligase, methylase, polymerase and transport (Table 12). These functions are not specific to bacteria, as these proteins can also be encoded by viral genes. The remaining annotations were checked manually to see if they are bacteria-specific. More uninformative annotations were identified, such as "Protein NrdI" and "Protein TonB". In total there were nine VOGs annotated with a description starting with "Protein", which does not provide functional information. While homologies exist between viral proteins and not just their host, but also other cellular proteins, the transfer of functional information is often hampered by annotations being too vague or too specific to be valuable. This same issue was described by Mahmoudabadi et al, who searched for homology-based annotations of proteins from viral NCBI RefSeq genomes in cellular databases [162].

Table 12: Most frequent descriptions of VOGs annotated with bacterial Swiss-Prot proteins

Description	Number of Proteins
uncharacterized | UPF	103
transferase	18
nuclease	14
synthase | synthetase	12
prophage	11
toxin	9
reductase	8
transposase	8
hydrolase	7
kinase	7
ligase	7
methylase	7
polymerase	7
transport	7

3.4.4 Analysis of VOGs annotated with viral proteins

2033 out of the 2435 VOGs annotated with SP have annotations that are derived from viral SP entries. These VOGs were then divided into two groups for further analysis, the first group consisting of VOGs that contain viral SP-proteins, and the second group lacking such proteins. 1623 of the 1737 VOGs containing SP-proteins were annotated with one of the SP-proteins present in the VOG, while the remaining 114 VOGs were annotated with viral SP-proteins not present in the VOG itself. Figure 14 shows how VOGs are subdivided and analyzed based on the origin of the SP-protein they are annotated with. For the VOGs annotated with a SP-protein from that VOG, 1497 have annotations matching the consensus of the SP-proteins, while for the 114 VOGs annotated with SP-proteins not present in the VOG, 60 had matching descriptions to the consensus function based on the SP-proteins present in that VOG. Examination of the VOGs with no description match between the current annotation and the consensus annotation from SP-proteins present in the VOG revealed that most of the descriptions are similar, suggesting the same function (see Analysis of Functional Homogeneity of VOGs containing more than one Swiss-Prot Protein).

Further analysis of the 296 VOGs annotated with viral SP-proteins, but not containing a viral SP-protein revealed that 83 of the proteins used for annotation were present in another VOG, 136 proteins were used for VOG construction, but not placed in VOGs, 24 proteins were not included in the construction of VOGs, and 53 annotations come from polyproteins and have been excluded from the SP-proteins searched for in the VOGs. They could also be the results of BLAST alignments covering the whole query sequence but only a portion of the target sequence (see Target Sequence Coverages of BLAST Alignments).

It is possible that homologies are identified between proteins from different VOGs and therefore a VOG can be annotated with a SP-protein found in another VOG. An example is VOG00004, which does not contain any SP-proteins, but a sequence similarity search with BLAST [70] identifies significant hits to SP-proteins present in VOG06147.

The homology-based annotations with proteins that have not been assigned to VOGs mostly come from viral lineages that are not well-covered in the VOGDB. 67 annotations come from Baculoviridae with a family coverage of just 6% in the VOGDB, 42 come from Mimiviridae with a coverage of 3%, and 160 proteins do not have lineage information on the family level.

Overall the quality of homology-based annotations is accurate when compared with viral SP-proteins present in the VOGs. 1497 + 60 descriptions were a match, while 126 + 54 descriptions were not an exact match, but could be manually attributed to be of similar function, or one of the proteins having uncharacterized function or a description not describing the function. No obvious mismatch in the annotations could be observed.

The presence of annotations with SP-proteins not present in the VOGDB shows that homology-based annotations can be a valuable approach to increase the number of annotated VOGs, but annotations derived from distant taxonomic lineages are potentially lineage-specific.

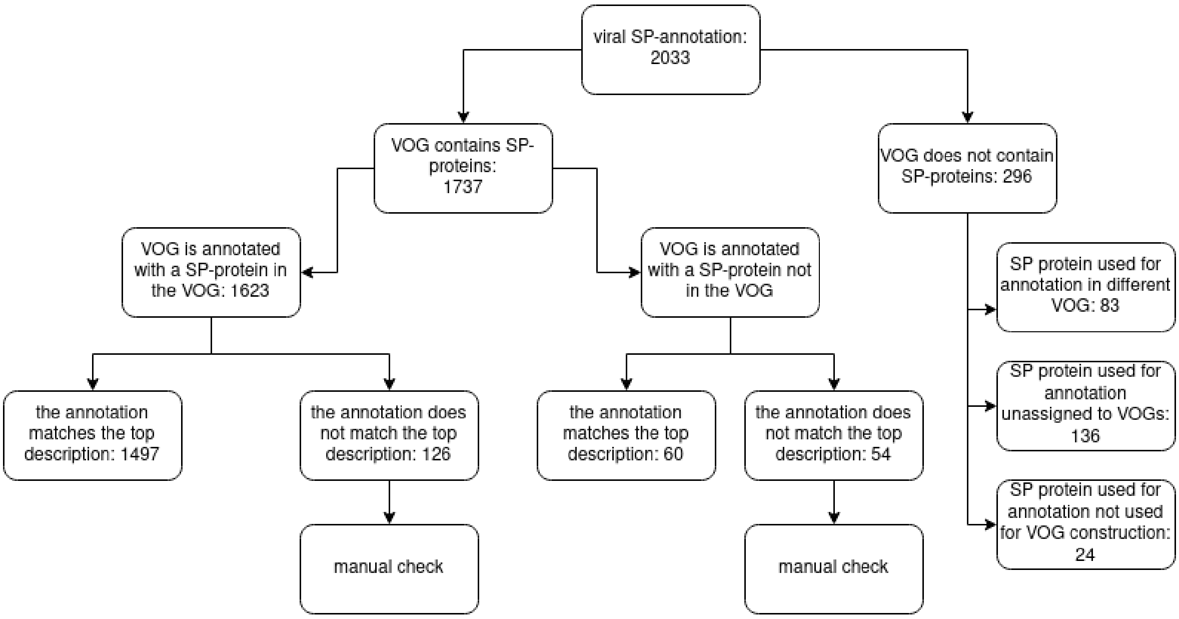

viral SP-annotation: 2033
VOG contains SP-proteins: 1737
VOG does not contain SP-proteins: 296
VOG is annotated with a SP-protein in the VOG: 1623
VOG is annotated with a SP-protein not in the VOG
SP protein used for annotation in different VOG: 83
the annotation matches the top description: 1497
the annotation does not match the top description: 126
the annotation matches the top description: 60
the annotation does not match the top description: 54
SP protein used for annotation unassigned to VOGs: 136
SP protein used for annotation not used for VOG construction: 24
manual check
manual check

Figure 14: VOG Annotations with viral Swiss-Prot Proteins. Flowchart describing the divisions of SP-annotations stemming from viral proteins. VOGs are divided first into VOGs that contain one or more SP-protein and those VOGs not containing SP-proteins. Further subdivisions indicate the origin of the protein used for annotation and the consensus between the functional descriptions. SP=Swiss-Prot

3.4.5 Target Sequence Coverages of BLAST Alignments

The analysis of average target sequence coverages for all BLAST hits of the protein used for annotation shows that most of the proteins are well-covered by the alignment. In the categories of annotation proteins from Archaea, SP-proteins in the VOG, SP-proteins not in the VOG and Bacteria, 83.33%, 91.56%, 71.05% and 66.76% of VOGs were annotated with a protein having an average BLAST coverage of >90%, respectively. For VOGs annotated based on eukaryotic proteins, the target sequence coverages are lower, with only 34.69% of all annotated VOGs having an average target sequence coverage >90%. Similarly, for VOGs that do not contain SP-proteins, the average target sequence coverage of the protein used for annotation is lower as well, with only 43.92% of VOGs exceeding an average BLAST target coverage >90% (Figure 15, S1).

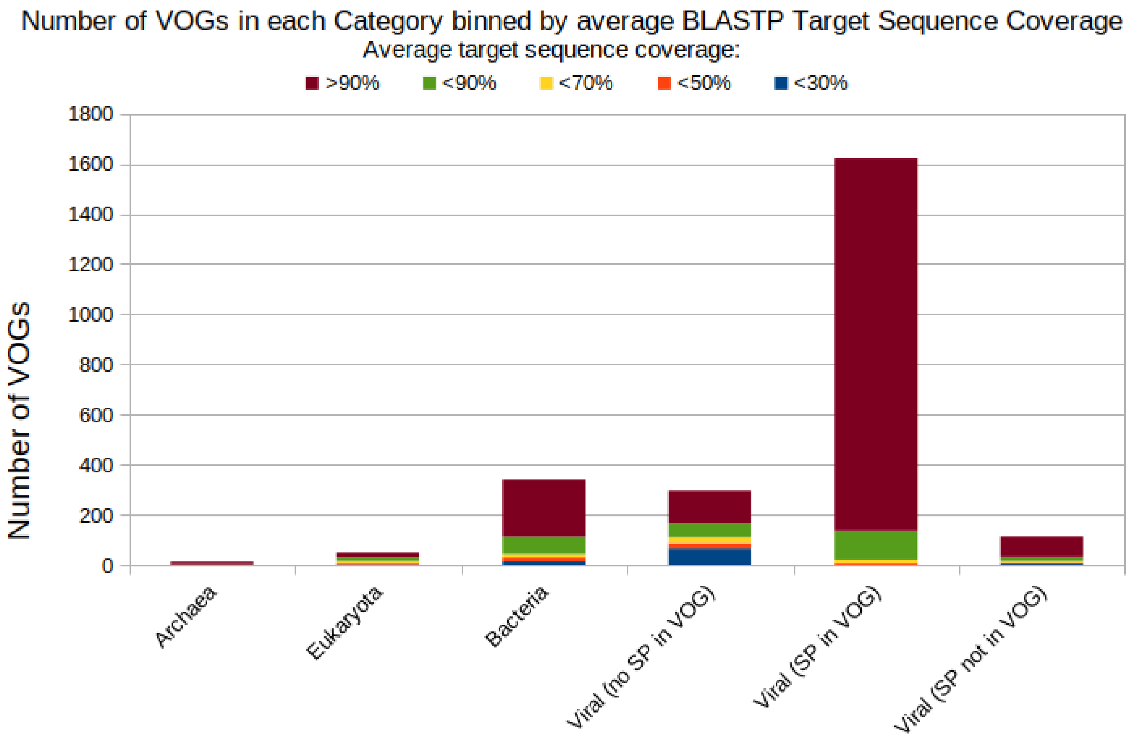


Figure 15: BLAST Average Target Sequence Coverage. VOGs were first categorized based on the origin of their annotation. Then the average BLAST target sequence coverage was determined for each VOG and VOGs were binned based on the coverage percentages. The dark purple bar indicates VOGs with an average target sequence coverage >90%. VOGs with target sequence coverages from 70-90% are shown in green, while yellow corresponds to coverages of 50-70%, orange to 30-50% and blue to average target sequence coverages <30%.

Interesting annotations are the ones inferred from homologies to SP-proteins with low target sequence coverages, as they could indicate that the query protein is aligned to a single domain and it is not a given that this domain is indicative of the function of the complete protein.

Several BLAST alignments with average target sequence coverages <30% are inspected to determine if a functional domain is present in the aligned region of the target sequence.

3 Eukaryotic Proteins with average target coverage < 30%:

VOG01678: Current Annotation: “sp | P08120 | CO4A1_DROME Collagen alpha-1(IV) chain”. This protein is 1779 amino acids long and has many “collagen triple-helix repeats” (IPR008160), as well as a “Collagen IV” domain in the C-terminal region. Individual BLAST hits are spread out along the sequence, covering the repeat domains as well as the Collagen IV domain.

VOG03160: Current Annotation: “sp | P08120 | CO4A1_DROME Collagen alpha-1(IV) chain”. The target protein is the same as in VOG01678, and BLAST hits again cover various regions of the protein and are not mapped to a single domain.

VOG22678: Current Annotation: “sp | Q93W20 | NIFU2_ARATH NifU-like protein 2, chloroplastic”. Analysis of the target sequence reveals that the VOG protein maps to the “NIF system FeS cluster assembly, NifU, C-terminal” domain (IPR001075). The sequence contains another copy of the same domain in the C-terminal region.

For both VOGs annotated with “sp | P08120 | CO4A1_DROME Collagen alpha-1(IV) chain”. The target sequence is covered well by different BLAST hits. In the third eukaryotic protein used to annotate VOG22678, the alignment maps to a functional domain of the protein. Therefore for all three annotations the transfer of function from the target sequence to the VOG is reasonable according to the presence of shared functional domains.

A closer look at BLAST hits of proteins in VOG00074:

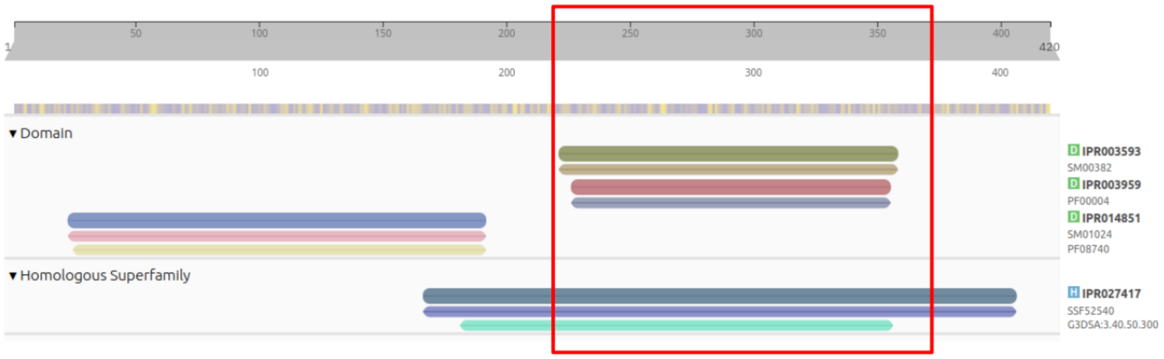


Figure 16: InterPro entry for sp | Q7ZV60 | BCS1_DANRE, currently used to annotate VOG00074. Different domains and homologous superfamilies are shown. The “AAA+” superfamily domain is shown in olive color, the “AAA+ ATPase” domain is depicted in brown along the same alignment coordinates [220-355]. The red rectangle indicates the alignment region.

