Computational Exploration of Virus Diversity on Transcriptomic Datasets

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Simon Käfer

aus Andernach

Bonn 2019

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

1. Gutachter: **Prof. Dr. Bernhard Misof**
Zoologisches Forschungsmuseum Koenig, Lehrstuhl Molekulare Biodiversitätsforschung, Universität Bonn

2. Gutachter: **Prof. Dr. Christian Drosten**
Institut für Virologie, Charite - Universitätsmedizin Berlin

1. Kommissionsmitglied (fachnah): **Prof. Dr. Lukas Schreiber**
Institut für Zelluläre & Molekulare Botanik, Universität Bonn

2. Kommissionsmitglied (fachfremd): **Prof. Dr. Ullrich Wüllner**
Klinik und Poliklinik für Neurologie des Universitätsklinikum Bonn

Tag der Promotion: 29.10.2019
Erscheinungsjahr: 2019

Table of Contents

1 Introduction

1.1 Viruses

Diseases caused by viruses, as well as their treatments, were known before the concept of viruses as pathogens. Applying dried scabs of smallpox onto the skin of a healthy person was used to prevent smallpox infection in the 18^{th} century. Edward Jenner used the same principle with smallpox from cows in 1796 to induce immunity to smallpox in humans. This has been the first documented case of a vaccination (from 'vacca', *latin*: cow; Modrow *et al.*, 2010).

Viruses have been identified as a potential cause for diseases in the late 19^{th} century by Louis Pasteur. After successful establishment of vaccination against Rosenbach's disease and anthrax, both caused by bacteria, he tried to find the causing agent of rabies. Since it was not possible to use dilution or ultra-filtration to eliminate the pathogenic effect of the solutions he was working with, he stated that rabies must be caused by a 'virus' (from *latin*: poison, mucus; Modrow *et al.*, 2010; Fields *et al.*, 2007). He succeeded to develop a vaccine in 1885. Later, in 1898, Dimitri I. Iwanowski and Martinus Willem Beijernick developed the concept of the 'contagium vivum fluidum', a self replicating liquid pathogenic agent. Eventually, Friedrich Loeffler and Paul Frosch discovered and verified the existence of the *Foot-and-Mouth-disease virus* in 1898 (Modrow *et al.*, 2010). Frederick Twort and Felix d'Herelle discovered that not only animals and plants but also bacteria could be infected with viruses and coined the term 'bacteriophages' in 1916/1917. Having easily cultivable bacteria as hosts and their respective phages, d'Herelle was able to establish experimental laboratory procedures like plaque essays to study virus propagation and derive infection cycles. He recognized that viruses had to enter their host cells to disseminate and that they were host-specific (Fields *et al.*, 2007). Some of his methods are still in use today.

However, the structure of viruses remained unclear as they were not visible under the light microscope. Clarification took until 1939, when d'Herelle was able to get electron micrographs of the *Tobacco mosaic virus*. The *in vitro* experiments with viruses combined with the characterization of DNA by Watson, Crick, and Franklin lead to various invaluable discoveries in molecular biology like episomes, transposons, insertion elements, retroviruses, viroids and prions (Watson and Crick, 1953; Fields *et al.*, 2007). These elements are spread by various mechanisms - including transmission by viruses - between different genomes and thus are thought to play an important role in evolution.

Yet, the origin of viruses is still unclear. There are several hypotheses that are not mutually exclusive and hence may all be correct to some extend (Wessner, 2010).

First, the progressive hypothesis. Here viruses have their origin within their host genomes. Small fragments are transferred from cell to cell due to slight mutations. Then these

fragments form groups that eventually interact with each other and are able to create virus particles and thus can be transmitted from host to host. Since retrotransposons make up an estimated ca. 42% of the human genome, these elements are potential candidates to support this hypothesis (Lander *et al.*, 2001).

Second, the regressive hypothesis, where obligate cellular parasitic organisms have lost most of their own genome that was not necessary to propagate within a host cell. Nucleocytoplasmic Large DNA Viruses (NCLDVs) are thought to be evidence for this hypothesis, especially *Mimivirus* (Raoult *et al.*, 2004). This virus is by far the largest virus that has been discovered yet. Its genome consists of a double-stranded DNA of 1.2 million basepairs (bp) that is contained in a icosahedral capsid of 400 nm in diameter. The authors describe it to be fairly similar to *Mycoplasma sp.*, small common facultative intracellular parasitic bacteria.

Third, the virus-first hypothesis. Here, the assumption is that RNA evolved before DNA. RNA carries information but can also perform catalytic functions. The first biological molecules that replicated themselves might have been viroids, *i.e.*, RNA molecules with catalyzing their own replication. Cells with membranes, inner cellular structures and cell walls evolved later. Thus viruses existed before Archaea, Bacteria, and Eukarya.

Especially in context with the endosymbiotic theory (Zimorski *et al.*, 2014), the origin of viruses and the evolution of multicellular life are possibly more intertwined than previously anticipated. Giant viruses like *Mimivirus* and other NCLDVs could have been precursors to the eukaryotic nucleus by symbiosis with a proto-eukaryote (Forterre and Gaïa, 2016). While bacteria and archaea mostly harbour larger DNA viruses, eukaryotes are more prone to be associated with small RNA viruses. Huge parts of eukaryotic organisms are comprised of retrotranscribing elements and ancient NCLDVs probably contributed a lot to the gene pool of modern eukaryotes (Goodier and Kazazian Jr, 2008; Koonin *et al.*, 2015). These integrated viral sequences were termed endogeneous viral elements (EVEs; Benveniste and Todaro, 1974; Goodier and Kazazian Jr, 2008; Holmes, 2011; Katzourakis and Gifford, 2010). However, nothing similar has been discovered in Bacteria and Archaea yet.

The large amount of detected EVEs shows that viruses play an important role in evolution, no matter which hypothesis of virus origin reflects the truth best. However, it is unknown whether viruses can still have such a large influence on human evolution today. In the modern world, virus epidemics are a global threat despite all advancements in medicine. For example, *Influenza A* has caused several documented pandemics in the 20^{th} century, starting with the 'Spanish Influenza' (H1N1) of 1918-1919 followed by the 'Asian Influenza' (H2N2) in 1957-1958, the 'Hong Kong Influenza' (H3N2) of 1968-1970 and the 'Russian Influenza' (H1N1) of 1977-1978 (Neumann *et al.*, 2009). Although vaccines are developed and adapted regularly today, highly infectious strains of *Influenza A* with pandemic potential

can emerge. Examples for this are H5N1 since 2005 (Chen *et al.*, 2006), H1N1 since 2009 (Hancock *et al.*, 2009; Neumann *et al.*, 2009), and H7N9 since 2013 (Gao *et al.*, 2013). These infections usually are spread from human to human yet especially re-assorted strains from pigs or birds are a major threat to humans. Other examples of respiratory viruses that originate from animals are the Severe Acute Respiratory Syndrome (SARS; Peiris *et al.*, 2003; Lee *et al.*, 2003) and the Middle East Respiratory Syndrome (MERS; de Wit and Munster, 2013). These viruses from the genus *Coronavirus* have emerged from their animal reservoir and cause severe illnesses in humans.

Also arthropod-borne diseases show pandemic potential associated with changes in their natural history. For instance, *West Nile virus* is usually transmitted by *Culex pipiens* from bird-to-bird but showed a shift in geographic range leading to massive amplification in non-immune bird populations, adaptation to local mosquito species, and perhaps gradual adaptation to additional vertebrate hosts including humans (Kilpatrick *et al.*, 2006). Dengue fever is considered to be a tropical disease that is transmitted by *Aedes aegypti*. Its geographic range is expected to further expand due to climate change, enabling its mosquito vector to thrive in regions that were too cold before (Hales *et al.*, 2002). Another possibility is that virus reservoirs in the Arctic or boreal areas, where low minimum temperatures have so far limited virus maintainance in insect hosts, may undergo particularly drastic changes due to the dependence of crucial mechanisms of virus-host interaction on minimal temperature thresholds (Ballinger *et al.*, 2014).