The protein used for annotation is “sp | Q7ZV60 | BCS1_DANRE Mitochondrial chaperone BCS1” from Danio rerio. The aligned regions of the BLAST hits contain the “Mitochondrial chaperone” domain (PTHR23070:SF151, not shown), but also The “AAA+ superfamily of ATPases” (IPR003593) (Figure 16). This domain is found in all kingdoms of living organisms, participating in diverse cellular processes. Another domain found in the aligned region of the target sequence is the "P-loop containing nucleoside triphosphate hydrolases superfamily" (SSF52540, IPR027417). Running InterProScan [69] on all VOG00074 proteins revealed that this superfamily domain is present in 49 out of 51 proteins in the VOG. The same is true for InterPro domains IPR003959 and IPR003593 ("AAA+ ATPase domain"). The annotation of VOG00074 with those domains would be more representative of the consensus function than the current annotation with the eukaryotic protein.

1 Viral Swiss-Prot protein assigned to the VOG with average BLAST target coverage < 30%:
VOG05939: Current Annotation: "sp|Q5UNS9|COLL7_MIMIV Collagen-like protein 7". This protein has many "collagen triple-helix repeats" (IPR008160), and the BLAST alignments match the repeat regions. VOG05939 only contains 2 proteins, but they account for multiple significant BLAST hits to different regions of the protein.

64 Viral Swiss-Prot proteins used for annotation of VOGs that do not contain SP-proteins with average BLAST target coverage < 30%:
47 of the proteins used for annotation are Polyproteins, explaining the low target sequence coverage. Inspection of VOG06122, annotated as "Replicase polyprotein 1ab" revealed that the VOG proteins map to the region of the polyprotein that is described as "RNA-dependent RNA polymerase" (IPR044356). This VOG would more accurately be annotated based on the functional domain it aligns to in the polyprotein.
VOG14556 contains 3 proteins with a length of 55 amino acids. The endonuclease protein with which the VOG is annotated counts 245 amino acids. The queries do not align to a proper domain in the sequence.
VOG00049 is annotated as "sp|Q6XQB2|FIBER_BPT1 Probable tail fiber protein", which has a length of 728 residues, but the BLAST alignment only covers amino acids 1-118. Notably, there are no InterPro domains defined in the SP entry.

6 low target sequence coverage VOG annotations of VOGs containing SP-proteins but annotated with another protein:
VOG00314: Current Annotation: "sp|P00581|DPOL_BPT7 DNA-directed DNA polymerase". The SP-protein used for annotation is 704 amino acids long, but the BLAST alignments only cover roughly the first 200 residues. This region corresponds to a "Ribonuclease H-like

superfamily" (IPR012337, IPR036397). The "DNA-directed DNA polymerase" domain is located further towards the C-terminal of the target protein. The SP-protein in the VOG is the second most frequent BLAST hit, with alignments covering the full length of the target protein. The description "3'-5' exonuclease gp74" would therefore be more suitable, as the full domain (IPR002562) is covered.

VOG00421: Current Annotation: "sp | P13390 | FIBL1_BPT5 L-shaped tail fiber protein". This VOG contains the SP-protein "sp | Q6XQB2 | FIBER_BPT1 Probable tail fiber protein". The protein used for annotation is only partially covered by the alignment, as only 209 out of 1396 residues are aligned. There is no functional domain in that part of the sequence.

VOG00510: Current Annotation: "sp | Q5UPF8 | YL088_MIMIV Putative ankyrin repeat protein L88". The BLAST hits align to the different repeat regions in the protein, which contains several "Ankyrin repeats" (IPR002110).

VOG00895: Current Annotation: "sp | P26700 | FIBH_BPP2 Probable tail fiber protein". The BLAST hit partially covers the "tail fiber" domain (PTHR35191) of the target protein.

VOG01878: Current Annotation: "sp | G3FEX6 | POLG_JAEVM Genome polyprotein". This protein is annotated as a polyprotein, and the alignment coordinates (1505-2123) map to the peptidase and helicase domains in the target sequence (IPR001850, IPR014001, IPR001650).

VOG23482: Current Annotation: "sp | O64203 | ENLYS_BPMD2 Endolysin A". The alignment region does not cover any functional domains in the protein. The target protein does not have described domains in this region, but is flanked by two aligned regions to one of the VOG protein queries.

14 low target sequence coverage VOG annotations of VOGs annotated with bacterial proteins:

3 of the VOGs are inspected in greater detail:

VOG00207: Current Annotation: "sp | P15032 | RECE_ECOLI Exodeoxyribonuclease 8". The alignment maps to the "Putative exodeoxyribonuclease 8, PDDEXK-like" domain (IPR024432) of the target protein, which has a functional domain at its N-terminus that is not covered by the alignment.

VOG00328: Current Annotation: "sp | Q2FX77 | LYTO_STAA8 Probable autolysin LytO". The target protein has amidase and peptidase domains, and the alignment covers the "CHAP (cysteine, histidine-dependent amidohydrolases/peptidases)" domain in the N-terminal region.

VOG00354: Current Annotation: “sp | B2J384 | SYK_NOSP7 Lysine–tRNA ligase”. The alignment does not cover the “Lysine-tRNA ligase” domain of the target protein, but only a short “KTSC” domain (IPR025309) of about 60 amino acids. The VOG query protein has a length of 70 amino acids.

A closer look of BLAST hits of proteins in VOG00001:

VOG00001 is a large VOG containing 1059 proteins from 934 species. It is currently annotated as “sp | P03041 | RPC1_BPP22 Transcriptional activator protein C1”, which is a protein from Salmonella phage P22 (Bacteriophage P22).

Target sequence coverage of BLAST hits of the VOG proteins to the P03041 | RPC1_BPP22 protein was found to be high, the average sequence coverage being 97.5%. P03041 | RPC1_BPP22 belongs to the “Transcription activator CII” (IPR007933) family, with its sequence being completely covered by that InterPro domain. Additionally, its sequence coordinates 2-82 are covered by the “lambda repressor-like, DNA-binding domain” superfamily (IPR010982) domain. The InterPro entry for the IPR007933 family describes the CII protein as a transcription activator, which is conserved in bacteriophage lambda and related phages, playing a key role in the decision between lytic or lysogenic phage development.

The protein used to annotate the VOG is only 92 amino acids long, and many proteins assigned to VOG00001 exceed that length. A total of 625 VOG proteins exceed a length of 92 amino acids. Another BLAST search was performed for these proteins, and 4 proteins were hits to the P03041 | RPC1_BPP22 protein with an e-value < 1e-10. The query coverages were 76% for two proteins and 82% for the other two. All four alignments covered the whole target sequence.

The BLAST hits of the proteins longer than 102 amino acids with query coverages >90% are mostly “Uncharacterized HTH-type transcriptional regulator” (sp | A6U5H5 | Y045_SINMW) and “HTH-type transcriptional regulator” (multiple hits in the SP database).

Analysis of additional BLAST hits:

The domains of other BLAST hits were viewed on InterPro [163] to check for functional domains in the alignment region. Dominant domains in that region are the “Ner, winged helix-turn-helix DNA-binding domain” (IPR038722), which is a transcriptional regulator, as well as the “lambda repressor-like, DNA-binding domain superfamily” (IPR010982). These domains are found in viral proteins, as well as in bacterial proteins, e.g. in the Sugar fermentation stimulation protein B (sp | P0ACH4 | SFSB_SHIFL). The InterPro entry for the “lambda-repressor-like" InterPro Domain (IPR010982) states that "Bacteriophage lambda C1 repressor" controls the expression of viral genes as part of the lysogeny/lytic growth switch [163]. Other DNA-binding domains display similar structural folds to that of Lambda C1. These include bacterial regulators such as the "purine repressor (PurR)", the "lactose repressor (Lacr)" and the "fructose repressor (FruR)". This explains the identification of bacterial sugar fermentation proteins as BLAST hits, even though the function of these proteins would not accurately describe the VOG function.

A closer look at VOG00130:

VOG000130 is annotated with "sp|Q3T4L9|HOLIN_BPPRD Holin" from Enterobacteria phage. The SP-protein has been identified to be assigned to the VOG (query and subject coverage as well as sequence identity of 100%). 53 (e-value 1e-04) and 38 (e-value 1e-10) out of 78 proteins have the "LydA-like holin" InterPro domain (IPR032126). In the currently implemented annotation approach using sequence similarity, only four proteins had significant hits in the SP database. The most frequent SP-protein also contains the InterPro domain and is completely covered in three of the alignments and almost completely covered in the fourth alignment (coordinates 3-108). The annotation of VOG00130 seems to provide a reasonable consensus of the VOG protein function due to the presence of the same domain in the majority of the VOG proteins.

Examination of VOGs annotated with eukaryotic Swiss-Prot proteins with high target sequence coverages of BLAST alignments:

A closer look of BLAST hits of proteins in VOG05261:

This VOG is annotated with "sp|Q0Z972|IL10_CALJA Interleukin-10" from Callithrix jacchus. Homologs to human Interleukins and viral Interleukin homologs were also identified by BLAST. The coverages of the subjects are high (>90%). The homologs contain the "Interleukin-10/19/20/22/24/26 family" (IPR020443) domain. According to the information provided in the InterPro [163] entry, Proteins encoded by viruses such as Epstein-Barr virus and equine herpes virus show a high degree of sequence identity to "IL-10", and are thought to be involved in evasion of host immune responses. Inspection of the SP entry of a viral BLAST hit (SP-accession: P68677) showed that the protein aids in the evasion of the host immune system. This information is backed by GO terms and Keywords.

3.4.6 Percentage of VOG Member Proteins identifying the Swiss-Prot protein as a Homolog via BLAST

Figure 17a shows how well the SP-proteins used for annotation are represented in the BLAST search results. In all six categories, the fraction of VOGs with more than 90% of members identifying the SP-protein as a homolog is less than 40% (shown in purple). Additional analysis revealed that the average VOG size for VOGs with low percentages of members identifying the BLAST hit is larger than the VOG size for VOGs with high fractions of proteins identifying the SP-protein via BLAST (Figure 17b). This is reasonable, as larger VOGs display greater sequence divergence than VOGs that contain low numbers of proteins.

While annotations based on homologs in cellular organisms often provide a good insight into the VOG proteins' function, some of these annotations are specific to the organism. In these cases a domain-based annotation approach would be more appropriate.

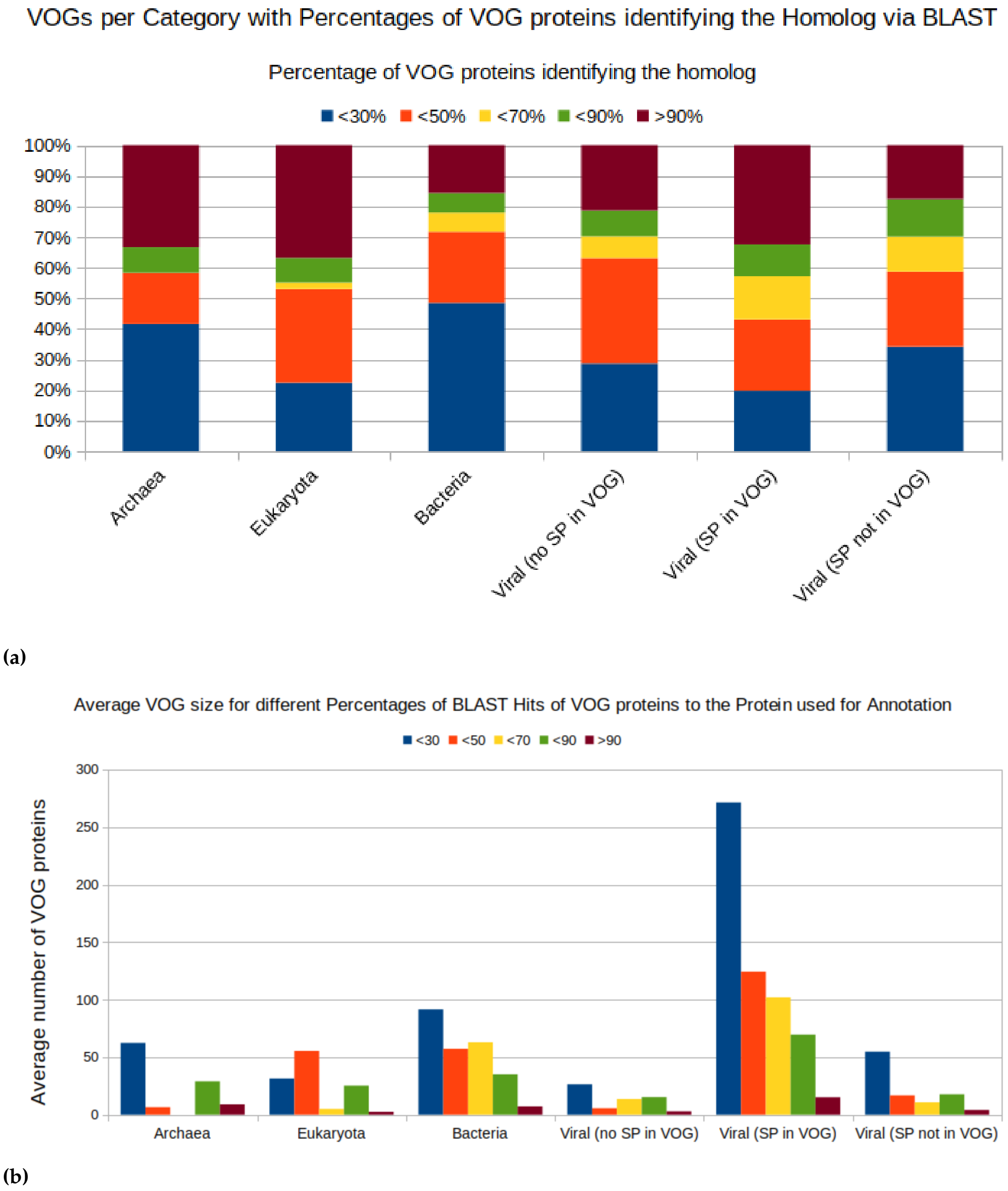
In the current annotation process, the query coverage is set to >90% for significant hits in the BLAST search, but the subject coverage is not considered. As viral proteins are often of short length, potential short alignments covering large portions of the query but not of the target sequence can be reported. This is especially interesting when performing homology searches to other organisms, that have longer protein sequences than viruses.

A closer look at the BLAST hits for VOGs currently annotated based on homologies to SP-proteins revealed that proteins with repeats are better covered than shown by sequence coverages of individual hits, because the same query protein maps to the target multiple times, thus covering it. This is true for e.g. Collagen alpha-1(IV) chain and Ankyrin repeats.

VOGs with proteins mapping to polyproteins are more accurately annotated with the functional domain in the aligned region. Many tail fiber proteins are long, with the alignment only covering parts of the target sequence. Most viral tail fiber proteins do not have described functional domains, rendering it difficult to support or reject functional relationships between query and target proteins. A target sequence coverage could be set to avoid transfer of functional information if the viral query protein only covers small portions of the target sequence. Here, a domain-based approach could be taken to only consider the matching domain for the annotation.