There is growing consensus that preparedness for epidemics should involve approaches to monitor viral diversity globally. Making viruses easier to detect is a first step towards that monitoring. Knowledge of broad virus diversity may subsequently enable predictions of virus spread and diversification (Jones *et al.*, 2008; Morse *et al.*, 2012; Anthony *et al.*, 2015). Additionally, if insects are already known vectors of other diseases, estimation on the pathogenicity of newly identified viruses can be made (Attoui *et al.*, 2006b). If viral evolution and diversity is ultimately shaped by environmental and ecological conditions, crucial aspects of viral emergence may become tractable by monitoring environmental change. This can lead to a whole new way of preventing, treating, and potentially eliminating virus-borne diseases (Fricke *et al.*, 2009). Emerging human epidemics could thus be identified early on (Mokili *et al.*, 2012).

Virus research has traditionally focused on human-relevant pathogens or viruses affecting livestock or agricultural products. Only recently, the exploration of viral diversity within all kinds of organisms has gained increasing attention (Mokili *et al.*, 2012). It may help in treating diseases and preventing epidemics, and may additionally indicate a way to extrapolate evolutionary processes and enable novel insight into the early evolution of life (Goodier and Kazazian Jr, 2008; Koonin *et al.*, 2015).

1.2 Insects

The evolutionary origin of Insects has been dated to about 479 million years ago (Misof *et al.*, 2014). Since then, they have successfully spread across the globe and conquered virtually all niches. Insects are the most diverse animal group on earth and can be found in nearly every habitat (Samways, 1993; Mora *et al.*, 2011). Thus they play a very important role in ecosystem health and can be used for setting the basis for many environmental impact assessment studies (Rosenberg *et al.*, 1986). Reasons for the choice of insects as ecosystem monitoring are obvious. They are predators, prey, decomposers, and pollinators that are important in every ecosystem and thus allow the comparison of different sites even across different studies.

However, in the modern western world, there recently have been multiple reports on a drastic decline in insect abundance (Leather, 2018). This change is probably caused by humans. Insects are often considered as pests that transmit diseases and harm crops. Pesticides were and are still being used to maintain the level of food production. Yet insects are also necessary for pollination (Pellmyr, 1992) and pesticides do not discriminate between beneficial and harmful insects. The decline in insects has severe impacts. Most obvious is the loss of pollinators that has a huge impact on food supply. Not only agricultural crops are at risk but also wild plants that depend on insect pollinators. Additionally, a lot of wild living animals like birds, bats and rodents feed on insects. Countermeasures have to be initiated to keep the ecosystems alive and diverse. The German Government *e.g.* has officially agreed to take part in this endeavor (Deter, 2017; Bundesregierung, 2017).

In some countries, insects are part of the daily diet. Efforts to include them into the diet of other countries have been made to counter food scarcity especially in overpopulated or inarable areas where conventional agriculture cannot provide enough food. The biggest dissent in these efforts concerns food safety and the unknown presence of potential pathogens (Halloran *et al.*, 2015). In recent years, growing evidence that insects contain large spectra of new unidentified viruses has mounted, asking for further studies (Cook *et al.*, 2013; Junglen and Drosten, 2013; Coffey *et al.*, 2014; Li *et al.*, 2015; Junglen, 2016; Shi *et al.*, 2016a,b).

As the known virus diversity is mainly derived from studies on pathogenic viruses, there is a bias towards these viruses within databases. However, viruses not necessarily cause disease. Some organisms even live in heritable symbiosis with viruses (Jaenike, 2012). For example, the parasitic wasp *Microplitis demolitor* relies on the symbiosis with *Microplitis demolitor bracovirus*. Female wasps inject the virus into other arthropods together with their eggs. The virus then allows the eggs to hatch and feed on the host by interfering with the hosts immune system so that it does not fight the eggs and larvae (Burke *et al.*, 2014; Burke, 2016).

It is obvious that using genomic and transcriptomic insect data to look for new and divergent viruses is promising and important. Especially non-blood-feeding insects probably contain vast amounts of viruses that have been neglected because they are not known to transmit diseases that are affecting human health and well-being.

1.3 Exploration of Viral Diversity

Since the initiation of the Human Genome Project (Watson, 1990), numerous large deep sequencing projects have collected enormous amounts of data, *e.g.* within Genome 10K (`genome10k.soe.ucsc.edu`), 1KITE (`www.1kite.org`; Misof *et al.*, 2014), i5k (Robinson *et al.*, 2011) and Bird 10K (Zhang, 2015). Recent advances in metagenomics with rapid growth of available gene databases have begun to facilitate the exploration of viral diversity using bioinformatic tools (Rosario and Breitbart, 2011; Mokili *et al.*, 2012; Bibby, 2013; Stephens *et al.*, 2015; Munang'andu *et al.*, 2017). Although the data of the aforementioned projects is well curated and annotated, it is expected to contain sequences of viral origin that remain undiscovered because these viruses do not yet exist in the search databases. These data can be used for a systematic analysis and exploration of viral diversity based on sensitive algorithms. Obviously, it is necessary to automate most of the process when facing vast amounts of data.

While the identification of potential viruses can be done using existing search tools (see chapter 2.1.1), the verification of these viruses is more difficult. Especially in the case of putative viral sequences that are very distantly related to known viruses, human interpretation of the results is necessary to verify the findings. Despite machine learning algorithms have improved in recent years (Dunjko and Briegel, 2018), there are still security measures like CAPTCHAs implemented in websites to tell humans apart from machines because algorithms cannot yet comprehend and solve many issues that the human brain is capable of (Jagadish *et al.*, 2014). The genome structure is an important aspect to consider when classifying a virus (Attoui *et al.*, 2006a,b) and is often too complex for algorithms to interpret. Here, the term genome structure refers firstly to the number of segments, the length of these and the combination of open reading frames (ORFs) therein, and secondly the proteins encoded by the ORFs and their relative position on the segment. The more complex such a genome structure is, the more necessary is human interpretation of those potential viral sequences.

Human interpretation of big data is time consuming and therefore a bottleneck in data analysis (Green and Guyer, 2011). It is necessary to summarize and visualize the data into a human readable and comprehensible format for faster and more reliable evaluation (Jagadish *et al.*, 2014). Creating a software that can be used by beginners and provides enhanced functionality and customizability for experienced users should be a primary goal.

This will on the one hand allow to have studies that are easier to compare and on the other hand let researchers tailor the settings to be more appropriate for their subjects.

With such software at hands, especially transcriptomic data can be used for viral studies. In contrast to genomes, transcriptomes contain only genes that are actually expressed within that organism, including viruses, and enable interpretations of the metabolic state of tissues or whole organisms (Fullwood *et al.*, 2009; Birol *et al.*, 2009). It would be impossible to find RNA viruses in genomic data as they do not have a DNA-stage. A recent example for the use of already existing transcriptomic data showed that near full virus genomes in the bivalves *Crassostrea gigas* and *Mytilus galloprovincialis* could have been identified and characterized using currently available bioinformatic tools (Rosani and Gerdol, 2017). These viruses were additionally confirmed by subsequent PCR. Transcriptomic data from the 1KITE-project has already been used for the identification of viral splicing variants (Zhou *et al.*, 2018).