The closer inspection of the BLAST results from VOG00001 showed that even if target sequences are well-covered by an alignment, the domains can be connected to different functions in cellular organisms and viruses. Therefore not all proteins identified by BLAST with low e-values and high query and sequence coverages actually describe viral protein functions. As proteins from cellular organisms are much better covered in the SP database than viral proteins (see The UniProtKB/Swiss-Prot Database), a possible improvement in annotation quality could be the annotation with domain names rather than protein descriptions from the SP entries. This is especially true when transferring annotations form cellular organisms to viral groups based on homologies.

Even for VOGs that contain SP-proteins, this protein is not always identified as a homolog of other VOG members by BLAST. This is little surprising, as viral sequences are known to be divergent, which is why VOG clustering is based on remote homology concepts. This reveals that current annotations are based on a homology search that misses true positives. New approaches should apply remote homology searches to determine the consensus function of VOGs.



(a)

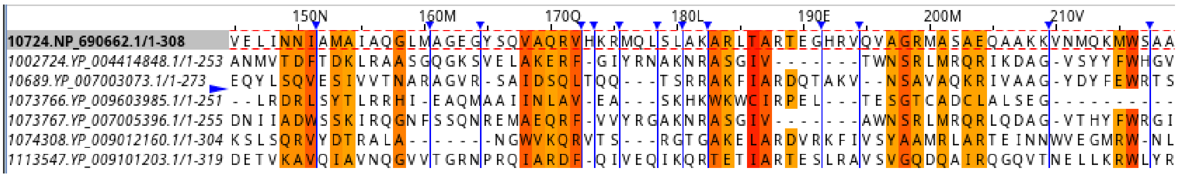
(b)

Figure 17: VOG Members identifying BLAST Hit and average VOG-size. a) Percentage of VOG proteins identifying the homolog in a BLAST search. Blue: Percentage of VOGs per category with less than 30% of member proteins identifying the SP-protein via BLAST. Orange: VOGs with 30%-50% of members identifying the SP-protein. Yellow: VOGs with 50%-70% of members identifying the SP-protein. Green: VOGs with 70%-90% of members identifying the SP-protein. Purple: VOGs with >90% of members identifying the SP-protein. **b)** Average number of VOG members binned by the fraction of proteins per VOG identifying the Swiss-Prot protein used for annotation via BLAST. The Coloring scheme is the same as described in a). SP=Swiss-Prot

3.5 Quality Check of the current Annotation Pipeline based on Literature-reviewed VOG Members

SPP1 Proteome:

Out of the 38 annotated genes, 8 were excluded due to being absent in the Genbank entry (genes 24.1*, 26*, 36.1, 37.1) or being unassigned to VOGs (genes 12, 22, 34, 44). For 18 out of the 30 remaining genes there was good consensus between the functional annotations of the protein and the VOG it is assigned to (Table 15). 6 of those proteins were assigned to VOGs that were annotated with the SP description of that same protein. Figure 18 shows a multiple sequence alignment of VOG02406 member proteins with the protein described by Godinho et al, which is also the protein used to annotate VOG02406, as the reference. The most conserved region in the alignment is in the region of 150-270AA, which corresponds to the "Phage head morphogenesis" domain (IPR006528) in the reference protein.



```
                                   150N      160M      170Q      180L      190E      200M      210V
10724.NP_690662.1/1-308            VELINNIAMAIAQGLMAGEGYSQVAQRVHKRMQLSLAKARLTARTEGHRVQVAGRMASAEQAAKKVNMQKMWSAA
1002724.YP_004414848.1/1-253       ANMVTDFTDKLRAASGQGKSVELAKERF-GIYRNAKNRASGIV-------TWNSRLMRQRIKDAG-VSYYFWHGV
10689.YP_007003073.1/1-273         EQYLSQVESIVVTNARAGVR-SAIDSQLTQQ---TSRRAKFIARDQTAKV--NSAVAQKRIVAAG-YDYFEWRTS
1073766.YP_009603985.1/1-251       --LRDRLSYTLRRHI-EAQMAAIINLAV-EA---SKHKWKWCIRPEL---TESGTCADCLALSEG----------
1073767.YP_007005396.1/1-255       DNIIADWSSKIRQGNFSSQNREMAEQRF-VVYRGAKNRASGIV-------AWNSRLMRQRLQDAG-VTHYFWRGI
1074308.YP_009012160.1/1-304       KSLSQRVYDTRALA-------NGWVKQRVTS---RGTGAKELARDVRKFIVSYAAMRLARTEINNWVEGMRW-NL
1113547.YP_009101203.1/1-319       DETVKAVQIAVNQGVVTGRNPRQIARDF-QIVEQIKQRTETIARTESLRAVSVGQDQAIRQGQVTNELLKRWLYR
```

Figure 18: VOG02406 Proteins aligned to the SPP1 VOG Member. Proteins of VOG02406 aligned to the SPP1 protein identified by Godinho et al as a initiation of infection protein (Swiss-Prot: Q38442 "Minor head protein GP7"). The alignment shows the conserved region at 150-210 Amino Acids. Only 6 out of the 545 VOG member proteins are shown in the alignment for better visibility

The remaining VOGs with agreeing functional descriptions had annotations based on SP-proteins other than the SPP1 protein. A special case occurred with gene 37.3, which is assigned to VOG00003, where the functional description of the VOG is "Excisionase", while Godinho et al described the protein as "putative DNA-binding protein". As excisionase proteins also bind DNA, the annotations were considered as matching. The 11 VOGs with no agreement with the functional description of their member protein described by Godinho et al had annotations that were either "REFSEQ hypothetical protein" (VOG01454, VOG00310 (contains 2 genes), VOG01488, VOG03856 and VOG01747), descriptions containing "propahge-derived" from bacterial SP-proteins (VOG00259 and VOG00413), or descriptions from SP that were not in complete agreement with the protein function predicted by Godinho et al (VOG06194, VOG00605, VOG00260, VOG07812).

VOGs with SP annotations not in agreement with the functions of the manually annotated proteins:
VOG06194 is annotated as "sp|Q05219|VG13_BPML5 Gene 13 protein", this SP entry has an annotation score of only 1, and the name does not provide information about the protein's function. Analysis of the entry in the SP database revealed that the "Gene 13 protein" belongs to the "terminase" family. The determined function of SPP1 gene product 2 is also "Terminase", which would be a more accurate description for the VOG function. VOG06194 would be considered as under-annotated, as the functional description does not reflect the true function of the proteins in the VOG.
VOG00605 is annotated as "sp|P19727|CAPSB_BPT7 Minor capsid protein", while the corresponding SPP1 protein is described as "major capsid protein". There is, however, a partial agreement between the descriptions, and consensus that the protein is a capsid protein.
VOG00260 is annotated as "sp|P16009|NEEDL_BPT4 Pre-baseplate central spike protein Gp5". The annotation score for this entry is 5, thus the annotation is very specific for that bacteriophage and might not represent the consensus function of the VOG. The gene product from the SPP1 phage that is assigned to VOG00260 is described functionally with "endolysin, cell lysis". For this VOG a more generic functional description as "lysin" or "endolysin" would be more appropriate to describe the VOG consensus function. VOG00260 is an example of an over-annotated VOG.
VOG07812 is annotated as "sp|P36549|YAF2_BACLI Uncharacterized 9.7 kDa protein in cwlL 5'region". This annotation comes from a bacterial protein, and the description does not provide information about the protein's function as a cell lysin. This VOG is an example of under-annotation with a non-viral protein.
For none of the VOGs there was a complete contradiction between the VOG consensus function and the functional description of the SPP1 protein assigned to it.

VOGs annotated as "REFSEQ hypothetical protein": Only one of the 5 VOGs annotated with "REFSEQ hypothetical protein" contains no proteins with descriptions other than "hypothetical protein". The four remaining VOGs contain some proteins with RefSeq descriptions that match the SPP1 protein. Notably, some proteins that are found in the SP database, are also described as "hypothetical proteins" in the RefSeq database. Therefore the RefSeq description does not always accurately represent what is known about the protein's function.
VOG00310 is annotated as a "phage tail assembly protein" in the paper, while the VOG's consensus function is "hypothetical protein". A closer inspection of the RefSeq names of

VOG00310 members revealed that out of 141 member proteins, 46 have names associated with tail proteins.

Additional literature-reviewed proteins:
Out of eight proteins belonging to eight different VOGs, six had functional descriptions matching the VOG function (Table 16). The VOGs with descriptions not matching the manual annotations were VOG06194 and VOG00260. These VOGs also contain a manually curated SPP1 protein, for which the annotations match those of the proteins in the non-SPP1 publications. The discrepancies for these VOGs are described above.

Based on the analysis of the functionally described SPP1 and literature-reviewed proteins, VOG annotations with viral SP-proteins are mostly in agreement with the manually curated protein annotations. While annotations from SP entries with low annotation scores do not always provide a lot of information about the protein's function, as protein names do not always reflect its function (VOG06194, VOG07812), using SP proteins with high annotation scores to describe a VOG's function might not be a good representation of the consensus function of all proteins in the VOG, as the description is rather specific (VOG00260).

Annotations derived from bacterial SP-proteins often do not provide informative descriptions, as they frequently contain the terms "prophage-derived" or "uncharacterized".

For RefSeq-based annotations, the most frequent description is often "hypothetical protein", but the next-frequent RefSeq descriptions matched the function of the SPP1 protein. Notably, some proteins that are described as "hypothetical proteins" can also be found in the SP database with a functional description. Examples are the proteins encoded by genes 17 (RefSeq Accession: NP_690679.1, SP Accession: O48448), and 17.1 (RefSeq Accessions: NP_690680.1, YP_710298.1, SP Accession: O48449). The current VOG annotation could potentially be improved by also considering the second and third most frequent RefSeq descriptions, as the most frequent ones are often "hypothetical" or "ORF". A score could be added to the annotations to show how many of the VOG proteins share the same RefSeq description.

This literature-based annotation quality check highlighted that VOG annotations based on SP-protein names can be both too specific, or not provide a high level of functional information.

3.6 Possible Improvements of VOG Annotations

One of the main problems of the current VOG annotations is that many VOGs with SP-based annotations do not have informative functional descriptions (see Overview of the current VOG Annotation). For VOGs annotated with viral SP-proteins the current annotation does not provide information about the protein used for annotation being present in the VOG or if the VOG function was inferred by homology due to sequence similarity to a SP-protein not present in the VOG. A new approach would include different confidence intervals of the annotations, i.e. it will be specified if the protein(s) used for annotation are present in the VOG, or if they have been identified as homologs to the VOG proteins. Strict cutoffs for query and sequence coverages will be applied to the homology searches to ensure high quality annotations as well as to avoid annotations with polyproteins. The current BLAST-based annotations are likely to miss remote homologies with low sequence identities that are often present between viral proteins due to rapid evolution and high mutations rates (see Introduction). Thus many of the proteins identified by a BLAST homology search are potentially significant hits for only a low number of VOG member proteins (Figure 17a). Additionally, some VOG annotations are derived from cellular organisms, and even though the target sequence coverages are high and the VOG proteins and the target protein contain the same domains, functional descriptions of identical domains can vary between eukaryotic, bacterial and viral proteins. Therefore annotations coming from viral proteins should be favored over those coming from cellular organisms.

Domain-based annotations could be applied to VOGs with no homologs in the SP database meeting the filtering criteria.

In the current RefSeq-based VOG annotation, the most frequent RefSeq description in a VOG is used to describe the function. “hypothetical protein” is the default description in RefSeq, and it is often not changed to a more appropriate one, even if the protein function has been manually curated. A few random examples are NP_043509.1, YP_002004528.1, NP_043508.1, all of which are in the SP database, but remain “hypothetical proteins” in the RefSeq database. As a result, the current annotation approach describes many VOGs with “hypothetical protein”. The fasta files containing VOG proteins can contain duplicates, which are currently not removed in the annotation process. Additionally a count of how many proteins in the VOG share that description can be added.

A description of the newly implemented VOG annotation pipeline is detailed in the Methods section.

3.7 Quality Checks of the new Annotation Approaches

3.7.1 Quality Check of the new Annotation Approach based on Swiss-Prot VOG Members

The annotation pipeline ensures that proteins with uninformative descriptions are no longer used to annotate the VOG. Because of the functional homogeneity of SP-proteins (see Analysis of Functional Homogeneity of VOGs containing more than one SP Protein) in VOGs, protein descriptions from SP entries that provide higher levels of information can be used. As SP-proteins in the VOG are already assumed to be a high quality standard in terms of functional description of the VOG, several MSAs of VOGs with manually curated SPP1 proteins were inspected to check how well the protein used for annotation is conserved in the alignment. While some alignments are of high quality with high conservation and consensus such as the MSA for VOG07812 (Figure 19), other alignments, especially when belonging to larger VOGs with long proteins are less conserved. Figure 19 displays a segment of the MSA for VOG00632, a large VOG comprised of almost 4000 proteins, most of which exceed a length of 500 amino acids. Conserved regions are not visible in the MSA, as would be expected for large alignments with little sequence identity. From this alignment the quality of the VOG annotation with the SP-proteins assigned to them cannot be judged. The annotation, however, matches the function of the SPP1 protein in the VOG. This is true for all 16 VOGs containing both SP-proteins and manually annotated SPP1 proteins, even if in some cases the two proteins are identical, as several of the SPP1 proteins are found in the SP database.

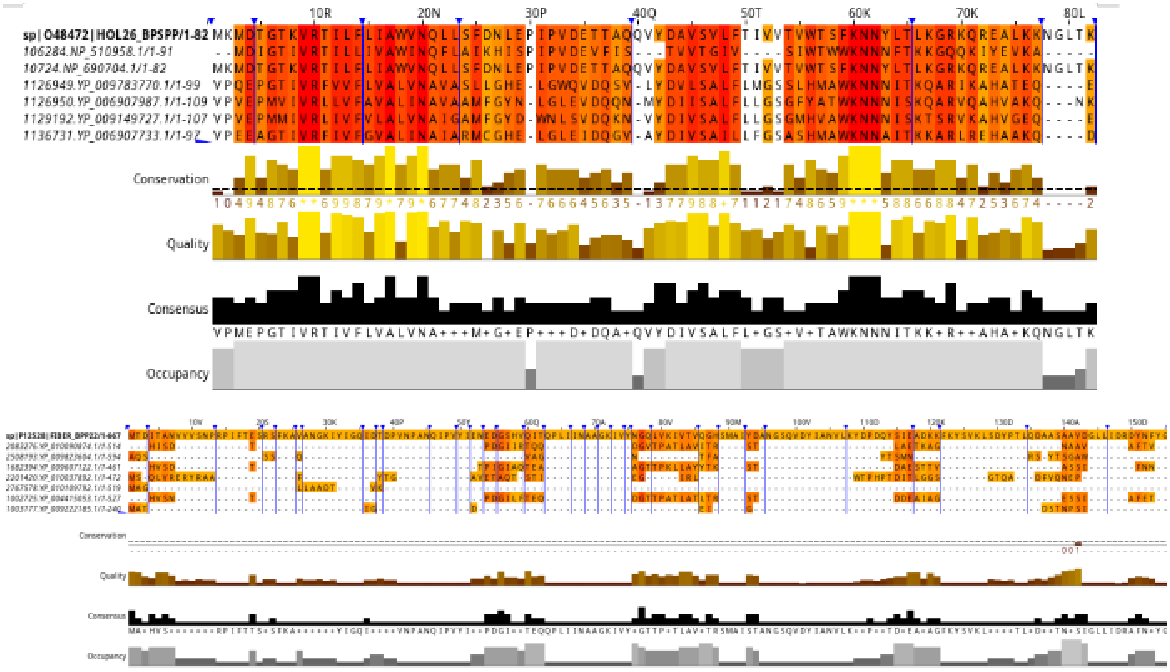

sp|O48472|HOL26_BPSPP/1-82
106284.NP_510958.1/1-91
10724.NP_690704.1/1-82
1126949.YP_009783770.1/1-99
1126950.YP_006907987.1/1-109
1129192.YP_009149727.1/1-107
1136731.YP_006907733.1/1-97
Conservation
Quality
Consensus
Occupancy
sp|P12528|FIBER_BPP22/1-667
Conservation
Quality
Consensus
Occupancy

Figure 19: MSAs of VOG07812 and VOG00632. top) Proteins of VOG07812 aligned with ClustalW. A VOG member protein used for the annotation (SP accession: O48472, Putative antiholin) is used as the reference and shown in the first row of the alignment, hiding insertions. The alignment shows the high conservation over the whole length of the protein. Only 6 out of the 157 VOG member proteins are shown for better visibility. **bottom)** Alignment of VOG00632 members. A SP-protein (Accession: P12528, Tail spike protein) used for annotation is shown as the reference in row 1. Redundancy has been removed from the alignment. Only 8 proteins are shown to keep the visibility high. The quality, consensus and conservation of the alignment are low. Higher conserved residues are shown in darker shades of red.

3.7.2 Quality Check of the new Annotation Approach with hmmsearch

A quality check of the remote homology based annotation approach was performed by comparing the annotations inferred by remote homology to the annotations of the 952 VOGs annotated with SP-proteins assigned to them. Out of 952 SP-annotated VOGs, the hmmsearch [72] approach identified hits for 794 of them, while no significant homologs were found for 158 VOGs. Hmmsearch does not identify all primary sequences that were used to build the HMM. This is especially true for large VOGs, where SP-proteins in the VOG were not detected. An example is VOG00001, where a BLAST [90] search was performed as described in Identification of Viral Swiss-Prot Proteins in VOGs, but with a word size of 4. The BLAST search identified 18 SP-proteins meeting the filtering criteria, but only 9 of them were found in the hmmsearch output (e-value 1e-04, all other parameters are the default parameters). For VOG00003 only 6 out of 9 SP-proteins in the VOG are found by hmmsearch.

In addition, the strict query profile coverage and target sequence coverage cut-offs limit the number of results further, while ensuring that hits do not cover just one small domain of the protein.

On the small dataset of SPP1 proteins, the hmmsearch-based annotation described 10 VOGs according to the manually annotated functions. For the remaining 20 VOGs hmmsearch did not provide an annotation. None of the VOGs were annotated contradictory to the SPP1 protein's function.

When compared with BLAST, hmmsearch was able to identify homologous hits that could not be detected with BLAST. An example is the existence of homologous relationships between VOG02561 proteins and endonuclease (SP accession: P32286). None of the proteins identified the endonuclease as a homolog with BLAST. InterProScan [69] confirmed the presence of endonuclease domains in two out of the four VOG proteins. BLAST, however, performed superior in VOGs containing SP-proteins, as those are by default always identified as hits with the more lax BLASTP cut-off values than in the original search to identify the curated proteins in the VOGDB (see Identification of Viral Swiss-Prot Proteins in VOGs). Nevertheless hmmsearch provides valuable information about remote homologies, being able to identify SP-proteins not in the VOGDB that can be used to annotate VOGs.

3.7.3 Quality Check of the new Annotation Approach with InterProScan

A quality check of the InterProScan [69] domain-based annotation was performed using the manually annotated SPP1 proteins. In 25 out of 30 VOGs the most frequent InterPro [163] domains describe the same function as the SPP1 protein. Some domains are well-covered in the VOG, e.g. the Portal protein is present in 956 out of 1166 proteins in VOG00039. Others have lower coverages, such as the "IPR024659 Major capsid protein Gp5" domain, which was detected in only 188 out of 1058 VOG00605 proteins, or the "Receptor-binding domain of short tail fiber protein gp12", detected in 317 out of 3951 VOG00632 proteins. Despite the low coverage, there is a match between the manually curated SPP1 function and the InterPro domain function. Two VOGs (00003 and 00621) had InterPro descriptions more generic than the SPP1 protein's function. Less frequent InterPro domains, however, matched the function of the SPP1 protein. For VOGs 00413 and 03856 the function of the SPP1 protein was described as "putative" and no comparison could be made. Only one VOG (VOG01488) had no domains detected in its proteins, and the SPP1 protein function is also not described in much detail, as it is a "putative DNA binding protein". In the 30 analyzed VOGs there were no contradictions between the InterPro annotation and the SPP1 protein annotation, however, for 5 VOGs no accurate comparison could be made due to the lack of information in either the SPP1 protein annotation or the InterPro annotation. While even domains with low coverages in VOGs match the description, the number of VOG proteins containing that domain can be regarded as a confidence score for the InterPro annotation.

3.7.4 Quality Check of the new Annotation Approach with RefSeq Names

The quality of the new RefSeq annotation is validated by comparing descriptions extracted with the current RefSeq annotation approach and the new approach with the functional descriptions of VOGs annotated with SP-proteins assigned to them. In 235 of the 961 VOGs containing SP-proteins the new and current RefSeq annotation approaches yield different descriptions. Analysis of these VOGs revealed that the new RefSeq annotation matched the SP-based annotation in 166 cases, was less specific than the SP-based annotation in 35 cases while being more specific in 7 VOGs, and contradicting it in just one VOG. The remaining 26 VOGs had descriptions such as "domain-containing protein" and "160 kDa protein" that could not be compared on the functional level and thus these VOGs could not be assigned to any category. Notably, the current RefSeq annotation approach described 202 of the 235 VOGs as "hypothetical protein", while with the new approach there were no VOGs described as "hypothetical protein". Only in VOG03302 the current RefSeq annotation "putative helicase" was a more accurate match to the SP-inferred annotation "Putative helicase R592" than the description "leucine rich repeat gene family" provided by the new annotation.

18 of the 961 VOGs were described as "hypothetical protein" with the new RefSeq annotation due to the absence of other descriptions in the VOG member descriptions. With the previous approach 220 VOGs were described as "hypothetical protein".

Additionally the RefSeq annotations were compared to the functional descriptions of SPP1 proteins belonging to VOGs (see Representation of Viral Swiss-Prot Proteins in the VOGDB). The current RefSeq annotation approach provided descriptions matching the SPP1 protein's function in 15 in VOGs, provided less specific annotations in 14 cases (12x "hypothetical protein", "DUF3168 domain-containing protein", "RecT-like ssDNA binding protein"), and for VOG02406 the descriptions "initiation of infection; binds to portal" and "head morphogenesis protein" were neither contradicting nor clearly describing the same function. There were no annotations that were completely contradictory. The new RefSeq annotation provided matching descriptions for 23 VOGs, while providing less descriptive annotations for 4 VOGs (1x "hypothetical protein", "DUF3168 domain-containing protein", "RecT-like ssDNA binding protein", "phage protein"), two more detailed descriptions, and VOG02406 again remained unassigned. The SPP1-proteins belonging to the VOGs where RefSeq annotations contained a higher information content were both annotated as "putative" by Godinho et al [161]. The new RefSeq annotation also did not provide any descriptions contradicting the SPP1-inferred function.

Overall, there is reasonable consensus between the novel RefSeq annotations and the annotations based on SP-proteins and manually annotated SPP1 proteins assigned to the VOGs. This renders the new approach a valid strategy to assign functional descriptions to VOGs that could not be annotated based on SP-proteins assigned to them or homologous proteins in the SP database. In addition, the reported number of VOG proteins sharing that name allows users to choose VOGs based on different levels of consensus.

3.8 Overview of the new VOG Annotation

With the new annotation approach, no VOGs are annotated based on homologies to cellular organisms, as functional descriptions were at times not applicable to viral proteins, and for some inferred homologies the alignment of the short viral protein to the target sequence only covered a small portion of the target. A total of 952 VOGs that contain SP proteins were annotated based on functional information from those members. These VOGs contain a total of 29.75% of all VOG proteins, meaning that almost one third of the VOGDB has high quality annotations. Homology-based annotations with hmmsearch attributed functions to 123 additional VOGs, combining for a modest total of 906 proteins. Closer inspection of those VOGs showed that 32 of them were annotated based on SP-proteins assigned to other VOGs, while 69 and 14 VOGs were annotated based on proteins unassigned to VOGs and proteins not considered for VOG construction, respectively. Homology search for the remaining homology-based annotated VOGs identified homolog SP-proteins sharing the same description but belonging to more than one of the classification classes of SP-proteins present in other VOGs, unassigned SP-proteins, and excluded SP-proteins. 3731 VOGs contain members with Pfam, Prosite of SUPERFAMILY domains and were described based on those domains. 4105 of the remaining VOGs had assigned proteins with RefSeq descriptions other than the default "hypothetical protein", these VOGs contain 20.04% of all proteins. The fraction of VOGs annotated as "hypothetical protein" remains large, counting more than two thirds of all VOGs, with 27.28% of all proteins in the VOGDB (Table 13, Figure 20).

Table 13: Overview of the new VOG Annotation. VOGs annotated with different approaches with numbers of VOGs and proteins assigned to them, as well as percentages of total VOGs and proteins in the VOGDB are shown.

Annotation	# VOGs	# proteins	% VOGs	% proteins
Annotation with SP protein in VOG	952	130571	3.35%	29.75%
Homology-based annotation	123	906	0.43%	0.20%
Domain-based annotation	3731	99648	13.14%	22.7%
RefSeq Annotation not "hypothetical protein"	4105	87967	14.46%	20.04%
RefSeq hypothetical protein	19475	119760	68.60%	27.28%

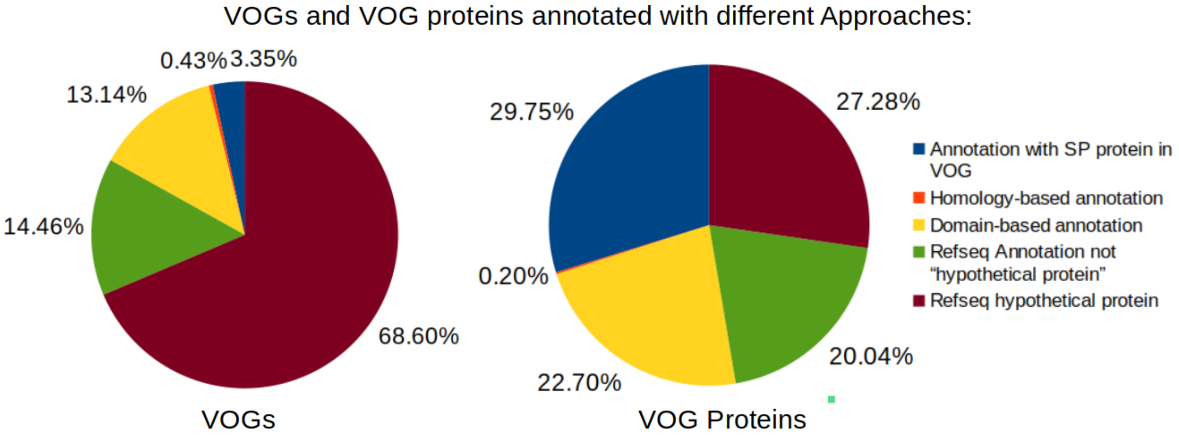


Figure 20: 68.60% of all VOGs are annotated as "hypothetical protein" based on protein descriptions from RefSeq, containing 27.28% of all VOG proteins (dark red). 14.46% of VOGs are annotated based on other RefSeq descriptions and contain 20.04% of all proteins (green). Only a fraction of VOGs (3.35%) is annotated based on VOG members in the UniProtKB/Swiss-Prot database that do not contain the patterns specified as uninformative in the description, but they account for 29.75% of all proteins (blue). An additional 0.43% of VOGs are annotated based on homologies to SP-proteins, containing just 0.2% of member proteins (red). 13.14% of VOGs have domain-based annotations. 22.70% of the VOG proteins are assigned to those VOGs (yellow). SP=Swiss-Prot.

3.8.1 Annotation of the Top 10 most wanted VOGs

With the new annotation approach functional annotations could be assigned to four of the ten VOGs (VOG06194, VOG06147, VOG00011 and VOG00153) based on SP-proteins in the VOG with descriptions of high information content. 5 VOGs (VOG02932, VOG00061, VOG00147, VOG00934 and VOG00437) are annotated based on domains present in member proteins, however, only in VOG00934 is the domain present in a significant portion of members. For the remaining VOGs the domains were identified in few of the member proteins. Just one of the ten VOGs is annotated based on RefSeq descriptions with the new approach – VOG06673 is now described as a minor tail protein, with about one third of all VOG members sharing that RefSeq name. Table 14 shows the 10 VOGs with the highest number of species that did not contain informative descriptions with the previous annotation approach. The new annotations are shown in the rightmost column. The comparison of current and new annotations for theses VOGs highlights the troublesome information transfer based on SP-proteins that have a poor entry description. Often a VOG has SP-proteins of both high and low annotation quality assigned to it, and filtering out uninformative descriptions can yield more meaningful functional annotations.

Table 14: Current and new annotations of the top 10 most wanted VOGs. Only partial new annotations including the annotation source and the most frequent domain are shown to enhance readability of the table. # Prot: number of proteins in the VOG. # Spec: number of VOG species.