The currently available virus detection pipelines are mainly designed for identification of known viruses with a view on disease-causing agents. Their general approach is to remove reads that are of host origin and then use an implementation of the Basic Local Alignment Search Tool (BLAST; Altschul *et al.*, 1990) for the remaining sequences (Wang *et al.*, 2013; Zhao *et al.*, 2013; Li *et al.*, 2016; Zhao *et al.*, 2017; Zheng *et al.*, 2017; Lin and Liao, 2017). This is a reasonable approach to reduce computing time for deep sequenced samples where the genome of the host is known. However, if the host genome is not known the search space cannot be reduced as much and using BLAST for virus search can take a very long time, especially for many large samples. Additionally, BLAST is able to detect diverse sequences only to a certain degree, so that it is only possible to identify sequences that are already in a database. Of course this is true for other algorithms as well but it is worth to think about implementation of various algorithms into pipelines that should be able to not only find diverse sequences but also agree on whether the identified sequences are of potential viral origin. Especially an implementation of Hidden Markov Models using HMMER3 (Eddy, 1998, 2011) has a promising outlook in virus research by providing a higher precision in metagenomic-based virus detection studies (Skewes-Cox *et al.*, 2014).

As viruses have very high mutation rates (Holland *et al.*, 1982), even closely related genera do not always show very high similarity and thus cannot always be easily detected via a single conventional method. Viruses that were extracted from *e.g.* cell culture and show unequivocal relatedness to known viruses on morphological fetures can sometimes be characterized and annotated using reference sequences despite very low identities (Attoui *et al.*, 2006a,b). However, morphology is not always conserved between relatively closely related viruses. Another issue is, in general, that virus taxa as distinct as genera have low sequence identity compared to prokaryoic and eucaryotic organisms. Despite this low

identity, virus characterization based on pure sequence information has been used early on (Anzola *et al.*, 1987, 1989). Additionally, laboratory studies have confirmed that the functions of strongly divergent proteins like the hemagglutinin of influenzaviruses were actually the same and that they likely share a common origin (Nakada *et al.*, 1984). It is also possible to apply proper annotation based on known protein families (Attoui *et al.*, 2001; Duncan *et al.*, 2004; Attoui *et al.*, 2005, 2006a,b, 2009). However, it is a logistical challenge to deal with masses of samples that have to go through several passages of virus isolation in the laboratory. Additionally, not all viruses can be cultivated in cell culture. Despite that, mass screening of deep sequencing data will allow to predict virus infections of the respective host and eventually improve databases for future reference.

An additional important aspect of having the ability to mass-screen metagenomic data for viruses is to study syndromes that are not obviously caused by a specific virus but rather an array of viruses in relation to bacteria and other microbiota *e.g.* in the gut microbiome of humans. There are speculations that viruses are an important driver of microbiomes (Weinbauer and Rassoulzadegan, 2004; Green and Guyer, 2011). Such influence has been reported in *Aedes albopictus*, a vector for *Chikungunya virus*, where the virus interferes with the diversity of symbiotic bacteria (Zouache *et al.*, 2012). In humans, alpha-synuclein acts as an anti-viral protein in the central nervous system. This protein has also a prion counterpart that contributes to Parkinson's disease (Massey and Beckham, 2016; Beatman *et al.*, 2016). In relation to that, patients suffering from Parkinson's disease show a significant difference in their gut microbiome compared to healthy individuals. Interestingly, virus abundance of DNA viruses was higher in healthy patients (Bedarf *et al.*, 2017). The composition of the gut microbiome is also considered in relation to multiple sclerosis. Some products of commensal and pathogenic microbiota are known to cause changes in expression of specific inflammatory proteins (Bhargava and Mowry, 2014). Imbalances of the microbial community and genetic susceptibility may eventually influence the risk and manifestation of multiple sclerosis (Brahic, 2010). However, research is just at the beginning of exploring the gut microbiome and future studies will give more insight on the issue.

1.4 Aim of this Study

The main aim of this work is to create a bioinformatic pipeline that enables mass-screening of deep sequencing data for specific and highly divergent virus groups with the focus on transcriptomic data. Sequencing the whole DNA from a eucaryotic organism using Next-Generation Sequencing (NGS) yields a so called 'genome' and contains all information stored in the DNA including non-coding regions and inactive genes (Xiong *et al.*, 2011). 'Transcriptome' refers to the corresponding sequencing of (m)RNA. Here, the (m)RNA is extracted from the organism or specific tissue, reversely transcribed into DNA and can then be sequenced using the same techniques as for DNA. This allows to identify expressed genes because inactive genes and non-coding regions are not represented in a transcriptomic dataset (no RNA-stage created within cells). This means that successfully reproducing DNA- and RNA-viruses are a part of transcriptomes as well. Genomic and transcriptomic data allow all kinds of large-scale studies on organisms (Reis-Filho, 2009) and it is necessary to make sure that only the sequences of the targeted organism are further processed in order to keep the respective study as correct as possible. However, the identification of yet unknown viral sequences enables not only the cleaning of NGS-data but also the exploration of virus diversity.

There are a few assumptions that this study is based on:

- If the RNA of an organism that is infected with a virus is sequenced, viral RNA is sequenced as well.
- If viral RNA is related to known viruses up to a certain degree, it should be detectable by different methods.
- If viral RNA is detected, not only small areas of that sequence should match known viruses, but also functional protein domains should be detectable.
- If viral RNA is supposed to be related to a known virus that is segmented, other related segments similar to that virus should be detectable as well.

The pipeline is supposed to be easy to use yet customizable to specific needs. It should be scalable and deliver a readable and comparable output. The used reference data ought to be up-to-date and use appropriate methods for the given data.

The first part (Preliminary Work, chapter 2.2, chapter 3.1) shows the proof of concept. Here, prototype search and sorting tools, possible data annotation and interpretation were tested on several RNA-virus groups on a big transcriptomic dataset.

The second part (TRAVIS, chapter 2.3, chapter 3.2) covers the pipeline algorithm and efficiency. Here, the prototype script elements have been combined and additional methods have been implemented to optimize work-flow and usefulness. Improvements have been made in terms of functionality, speed and reliability with the focus on another RNA-virus family using the same transcriptomic dataset.

2 Materials and Methods

2.1 Bioinformatic Tools Used in This Study

Apart from custom software scripts written in PERL and R, several third-party tools were used. This section contains a list of all used software including a short description.

2.1.1 Sequence Search and Comparison

In order to find similarities between two or more sequences, several algorithms have been developed for scoring and visualizing resemblances. The software described in this section covers well established methods as well as recent algorithms.

2.1.1.1 NCBIBLAST+

The Basic Local Alignment Search Tool algorithm (BLAST; Altschul *et al.*, 1990) is probably the most used algorithm for sequence comparison today. It is an essential part of the database service provided by the National Center for Biotechnology Information (NCBI; NCBICoordinators, 2016). The algorithm uses short *k*-mers ('words') to initiate the sequence comparison. *K*-mers are short snippets from a sequence, where *k* is an integer indicating the number of characters these snippets contain. The sequences that have to be compared are cut into all possible *k*-mers of *k* length (initial default size for BLAST is 5). If well-scoring matches are found, the word size is stepwise increased in order to get longer matches. BLAST assumes that the more similar two sequences are, the more *k*-mers will match along them. Several statistical values are provided for the individual matches to evaluate their significance. This allows to identify the closest known relative in a given database for a specific query sequence. This works for nucleotide and amino acid sequences and is considered to be fast and reliable (Altschul *et al.*, 1990).

2.1.1.2 HMMER3

Profile Hidden Markov Models (pHMMs) are an implementation of markov chains, where in a sequence of states the probability of the transition from one state to another is depending on the previous state. They are used for detecting remote sequence similarities on protein level where not only the identity of two sequences at a given position is considered but also the surroundings at a specific position based on the markov chain. In this study, HMMSEARCH and JACKHMMER from the HMMER3-suite are used (Eddy, 1998; Johnson *et al.*, 2010; Eddy, 2011). HMMSEARCH can use a pHMM created based on a multiple sequence alignment to look for specific matches to that profile in a protein database. It reports statistical parameters for inferring the significance of the match but is not able to identify the closest known relative from the particular alignment the profile is based on.

JACKHMMER (Johnson *et al.*, 2010) however is an implementation of a similar approach that can work with single reference sequences. Together with statistical values, it is possible to identify the closest known relative from a given database to a query sequence.