VOG	# Prot	# Spec	Consensus Functional Description (current)	Consensus Functional Description (new)
VOG06194	3200	3120	sp|Q05219|VG13_BPML5 Gene 13 protein	Terminase, large subunit (SP in VOG)
VOG06147	7764	3092	sp|O64200|VG04_BPMD2 Gene 4 protein	DNA endonuclease I-HmuI (SP in VOG)
VOG02932	953	952	sp|Q05240|VG31_BPML5 Gene 31 protein	SSF49785 Galactose-binding domain-like IPR008979 Galactose-binding-like domain superfamily (11 out of 953 proteins have this domain)
VOG00061	840	839	sp|O64214|VG20_BPMD2 Gene 20 protein	PF05521 Phage head-tail joining protein IPR008767 Bacteriophage SPP1, head-tail adaptor (2 out of 840 proteins have this domain)
VOG06673	1093	762	sp|Q05241|VG32_BPML5 Gene 32 protein	minor tail protein (315 out of 1093 RefSeq proteins have this description)
VOG00147	712	711	sp|O64216|VG22_BPMD2 Gene 22 protein	PF11367 Protein of unknown function (DUF3168) IPR021508 Tail completion protein (3 out of 712 proteins have this domain)
VOG00934	710	710	sp|O64262|VG69_BPMD2 Gene 69 protein	PF12705 PD-(D/E)XK nuclease superfamily IPR038726 PD-(D/E)XK endonuclease-like domain, AddAB-type (508 out of 710 proteins have this domain)
VOG00011	666	666	sp|P13004|YLYS_BPPHV Uncharacterized protein near lysin gene (Fragment)	Holliday junction resolvase (SP in VOG)

Continued on next page

Table 14 – *Continued from previous page*

VOG	# Prot	# Spec	Consensus Functional Description (current)	Consensus Functional Description (new)
VOG00437	651	650	sp|O64208|VG15_BPMD2 Gene 15 protein	PF18451 Contact-dependent growth inhibition CdiA C-terminal domain IPR040559 tRNA nuclease CdiA, C-terminal (1 out of 651 proteins have this domain)
VOG00153	664	641	sp|Q9T1Q7|VP41_BPAPS Putative protein p41	Early transcription factor 70 kDa subunit (SP in VOG)

3.9 Quality Check of the new Annotation Pipeline based on Literature-reviewed VOG Members

Table 15 shows a comprehensive overview of all proteins annotated by Godinho et al [161] that have been placed in VOGs (VOGDB version 208). New annotations that match the protein function described in the paper better than the current VOG annotations are colored in green. 3 VOGs (VOG00310, VOG01454 and VOG01747) previously annotated as "REFSEQ hypothetical protein" now have domain-based annotations that match the SPP1 protein. While most VOGs that contain SP-proteins or are annotated based on homologies to SP-proteins have matching descriptions, some improvement in annotation quality could be observed in a few VOGs. VOGs 00259 and 07812 have previously been annotated based on homologies to bacterial SP-proteins, which provided no functional information. In the new approach SP-proteins assigned to the VOG provide more functional information. The annotation of VOG00260 has previously been identified as an over-annotation, providing much detail about protein function. The new description "Endolysin" is based on the consensus of several manually curated SP-proteins assigned to the VOG. The consensus function of VOG00605 changes from "minor capsid protein" to "major capsid protein", and VOG06194 receives a more informative description, as the new annotation approach disregards SP descriptions with little informative content such as "Gene X" or "uncharacterized protein" in the process. No improvement could be seen in VOG00758. This VOG has been described as a Holin-like protein based on homology, and is now annotated based on a Haemolysin domain (IPR019715), which is found in bacteria and also in the viral Holin-like protein. A possible cause for this is that the protein did not pass the strict quality filters of the homology-based annotation. VOG03856 is now annotated as a DNA double-strand break repair helicase based on homology, but this function cannot be verified based on the vague description of the SPP1 protein as "putative ATP-binding protein". For VOGs annotated based on SP-proteins, it is indicated in parentheses if the Protein is assigned to the VOG or if it was detected to be a homolog to the VOG proteins. The RefSeq-based annotations now contain information about how many of the total VOG proteins share that description (see VOG01488). With the new annotation approach all VOG descriptions for the non-SPP1-proteins were a match to the manually annotated proteins (Table 16). Overall, the most significant improvement could be observed for VOGs that contain SP-proteins or homologs, but have been annotated with a SP-protein with an uninformative description.

Table 15: SPP1 Proteins - Current and new Annotation. Comparison of current and new VOG annotation approaches with the functions of SPP1 proteins functionally described by Godinho et al. New annotations show the consensus functional description and the origin of the annotation. Annotations based on Swiss-Prot proteins are shortened in the table, as they also contain Keyword and GO-term information, as well as a list of the identifiers of all Swiss-Prot proteins assigned to the VOG.

VOG	Protein Function (Godinho et al)	VOG current Functional Description	VOG new Functional Description
VOG00003	putative DNA binding protein	sp | O22001 | VXIS_BPMD2 Excisionase	Excisionase (SP in VOG)
VOG00039	st.; portal protein	sp | Q05220 | PORTL_BPML5 Portal protein	Portal protein (SP in VOG)
VOG00046	st.; connector stopper protein	sp | O48446 | HCP16_BPSPP Head completion protein gp16	Head completion protein gp16 (SP in VOG)
VOG00145	replicative DNA helicase; binds host DnaG and DnaX	sp | P04530 | HELIC_BPT4 DnaB-like replicative helicase	DNA helicase/primase (SP in VOG)
VOG00234	st.; connector adaptor protein	sp | Q38584 | HCP15_BPSPP Head completion protein gp15	Head completion protein gp15 (SP in VOG)
VOG00259	st.; distal tail protein (Dit)	sp | O31977 | YOMH_BACSU Spbetaprophage-derived uncharacterized protein YomH	Distal tail protein (SP in VOG)
VOG00260	endolysin; cell lysis	sp | P16009 | NEEDL_BPT4 Pre-baseplate central spike protein Gp5	Endolysin (SP in VOG)
VOG00310	tail chaperone protein	REFSEQ hypothetical protein	PF12363 Phage tail assembly chaperone protein, TAC IPR024410 Phage tail assembly chaperone protein, TAC (105 out of 141 proteins have this domain)
VOG00330	5'-3' exonuclease	sp | P03697 | EXO_LAMBD Exonuclease	Exonuclease (SP in VOG)

Continued on next page

Table 15 – *Continued from previous page*

VOG	Protein Function (Godinho et al)	VOG current Functional Description	VOG new Functional Description
VOG00413	putative bacteria surface binding protein	sp|O31912|YORA_BACSU SPbeta prophage-derived uncharacterized protein YorA	SSF51126 Pectin lyase-like IPR011050 Pectin lyase fold/virulence factor (30 out of 31 proteins have this domain); PF13229 Right handed beta helix region IPR039448 Right handed beta helix domain (21 out of 31 proteins have this domain); PF05048 Periplasmic copper-binding protein (NosD) IPR007742 Periplasmic copper-binding protein NosD, beta helix domain (3 out of 31 proteins have this domain); PF12708 Pectate lyase superfamily protein IPR024535 Pectate lyase superfamily protein (3 out of 31 proteins have this domain)
VOG00605	st.; major capsid protein (MCP)	sp|P19727|CAPSB_BPT7 Minor capsid protein	Major capsid protein (SP in VOG)
VOG00621	replicative DNA helicase; binds host DnaG and DnaX	sp|P37469|DNAC_BACSU Replicative DNA helicase	Replicative DNA helicase (Homology-based)
VOG00632	st.; tail tip protein; Tal; anti-receptor protein	sp|P18771|FIBP_BPT4 Long-tail fiber proximal subunit	Tail spike protein (SP in VOG)
VOG00758	component of holin; cell lysis	sp|O48470|HOL24_BPSPP Holin-like protein 24.1	PF10779, Haemolysin XhlA, IPR019715, Haemolysin XhlA, - (9/17 proteins have this domain);
VOG01025	st.; tape measure protein (TMP)	sp|Q6XQC4|TMP_BPT1 Tape measure protein	Tape measure protein (SP in VOG)
VOG01454	putative tail protein	REFSEQ hypothetical protein	PF04883 Bacteriophage HK97-gp10, putative tail-component IPR010064 Bacteriophage HK97-gp10, putative tail-component (271 out of 530 proteins have this domain)
VOG01488	putative DNA binding protein	REFSEQ hypothetical protein	hypothetical protein (7 out of 7 RefSeq proteins have this description)
VOG01603	st.; tail-to-head joining protein (THJP)	sp|O48448|COMPL_BPSPP Tail completion protein gp17	Tail completion protein gp17 (SP in VOG)

Continued on next page

Table 15 – *Continued from previous page*

VOG	Protein Function (Godinho et al)	VOG current Functional Description	VOG new Functional Description
VOG01700	SSB	sp|O21902|SSB_BPLSK SSB protein	SSB protein (SP in VOG)
VOG01747	gp40 helicase loader	REFSEQ hypothetical protein	SSF89064, Replisome organizer (g39p helicase loader/inhibitor protein), IPR036173, G39, N-terminal domain superfamily, - (3/31 proteins have this domain); PF11417, Loader and inhibitor of phage G40P, IPR024424, Replicative helicase inhibitor G39P, N-terminal, - (1/31 proteins have this domain);
VOG02406	st.; initiation of infection; binds to portal	sp|Q38442|GP7_BPSPP Minor head protein GP7	Minor head protein GP7 (SP in VOG)
VOG02589	recT-like recombinase	sp|P03698|VBET_LAMBD Recombination protein bet	Recombination protein bet (SP in VOG)
VOG03678	procapsid scaffolding protein	sp|Q38580|SCAF_BPSPP Capsid assembly scaffolding protein	Capsid assembly scaffolding protein (Homology-based)
VOG03856	putative ATP-binding protein	REFSEQ hypothetical protein	DNA double-strand break repair helicase HerA (Homology-based)
VOG06194	large terminase subunit (TerL)	sp|Q05219|VG13_BPML5 Gene 13 protein	Terminase, large subunit (SP in VOG)
VOG06470	small terminase subunit (TerS)	sp|P68928|TERS_BPSF6 Terminase small subunit	Terminase small subunit (Homology-based)
VOG06483	SPP1 origin binding protein and replication restart (PriA-like)	sp|P03688|VRPO_LAMBD Replication protein O	DNA replication protein gp18 (SP in VOG)
VOG06647	procapsid scaffolding protein	sp|Q05222|SCAF_BPML5 Probable capsid assembly scaffolding protein	Capsid assembly scaffolding protein (SP in VOG)
VOG07809	st.; tail tube protein (TTP)	sp|A9CRB8|TAIL_BPMR1 Putative tail protein	Putative tail protein (SP in VOG)
VOG07809	st.; tail tube protein; Cter FN3 motif	sp|A9CRB8|TAIL_BPMR1 Putative tail protein	Putative tail protein (SP in VOG)
VOG07812	component of holin; cell lysis	sp|P36549|YAF2_BACLI Uncharacterized 9.7 kDa protein in cwlL 5'region	Putative antiholin (SP in VOG)

Table 16: Literature-reviewed Proteins - Current and new Annotation. Evaluation of the current and new VOG annotation by comparison with the functions of non-Swiss-Prot proteins functionally described in publications.

RefSeqID	Protein Function	VOG	VOG current Functional Description	VOG new Functional Description
466052.YP_001468054.1	Terminase, large subunit [164, 165]	VOG06194	sp|Q05219|VG13_BPML5 Gene 13 protein	Terminase, large subunit (SP in VOG)
10390.YP_001033929.1	Envelope glycoprotein L [166]	VOG00536	sp|P09308|GL_VZVD Envelope glycoprotein L	Envelope glycoprotein L (SP in VOG)
10320.NP_045368.1	Protein kinase [167]	VOG00022	sp|Q32PI1|VRK1_BOVIN Serine/threonine-protein kinase VRK1	Serine/threonine-protein kinase UL13 (SP in VOG)
10510.NP_046328.1	Pre-hexon-linking protein VIII [168]	VOG05387	sp|P03280|CAP8_ADE02 Pre-hexon-linking protein VIII	Pre-hexon-linking protein VIII (SP in VOG)
10693.YP_002854084.1	Endolysin [169]	VOG00260	sp|P16009|NEEDL_BPT4 Pre-baseplate central spike protein Gp5	Endolysin (SP in VOG)
10359.YP_081537.1	small terminase subunit [170]	VOG02938	sp|P04295|TRM3_HHV11 Tripartite terminase subunit 3	Tripartite terminase subunit 3 (SP in VOG)
215158.NP_848215.1	Major capsid protein [171, 172]	VOG01587	REFSEQ major capsid protein	SSF56563 Major capsid protein gp5 (34 out of 61 proteins have this domain); PF05065 Phage capsid family IPR024455 Phage capsid (10 out of 61 proteins have this domain); PF19307 Phage capsid-like protein IPR045641 Major membrane protein I-like, C-terminal (6 out of 61 proteins have this domain)
11246.NP_048055.1	Fusion glycoprotein F0 [173, 174]	VOG05585	sp|O36634|FUS_HRSVB Fusion glycoprotein F0	Fusion glycoprotein F0 (SP in VOG)

3.10 Marker Gene VOGs

3.10.1 Major Capsid Proteins

The 64 selected VOGs to cover MCPs covered 91.27% of all species with proteins assigned to VOGs. The number of proteins per species is 1.16678, with 3918 species represented by 1 protein, 228 species by 2 proteins, 147 species by 3 proteins, 59 species by 4 proteins and

7 species by 5 proteins (Figure 21). Out of the 417 uncovered species, 303 have dsDNA genomes. 245 of these species belong to the order Caudovirales.

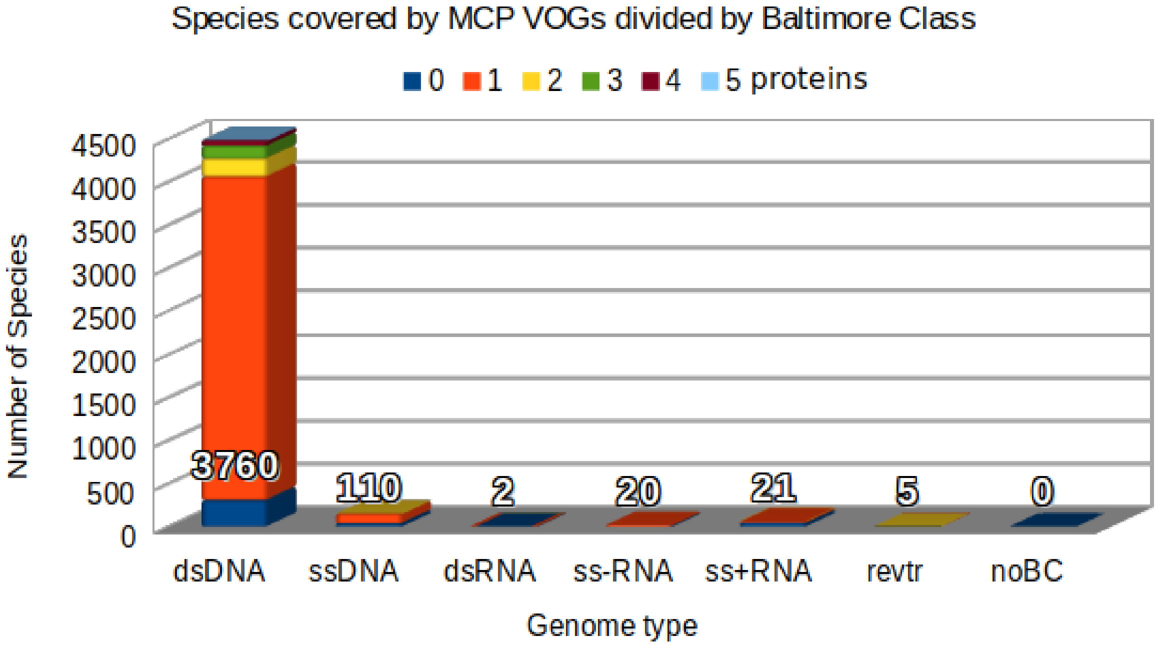


Figure 21: VOG Species covered by Major Capsid Protein VOGs. Most species are covered by a single Major Capsid Protein in the selected VOGs.

3.10.2 Terminase, large subunit

The 33 selected “Terminase, large subunit” VOGs cover 86.54% of all species with proteins assigned to VOGs. Notably, no species of non-dsDNA genome-types are covered, making these VOGs suitable markers for the identification Caudovirales. Only 201 Caudoviral species are not covered by the selected VOGs. 3547 species are covered by a single protein, while 534, 43, 4, 4 and 1 species are covered by 2, 3, 4, 5 and 6 proteins, respectively (Figure 22).

Identification of Terminase, large subunit VOGs based on the current VOG annotation file: Using the same approach to extract VOGs with Terminase, large subunit function with the current annotation file only led to a coverage of 19% of all VOG species with an average number of large Terminase subunit proteins of 1.03. With this approach only 20 VOGs were identified. With the new annotation the sensitivity of identifying Terminase, large subunit VOGs with a simple text search increased, while only minimally increasing the average number of proteins per species.

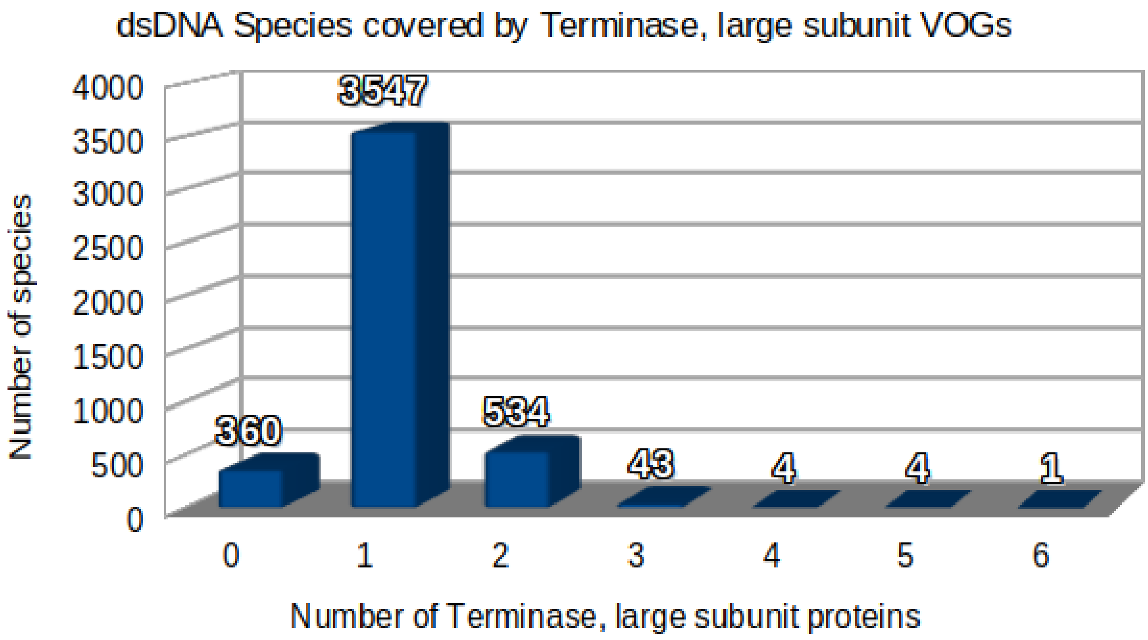


Figure 22: VOG Species covered by Terminase, large subunit VOGs. Most species are covered by a single "Terminase, large subunit" protein in the selected VOGs.

For the uncovered species it is possible that the proteins encoding a MCP are not assigned to VOGs. Even the best-represented species Caudovirales only has 92.5% of it's proteins assigned to VOGs (see coverage chart above…). In total 24% of all proteins are not placed into VOGs. Another possibility is that the VOGs the proteins are assigned to remain unannotated and are therefore not included in the MCP VOGs. A third possibility is that the major capsid protein is encoded as part of a structural polyprotein and therefore not found in the VOGDB after the application of the polyprotein filter, this is especially true for RNA viruses. While MCP are genetic markers for Caudovirales, they are also present in other virus types. The large terminase subunit is specific to Caudovirales and no other species are represented in the selected VOGs. The simple VOG selection by matching patterns in the annotations is more sensitive when using the new annotations file, as species coverage increased from 19.8% to 86.5%, while only increasing the average number of marker proteins per covered species by 0.1316. Thus the novel annotation approach enables a more user-friendly way to find VOGs belonging to gene families, as well-described proteins do not have to be manually assigned to VOGs in order to select VOGs to construct marker gene panels.

Chapter 4

Conclusion and Outlook

4.1 VOG Annotation

The new annotation provides insight into the origin of the SP-proteins used for annotation, as now it is specified if the protein has been assigned to the VOG or if it has been identified as a homolog. A major improvement can be seen in the functional description of the VOGs, as uninformative SP descriptions are preferably not used to indicate the VOG's consensus function. The added scores provide information about how many proteins share a name or domain. Annotations based on homologies to cellular organisms have been removed, as they potentially covered the target sequence only partially or could also be annotated with domains, avoiding organism-specific annotation propagation from cellular to viral proteins. The newly implemented domain-based annotation only uses a few databases of models, namely Pfam, Prosite and Superfamily. Another database, the CATH protein structure classification database, has been extended in recent years to include CATH+, which adds layers of derived data, such as predicted sequence domains, functional annotations and functional clustering (known as Functional Families or FunFams) [175]. At this point in time CATH+ has not been integrated into InterProScan, as only CATH-Gene3D models are available now. The available CATH and Superfamily domains are not well-annotated and also poorly integrated into InterPro, however, both the functional annotation and the integration are active fields of work [163]. Therefore CATH-Gene3D and Superfamily will likely provide valuable information for structure- and domain-based protein annotation in the future, and they could be integrated into future releases of the VOGDB.

4.2 Marker Gene Panels

With the current under-representation of non-dsDNA viruses in the VOGDB, the application of universal marker genes is not possible at this point in time. If diversity increases in future VOGDB releases, universal virus-specific single-copy markers could be valuable in determining contamination and completeness of viral genomes in a manner that does not depend on reference sequences.

Bibliography

1. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The Ancient Virus World and Evolution of Cells. *Biology Direct* **1,** 29. ISSN: 1745-6150 (Sept. 2006).
2. Microbiology by Numbers. *Nature Reviews Microbiology* **9,** 628–628. ISSN: 1740-1534 (Sept. 2011).
3. Sender, R., Fuchs, S. & Milo, R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. *Cell* **164,** 337–340. ISSN: 1097-4172 (Jan. 2016).
4. Pride, D. *Viruses Can Help Us as Well as Harm Us* https://www.scientificamerican.com/article/viruses can-help-us-as-well-as-harm-us/.
5. Clokie, M. R., Millard, A. D., Letarov, A. V. & Heaphy, S. Phages in Nature. *Bacteriophage* **1,** 31–45. ISSN: 2159-7073 (2011).
6. Fischetti, V. A. Development of Phage Lysins as Novel Therapeutics: A Historical Perspective. *Viruses* **10,** 310. ISSN: 1999-4915 (June 2018).
7. Verster, K. I. *et al.* Horizontal Transfer of Bacterial Cytolethal Distending Toxin B Genes to Insects. *Molecular Biology and Evolution* **36,** 2105–2110. ISSN: 0737-4038 (Oct. 2019).
8. Fiers, W. *et al.* Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene. *Nature* **260,** 500–507. ISSN: 0028-0836 (Apr. 1976).
9. Gelderblom, H. R. *Structure and Classification of Viruses* (University of Texas Medical Branch at Galveston, 1996).
10. Chaitanya, K. V. Structure and Organization of Virus Genomes. *Genome and Genomics,* 1–30 (Nov. 2019).
11. Schulz, F. *et al.* Giant Virus Diversity and Host Interactions through Global Metagenomics. *Nature* **578,** 432–436. ISSN: 1476-4687 (Feb. 2020).

12. Aherfi, S., La Scola, B., Pagnier, I., Raoult, D. & Colson, P. The Expanding Family Marseilleviridae. *Virology. Special Issue: Giant Viruses* **466–467,** 27–37. ISSN: 0042-6822 (Oct. 2014).

13. *Mimivirus | Learn Science at Scitable* https://www.nature.com/scitable/topicpage/discovery-of-the-giant-mimivirus-14402410/.

14. Oliveira, G., La Scola, B. & Abrahão, J. Giant Virus vs Amoeba: Fight for Supremacy. *Virology Journal* **16,** 126. ISSN: 1743-422X (Nov. 2019).

15. Scola, B. L. *et al.* A Giant Virus in Amoebae. *Science* **299,** 2033–2033 (Mar. 2003).

16. Boyer, M. *et al.* Giant Marseillevirus Highlights the Role of Amoebae as a Melting Pot in Emergence of Chimeric Microorganisms. *Proceedings of the National Academy of Sciences* **106,** 21848–21853. ISSN: 0027-8424, 1091-6490 (Dec. 2009).

17. Gallot-Lavallée, L. & Blanc, G. A Glimpse of Nucleo-Cytoplasmic Large DNA Virus Biodiversity through the Eukaryotic Genomics Window. *Viruses* **9,** 17. ISSN: 1999-4915 (Jan. 2017).

18. *Nucleocytoplasmic Large DNA Virus - an Overview | ScienceDirect Topics* https://www.sciencedirect.com/t
and-dentistry/nucleocytoplasmic-large-dna-virus.

19. Pagnier, I. *et al.* A Decade of Improvements in Mimiviridae and Marseilleviridae Isolation from Amoeba. *Intervirology* **56,** 354–363. ISSN: 0300-5526, 1423-0100 (2013).

20. Philippe, N. *et al.* Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes. *Science* **341,** 281–286 (July 2013).

21. Bacteriophage MS2. *Wikipedia* (May 2021).

22. Porcine Circovirus. *Wikipedia* (May 2021).

23. Colson, P. *et al.* Viruses with More Than 1,000 Genes: Mamavirus, a New Acanthamoeba Polyphaga Mimivirus Strain, and Reannotation of Mimivirus Genes. *Genome Biology and Evolution* **3,** 737–742. ISSN: 1759-6653 (June 2011).

24. Boutet, E. *et al.* UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. *Methods in Molecular Biology (Clifton, N.J.)* **1374,** 23–54. ISSN: 1940-6029 (2016).

25. Hellinger, H.-J. *Comparative Genomics of Viruses and Prokaryotes - Dissertation* 2019.

26. Verdaguer, N., Ferrero, D. & Murthy, M. R. N. Viruses and Viral Proteins. *IUCrJ* **1,** 492–504. ISSN: 2052-2525 (Oct. 2014).

27. Rodamilans, B., Shan, H., Pasin, F. & García, J. A. Plant Viral Proteases: Beyond the Role of Peptide Cutters. *Frontiers in Plant Science* **9,** 666. ISSN: 1664-462X (2018).

28. Yost, S. A. & Marcotrigiano, J. Viral Precursor Polyproteins: Keys of Regulation from Replication to Maturation. *Current Opinion in Virology* **3,** 137–142. ISSN: 1879-6257 (Apr. 2013).

29. *Phylogeny | Biology* https://www.britannica.com/science/phylogeny.

30. Gorbalenya, A. E. & Lauber, C. Phylogeny of Viruses. *Reference Module in Biomedical Sciences,* B978-0-12-801238-3.95723–4 (2017).

31. *Environmental Mutagens and Gene Expression | Learn Science at Scitable* http://www.nature.com/scitab mutagens-cell-signalling-and-dna-repair-1090.

32. Balch, W. E., Magrum, L. J., Fox, G. E., Wolfe, R. S. & Woese, C. R. An Ancient Divergence among the Bacteria. *Journal of Molecular Evolution* **9,** 305–311. ISSN: 0022-2844 (Aug. 1977).

33. Fox, G. E. *et al.* The Phylogeny of Prokaryotes. *Science (New York, N.Y.)* **209,** 457–463. ISSN: 0036-8075 (July 1980).

34. Woese, C. R. & Fox, G. E. Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* **74,** 5088–5090. ISSN: 0027-8424 (Nov. 1977).

35. Wu, D., Jospin, G. & Eisen, J. A. Systematic Identification of Gene Families for Use as "Markers" for Phylogenetic and Phylogeny-Driven Ecological Studies of Bacteria and Archaea and Their Major Subgroups. *PLoS ONE* **8,** e77033. ISSN: 1932-6203 (Oct. 2013).

36. Harris, H. M. B. & Hill, C. A Place for Viruses on the Tree of Life. *Frontiers in Microbiology* **11,** 3449. ISSN: 1664-302X (2021).

37. Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews* **84.** ISSN: 1092-2172, 1098-5557 (May 2020).

38. *Evaluation of 16S rRNA Gene Sequencing for Species and Strain-Level Microbiome Analysis | Nature Communications* https://www.nature.com/articles/s41467-019-13036-1.

39. Janda, J. M. & Abbott, S. L. 16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *Journal of Clinical Microbiology* **45,** 2761–2764. ISSN: 0095-1137 (Sept. 2007).

40. Harrison, R. G. Animal Mitochondrial DNA as a Genetic Marker in Population and Evolutionary Biology. *Trends in Ecology & Evolution* **4,** 6–11. ISSN: 0169-5347 (Jan. 1989).

41. Mercier, L. *ViralZone* https://viralzone.expasy.org/.

42. Gorbalenya, A. E. *et al.* The New Scope of Virus Taxonomy: Partitioning the Virosphere into 15 Hierarchical Ranks. *Nature Microbiology* **5,** 668–674. ISSN: 2058-5276 (May 2020).

43. Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and Evolution of Viruses of Eukaryotes: The Ultimate Modularity. *Virology* **479,** 2–25. ISSN: 0042-6822 (May 2015).

44. Kuhn, J. H. Virus Taxonomy. *Encyclopedia of Virology,* 28–37 (2021).

45. Wolf, Y. I. *et al.* Origins and Evolution of the Global RNA Virome. *mBio* **9,** e02329–18. ISSN: 2150-7511 (Nov. 2018).

46. Lefkowitz, E. J. *et al.* Virus Taxonomy: The Database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research* **46,** D708–D717. ISSN: 0305-1048 (Jan. 2018).

47. Simmonds, P. *et al.* Virus Taxonomy in the Age of Metagenomics. *Nature Reviews Microbiology* **15,** 161–168. ISSN: 1740-1534 (Mar. 2017).

48. *What's the Point of Virus Taxonomy?* July 2020.

49. Krishnamurthy, S. R. & Wang, D. Origins and challenges of viral dark matter. en. *Virus Res.* **239,** 136–142 (July 2017).

50. *Origins and Challenges of Viral Dark Matter | Elsevier Enhanced Reader* https://reader.elsevier.com/reader/
west-1&originCreation=20211118085522.

51. Emerson, J. B. *et al.* Host-Linked Soil Viral Ecology along a Permafrost Thaw Gradient. *Nature Microbiology* **3,** 870–880. ISSN: 2058-5276 (Aug. 2018).

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

53. Brum, J. R. & Sullivan, M. B. Rising to the Challenge: Accelerated Pace of Discovery Transforms Marine Virology. *Nature Reviews Microbiology* **13,** 147–159. ISSN: 1740-1534 (Mar. 2015).

54. Gregory, A. C. *et al.* Genomic Differentiation among Wild Cyanophages despite Widespread Horizontal Gene Transfer. *BMC Genomics* **17,** 930. ISSN: 1471-2164 (Dec. 2016).

55. Williamson, S. J. *et al.* Metagenomic Exploration of Viruses throughout the Indian Ocean. *PLoS ONE* **7** (ed Gilbert, J. A.) e42047. ISSN: 1932-6203 (Oct. 2012).

56. Wilson, W. H. *et al.* Genomic Exploration of Individual Giant Ocean Viruses. *The ISME Journal* **11,** 1736–1745. ISSN: 1751-7362, 1751-7370 (Aug. 2017).

57. Zhang, Y.-Y., Chen, Y., Wei, X. & Cui, J. Viromes in Marine Ecosystems Reveal Remarkable Invertebrate RNA Virus Diversity. *SCIENCE CHINA Life Sciences.* ISSN: 1674–730, 1869-1889 (2021).

58. Fawaz, M. *et al.* Duck Gut Viral Metagenome Analysis Captures Snapshot of Viral Diversity. *Gut Pathogens* **8.** Cited By :18. ISSN: 1757-4749 (2016).

59. Minot, S. *et al.* Rapid Evolution of the Human Gut Virome. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 12450–12455. ISSN: 0027-8424 (July 2013).

60. Reyes, A. *et al.* Gut DNA Viromes of Malawian Twins Discordant for Severe Acute Malnutrition. *Proceedings of the National Academy of Sciences* **112,** 11941–11946. ISSN: 0027-8424, 1091-6490 (Sept. 2015).

61. *SIV Infection-Mediated Changes in Gastrointestinal Bacterial Microbiome and Virome Are Associated with Immunodeficiency and Prevented by Vaccination | Elsevier Enhanced Reader* https://reader.elsevier.com/reader/sd/pii/S1931312816300518?token=F2010E075FF0C45FFE69B36
west-1&originCreation=20211118124721.

62. Seemann, T. Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* **30,** 2068–2069. ISSN: 1367-4803, 1460-2059 (July 2014).

63. Thibaud-Nissen, F. *et al.* P8008 The NCBI Eukaryotic Genome Annotation Pipeline. *Journal of Animal Science* **94,** 184–184. ISSN: 0021-8812. eprint: https://academic.oup.com/jas/article-pdf/94/suppl_4/184/23414490/184.pdf. https://doi.org/10.2527/jas2016.94supplement4184x (Sept. 2016).

64. Tatusova, T. *et al.* NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44,** 6614–6624 (Aug. 2016).

65. Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B. & Tatusova, T. FLAN: A Web Server for Influenza Virus Genome Annotation. *Nucleic Acids Research* **35,** W280–W284. ISSN: 0305-1048. pmid: 17545199. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933127/ (2022) (Web Server issue July 2007).

66. An Approach to Function Annotation for Proteins of Unknown Function (PUFs) in the Transcriptome of Indian Mulberry.

67. Zhang, K.-Y. *et al.* Vgas: A viral genome annotation system. en. *Front. Microbiol.* **10,** 184 (Feb. 2019).

68. Wang, S., Sundaram, J. P. & Spiro, D. VIGOR, an annotation program for small viral genomes. en. *BMC Bioinformatics* **11,** 451 (Sept. 2010).

69. Jones, P. *et al.* InterProScan 5: Genome-Scale Protein Function Classification. *Bioinformatics* **30,** 1236–1240. ISSN: 1367-4803 (May 2014).

70. Camacho, C. *et al.* BLAST+: architecture and applications. en. *BMC Bioinformatics* **10,** 421 (Dec. 2009).

71. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. en. *Nat. Methods* **12,** 59–60 (Jan. 2015).

72. Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7,** e1002195. ISSN: 1553-7358 (Oct. 2011).

73. Ruiz-Perez, C. A., Conrad, R. E. & Konstantinidis, K. T. MicrobeAnnotator: A User-Friendly, Comprehensive Functional Annotation Pipeline for Microbial Genomes. *BMC Bioinformatics* **22,** 11. ISSN: 1471-2105 (Jan. 2021).

74. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **33,** D501–D504 (2005).

75. Punta, M. *et al.* The Pfam Protein Families Database. *Nucleic Acids Research* **40,** D290–D301. ISSN: 0305-1048 (Jan. 2012).

76. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs Database of Protein Families. *Nucleic Acids Research* **31,** 371–373. ISSN: 0305-1048 (Jan. 2003).

77. *CoGe: Comparative Genomics* https://genomevolution.org/coge/.

78. Koonin, E. V. Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* **39,** 309–338. ISSN: 0066-4197, 1545-2948 (Dec. 2005).

79. Glover, N. *et al.* Advances and Applications in the Quest for Orthologs. *Molecular Biology and Evolution* **36,** 2157–2164. ISSN: 0737-4038 (Oct. 2019).

80. Nehrt, N. L., Clark, W. T., Radivojac, P. & Hahn, M. W. Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLoS computational biology* **7,** e1002073. ISSN: 1553-7358 (June 2011).

81. Stamboulian, M., Guerrero, R. F., Hahn, M. W. & Radivojac, P. The Ortholog Conjecture Revisited: The Value of Orthologs and Paralogs in Function Prediction. *Bioinformatics* **36,** i219–i226. ISSN: 1367-4803 (July 2020).

82. Galperin, M. Y. *et al.* COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens. *Nucleic Acids Research* **49,** D274–D281. ISSN: 0305-1048 (Jan. 2021).

83. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution. *Nucleic Acids Research* **28,** 33–36. ISSN: 0305-1048 (Jan. 2000).

84. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): A Resource for Comparative Genomics and Protein Family Annotation. *Nucleic Acids Research* **45,** D491–D498. ISSN: 0305-1048, 1362-4962 (Jan. 2017).

85. Altenhoff, A. M. *et al.* OMA Orthology in 2021: Website Overhaul, Conserved Isoforms, Ancestral Gene Order and More. *Nucleic Acids Research* **49,** D373–D379. ISSN: 0305-1048 (Jan. 2021).

86. Kriventseva, E. V. *et al.* OrthoDB V10: Sampling the Diversity of Animal, Plant, Fungal, Protist, Bacterial and Viral Genomes for Evolutionary and Functional Annotations of Orthologs. *Nucleic Acids Research* **47,** D807–D811. ISSN: 0305-1048 (Jan. 2019).

87. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28,** 27–30. ISSN: 0305-1048 (Jan. 2000).

88. Jensen, L. J. *et al.* eggNOG: Automated Construction and Annotation of Orthologous Groups of Genes. *Nucleic Acids Research* **36,** D250–D254. ISSN: 0305-1048 (Jan. 2008).

89. Rost, B. Twilight Zone of Protein Sequence Alignments. *Protein Engineering, Design and Selection* **12,** 85–94. ISSN: 1741-0126 (Feb. 1999).

90. *BLAST+: Architecture and Applications* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2803857/.

91. *Introduction* (National Center for Biotechnology Information (US), Jan. 2021).

92. Pearson, W. R. An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics* **42.** ISSN: 1934-3396, 1934-340X (June 2013).

93. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. en. *Nucleic Acids Res.* **25,** 3389–3402 (Sept. 1997).

94. Steinegger, M. *et al.* HH-suite3 for Fast Remote Homology Detection and Deep Protein Annotation. *BMC Bioinformatics* **20,** 473. ISSN: 1471-2105 (Sept. 2019).

95. *HMMER* http://hmmer.org/.

96. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Molecular Systems Biology* **7,** 539. ISSN: 1744-4292 (Jan. 2011).

97. Li, W., Jaroszewski, L. & Godzik, A. Clustering of Highly Homologous Sequences to Reduce the Size of Large Protein Databases. *Bioinformatics (Oxford, England)* **17,** 282–283. ISSN: 1367-4803 (Mar. 2001).

98. Letunic, I., Doerks, T. & Bork, P. SMART 7: Recent Updates to the Protein Domain Annotation Resource. *Nucleic Acids Research* **40,** D302–D305. ISSN: 0305-1048 (Jan. 2012).

99. Wu, C. H. *et al.* PIRSF: Family Classification System at the Protein Information Resource. *Nucleic Acids Research* **32,** D112–D114. ISSN: 0305-1048 (Jan. 2004).

100. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: Modeling the Evolution of Gene Function, and Other Gene Attributes, in the Context of Phylogenetic Trees. *Nucleic Acids Research* **41,** D377–D386. ISSN: 0305-1048 (Jan. 2013).

101. Pedruzzi, I. *et al.* HAMAP in 2013, New Developments in the Protein Family Classification and Annotation System. *Nucleic Acids Research* **41,** D584–D589. ISSN: 0305-1048 (Jan. 2013).

102. Sigrist, C. J. A. *et al.* New and Continuing Developments at PROSITE. *Nucleic Acids Research* **41,** D344–D347. ISSN: 0305-1048 (Jan. 2013).

103. Bru, C. *et al.* The ProDom Database of Protein Domain Families: More Emphasis on 3D. *Nucleic Acids Research* **33,** D212–D215. ISSN: 0305-1048 (Jan. 2005).

104. Attwood, T. K. *et al.* PRINTS and Its Automatic Supplement, prePRINTS. *Nucleic Acids Research* **31,** 400–402. ISSN: 0305-1048 (Jan. 2003).

105. Lees, J. *et al.* Gene3D: A Domain-Based Resource for Comparative Genomics, Functional Annotation and Protein Network Analysis. *Nucleic Acids Research* **40,** D465–D471. ISSN: 0305-1048 (Jan. 2012).

106. de Lima Morais, D. A. *et al.* SUPERFAMILY 1.75 Including a Domain-Centric Gene Ontology Method. *Nucleic Acids Research* **39,** D427–D434. ISSN: 0305-1048 (Jan. 2011).

107. *Release Notes - InterPro* https://www.ebi.ac.uk/interpro/release_notes/.

108. Hunter, S. *et al.* InterPro in 2011: New Developments in the Family and Domain Prediction Database. *Nucleic Acids Research* **40,** D306–D312. ISSN: 0305-1048 (Jan. 2012).

109. Ashburner, M. *et al.* Gene Ontology: Tool for the Unification of Biology. *Nature genetics* **25,** 25–29. ISSN: 1061-4036 (May 2000).

110. Lapidus, A. L. & Korobeynikov, A. I. Metagenomic Data Assembly – The Way of Decoding Unknown Microorganisms. *Frontiers in Microbiology* **12,** 653. ISSN: 1664-302X (2021).

111. Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I. & Watson, M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. *Frontiers in Genetics* **8,** 23. ISSN: 1664-8021 (2017).

112. Alneberg, J. *et al.* Binning Metagenomic Contigs by Coverage and Composition. *Nature Methods* **11,** 1144–1146. ISSN: 1548-7105 (Nov. 2014).

113. Nyström-Persson, J., Keeble-Gagnère, G. & Zawad, N. Compact and Evenly Distributed *k* -Mer Binning for Genomic Sequences. *Bioinformatics* **37** (ed Robinson, P.) 2563–2569. ISSN: 1367-4803, 1460-2059 (Sept. 2021).

114. Smits, S. L. *et al.* Recovering Full-Length Viral Genomes from Metagenomes. *Frontiers in Microbiology* **6,** 1069. ISSN: 1664-302X (Oct. 2015).

115. Shah, N., Molloy, E. K., Pop, M. & Warnow, T. TIPP2: Metagenomic Taxonomic Profiling Using Phylogenetic Markers. *Bioinformatics* **37,** 1839–1845. ISSN: 1367-4803 (July 2021).

116. Hunt, M. *et al.* IVA: Accurate de Novo Assembly of RNA Virus Genomes. *Bioinformatics* **31,** 2374–2376. ISSN: 1367-4803 (July 2015).

117. Yang, X. *et al.* De Novo Assembly of Highly Diverse Viral Populations. *BMC Genomics* **13,** 475. ISSN: 1471-2164 (2012).

118. Pappas, N. *et al.* Virus Bioinformatics. *Encyclopedia of Virology,* 124–132 (2021).

119. Nayfach, S. *et al.* CheckV Assesses the Quality and Completeness of Metagenome-Assembled Viral Genomes. *Nature Biotechnology.* ISSN: 1087-0156, 1546-1696 (Dec. 2020).

120. Guo, J. *Getting Started with VirSorter2* https://www.protocols.io/view/getting-started-with-virsorter2-bidpka5n. July 2020.

121. Akhter, S., Aziz, R. K. & Edwards, R. A. PhiSpy: A Novel Algorithm for Finding Prophages in Bacterial Genomes That Combines Similarity- and Composition-Based Strategies. *Nucleic Acids Research* **40,** e126. ISSN: 0305-1048 (Sept. 2012).

122. Starikova, E. V. *et al.* Phigaro: High-Throughput Prophage Sequence Annotation. *Bioinformatics (Oxford, England)* **36,** 3882–3884. ISSN: 1367-4811 (June 2020).

123. Morgan, G. J. Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959–1965. *Journal of the History of Biology* **31,** 155–178. ISSN: 1573-0387 (June 1998).

124. Schloss, P. D. & Westcott, S. L. Assessing and Improving Methods Used in Operational Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Applied and Environmental Microbiology* **77,** 3219–3226. ISSN: 1098-5336 (May 2011).

125. Di Bella, J. M., Bao, Y., Gloor, G. B., Burton, J. P. & Reid, G. High Throughput Sequencing Methods and Analysis for Microbiome Research. *Journal of Microbiological Methods* **95,** 401–414. ISSN: 0167-7012 (Dec. 2013).

126. Tyler, A. D., Smith, M. I. & Silverberg, M. S. Analyzing the Human Microbiome: A "How to" Guide for Physicians. *The American Journal of Gastroenterology* **109,** 983–993. ISSN: 1572-0241 (July 2014).

127. Bharti, R. & Grimm, D. G. Current Challenges and Best-Practice Protocols for Microbiome Analysis. *Briefings in Bioinformatics* **22,** 178–193. ISSN: 1477-4054 (Jan. 2021).

128. Saary, P., Mitchell, A. L. & Finn, R. D. Estimating the Quality of Eukaryotic Genomes Recovered from Metagenomic Analysis with EukCC. *Genome Biology* **21,** 244. ISSN: 1474-760X (Dec. 2020).

129. Gevers, D. *et al.* Re-Evaluating Prokaryotic Species. *Nature Reviews Microbiology* **3,** 733–739. ISSN: 1740-1534 (Sept. 2005).

130. Huse, S. M. *et al.* Exploring Microbial Diversity and Taxonomy Using SSU rRNA Hypervariable Tag Sequencing. *PLoS Genetics* **4,** e1000255. ISSN: 1553-7390 (Nov. 2008).

131. Klappenbach, J. A., Saxman, P. R., Cole, J. R. & Schmidt, T. M. Rrndb: The Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Research* **29,** 181–184. ISSN: 0305-1048 (Jan. 2001).

132. Větrovský, T. & Baldrian, P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE* **8** (ed Neufeld, J.) e57923. ISSN: 1932-6203 (Feb. 2013).

133. Case, R. J. *et al.* Use of 16S rRNA and rpoB Genes as Molecular Markers for Microbial Ecology Studies. *Applied and Environmental Microbiology* **73,** 278–288. ISSN: 0099-2240 (Jan. 2007).

134. *Doi:10.1016/j.Ympev.2004.01.014 | Elsevier Enhanced Reader* https://reader.elsevier.com/reader/sd/pii/S
west-1&originCreation=20211116180432.

135. Mollet, C., Drancourt, M. & Raoult, D. rpoB Sequence Analysis as a Novel Basis for Bacterial Identification. *Molecular Microbiology* **26,** 1005–1011. ISSN: 1365-2958 (1997).

136. Puhler, G. *et al.* Archaebacterial DNA-dependent RNA Polymerases Testify to the Evolution of the Eukaryotic Nuclear Genome. *Proceedings of the National Academy of Sciences* **86,** 4569–4573. ISSN: 0027-8424, 1091-6490 (June 1989).

137. Wu, M. & Eisen, J. A. A Simple, Fast, and Accurate Method of Phylogenomic Inference. *Genome Biology* **9,** R151. ISSN: 1465-6906 (2008).

138. *Phylogenomic Analysis of Bacterial and Archaeal Sequences with AMPHORA2 | Bioinformatics | Oxford Academic* https://academic.oup.com/bioinformatics/article/28/7/1033/210898.

139. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Research* **25,** 1043–1055. ISSN: 1088-9051, 1549-5469 (July 2015).

140. *Identification of New Molecular Markers for Assembling the Eukaryotic Tree of Life | Elsevier Enhanced Reader* https://reader.elsevier.com/reader/sd/pii/S1055790310001089?token=98CB8DBA
west-1&originCreation=20211117181745.

141. Yoon, H. S. *et al.* Broadly Sampled Multigene Trees of Eukaryotes. *BMC Evolutionary Biology* **8,** 14. ISSN: 1471-2148 (Jan. 2008).

142. Ren, R. *et al.* Phylogenetic Resolution of Deep Eukaryotic and Fungal Relationships Using Highly Conserved Low-Copy Nuclear Genes. *Genome Biology and Evolution* **8,** 2683–2701. ISSN: 1759-6653 (Sept. 2016).

143. Schoch, C. L. *et al.* Nuclear Ribosomal Internal Transcribed Spacer (ITS) Region as a Universal DNA Barcode Marker for Fungi. *Proceedings of the National Academy of Sciences* **109,** 6241–6246. ISSN: 0027-8424, 1091-6490 (Apr. 2012).

144. Walker, D. M., Castlebury, L. A., Rossman, A. Y. & White, J. F. New Molecular Markers for Fungal Phylogenetics: Two Genes for Species-Level Systematics in the Sordariomycetes (Ascomycota). *Molecular Phylogenetics and Evolution* **64,** 500–512. ISSN: 1095-9513 (Sept. 2012).

145. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067. ISSN: 1367-4803. eprint: https://academic.oup.com/bioinformatics/article-pdf/23/9/1061/761949/btm071.pdf. https://doi.org/10.1093/bioinformatics/btm071 (Mar. 2007).

146. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31,** 3210–3212. ISSN: 1367-4803. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/19/3210/17086320/btv351.pdf. https://doi.org/10.1093/bioinformatics/btv351 (June 2015).

147. Kristensen, D. M. *et al.* Orthologous Gene Clusters and Taxon Signature Genes for Viruses of Prokaryotes. *Journal of Bacteriology* **195,** 941–950. ISSN: 0021-9193 (Mar. 2013).

148. Frick, D. N. & Lam, A. M. I. Understanding Helicases as a Means of Virus Control. *Current pharmaceutical design* **12,** 1315–1338. ISSN: 1381-6128 (2006).

149. Rajagopal, V. & Patel, S. S. Viral Helicases. *Viral Genome Replication,* 429–466 (Nov. 2008).

150. Kristensen, D. M. *et al.* A Low-Polynomial Algorithm for Assembling Clusters of Orthologous Groups from Intergenomic Symmetric Best Matches. *Bioinformatics* **26,** 1481–1487. ISSN: 1460-2059, 1367-4803 (June 2010).

151. *UniProt* https://www.uniprot.org/.

152. *SWISS-PROT - an Overview | ScienceDirect Topics* https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/swiss-prot.

153. Bairoch, A. & Apweiler, R. The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL in 2000. *Nucleic Acids Research* **28,** 45–48. ISSN: 0305-1048 (Jan. 2000).

154. *How Do We Manually Annotate a UniProtKB Entry?* https://www.uniprot.org/help/manual_curation.

155. *UniProt Knowledgebase User Manual* https://web.expasy.org/docs/userman.html#ID_line. June 2021.

156. *Virus-Host Database* https://www.genome.jp/virushostdb/.

157. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. en. *Mol. Biol. Evol.* **33,** 1635–1638 (June 2016).

158. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PloS One* **11,** e0163962. ISSN: 1932-6203 (2016).

159. Sievers, F. & Higgins, D. G. Clustal Omega for Making Accurate Alignments of Many Protein Sequences: Clustal Omega for Many Protein Sequences. *Protein Science* **27,** 135–145. ISSN: 09618368 (Jan. 2018).

160. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. en. *Nucleic Acids Res.* **49,** D10–D17 (Jan. 2021).

161. Godinho, L. M. *et al.* The Revisited Genome of Bacillus Subtilis Bacteriophage SPP1. *Viruses* **10,** 705. ISSN: 1999-4915 (Dec. 2018).

162. Mahmoudabadi, G. & Phillips, R. A Comprehensive and Quantitative Exploration of Thousands of Viral Genomes. *eLife* **7,** e31955. ISSN: 2050-084X.

163. Blum, M. *et al.* The InterPro Protein Families and Domains Database: 20 Years On. *Nucleic Acids Research* **49,** D344–D354. ISSN: 0305-1048 (Jan. 2021).

164. Hilbert, B. J. *et al.* Structure and mechanism of the ATPase that powers viral genome packaging. en. *Proc. Natl. Acad. Sci. U. S. A.* **112,** E3792–9 (July 2015).

165. Hilbert, B. J., Hayes, J. A., Stone, N. P., Xu, R.-G. & Kelch, B. A. The large terminase DNA packaging motor grips DNA with its ATPase domain for cleavage by the flexible nuclease domain. en. *Nucleic Acids Res.* **45,** 3591–3605 (Apr. 2017).

166. Yoshida, S., Lee, L. F., Yanagida, N. & Nazerian, K. Identification and characterization of a Marek's disease virus gene homologous to glycoprotein L of herpes simplex virus. en. *Virology* **204,** 414–419 (Oct. 1994).

167. Labiuk, S. L. *et al.* Bovine herpesvirus-1 US3 protein kinase: critical residues and involvement in the phosphorylation of VP22. en. *J. Gen. Virol.* **91,** 1117–1126 (May 2010).

168. Reddy, P. S. *et al.* Nucleotide sequence, genome organization, and transcription map of bovine adenovirus type 3. en. *J. Virol.* **72,** 1394–1402 (Feb. 1998).

169. Dore, A. S. *et al.* Decoding corticotropin-releasing factor receptor type 1 crystal structures. en. *Curr. Mol. Pharmacol.* **10,** 334–344 (2017).

170. Couvreux, A. *et al.* Insight into the structure of the pUL89 C-terminal domain of the human cytomegalovirus terminase complex. en. *Proteins* **78,** 1520–1530 (May 2010).

171. Kropinski, A. M. *et al.* The genome of epsilon15, a serotype-converting, Group E1 Salmonella enterica-specific bacteriophage. en. *Virology* **369,** 234–244 (Dec. 2007).

172. Baker, M. L. *et al.* Validated near-atomic resolution structure of bacteriophage epsilon15 derived from cryo-EM and modeling. en. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 12301–12306 (July 2013).

173. Buchholz, U. J., Finke, S. & Conzelmann, K. K. Generation of bovine respiratory syncytial virus (BRSV) from cDNA: BRSV NS2 is not essential for virus replication in tissue culture, and the human RSV leader region acts as a functional BRSV genome promoter. en. *J. Virol.* **73,** 251–259 (Jan. 1999).

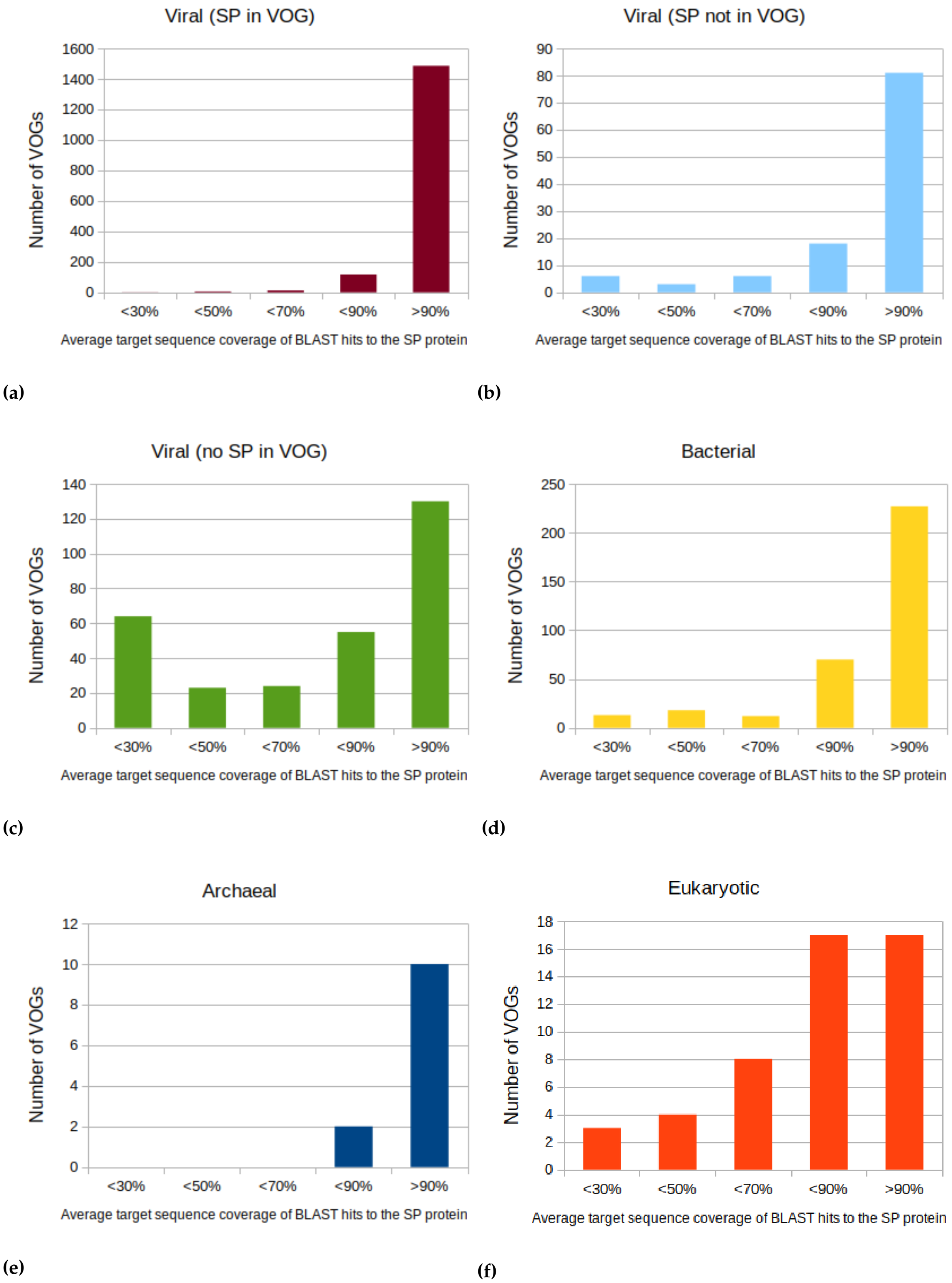
174. Zhang, B. *et al.* Protection of calves by a prefusion-stabilized bovine RSV F vaccine. en. *NPJ Vaccines* **2** (Dec. 2017).

175. Sillitoe, I. *et al.* CATH: Expanding the Horizons of Structure-Based Functional Annotations for Genome Sequences. *Nucleic Acids Research* **47,** D280–D284. ISSN: 0305-1048 (Jan. 2019).

Appendix A

AppendixA

A.1 Supplementary Figures



(a)

(b)

(c)

(d)

(e)

(f)

Supplementary Figure S1: Average target sequence coverages per Category. **a)** VOGs containing SP-proteins that are annotated with a SP-protein assigned to the VOG. **b)** VOGs containing SP-proteins, annotated with a SP protein not in the VOG. **c)** VOGs not containing SP-proteins that are annotated with a viral SP-protein. **d)** VOGs annotated with a SP-protein of eukaryotic **e)** archaeal and **f)** bacterial origin.