HMM-based sequence searches are implemented in several software packages and web-interfaces like Pfam (Bateman, 2004; Finn *et al.*, 2015), InterProScan (Zdobnov and Apweiler, 2001; Jones *et al.*, 2014), PROSITE (Hulo, 2006; Sigrist *et al.*, 2012) and TMHMM (Sonnhammer *et al.*, 1998).

2.1.1.3 InterProScan

InterProScan is the search tool provided for the InterPro database. It is a database containing predictive information about protein functions based on several third-party domain detection algorithms and databases (Finn *et al.*, 2016). These signatures are contributed by CATH-Gene3D (Lam *et al.*, 2015), HAMAP (Pedruzzi *et al.*, 2014), PANTHER (Mi *et al.*, 2015), Pfam (Bateman, 2004; Finn *et al.*, 2015), PIRSF (Wu *et al.*, 2004), PRINTS (Attwood *et al.*, 2012), ProDom (Bru *et al.*, 2005), PROSITE (Hulo, 2006; Sigrist *et al.*, 2012), SMART (Letunic *et al.*, 2014), SUPERFAMILY (Oates *et al.*, 2014), TIGRFAMs (Haft *et al.*, 2012), CDD (Marchler-Bauer *et al.*, 2014), and SFLD (Akiva *et al.*, 2013). InterProScan is a tool designed for searches within those signatures that relies on Hidden Markov Models using HMMER3 (Zdobnov and Apweiler, 2001; Jones *et al.*, 2014). It offers a web-interface and local installation. This tool is very powerful in prediction and annotation of proteins. However the calculations are very time consuming and the installation on a local machine requires additional software knowledge and the respective databases are very large.

2.1.1.4 MMSeqs2

MMSeqs2 is a new sequence comparison suite that is designed for large protein datasets (Steinegger and Söding, 2017). It is a *k*-mer-based approach that de-constructs reference and query sequences into 7-mers and creates temporal databases containing the positions of the individual *k*-mers and in which sequences they can be found. When comparing two sequences, the succession and position of identical words on both sequences are used to infer potential homology. The more similar two sequences are, the more sub-sequential words on both sequences match. It is possible to infer the closest known relative of the query sequence and several statistical values are given to evaluate the significance of the matches. An additional feature of MMSeqs2 is that it allows to cluster a given database by sequence similarity. This can be used to create bins of diverse sequences where the annotation is unknown, not sufficient or not applicable for the given task.

2.1.1.5 MAFFT

MAFFT is a multiple sequence alignment software that has various implemented alignment strategies (Katoh, 2002). In general, it first creates a distance matrix of the given sequences and infers a preliminary phylogenetic guide tree. Then the alignment is optimized by the guide tree progressively in multiple iterations where the guide tree is also refined multiple times. Depending on the composition of the sequences, appropriate variations can be used for the optimization of the alignment. For example, the E-INS-i algorithm is suitable for sequences that have several conserved motifs distributed over long un-alignable regions and hence is used in this study for the alignment of viral sequences. It is supposed to be the slowest but most accurate algorithm.

2.1.1.6 ASAP

ASAP (Kück, unpublished) codes amino acids based on their hydrophobicity and aligns the coded positions with MAFFT (see chapter 2.1.1.5; Katoh, 2002). The original amino acid states are then retranslated and can be used with other algorithms that require amino acid sequences. Because the three-dimensional structure of a protein is partially depending on the polar characteristics of amino acids, using this information can also be used to compare amino acids (Gaboriaud *et al.*, 1987). Especially in the case of very distantly related proteins, the three-dimensional structure might be more informative than the underlying sequence itself (Richards, 1977; Floudas *et al.*, 2006; Wright and Dyson, 1999).

2.1.1.7 T-Coffee

T-Coffee is a software suite for the generation of multiple sequence alignments (Notredame *et al.*, 2000). It follows a progressive approach and is able to combine data of different sources. These could *e.g.* be previously calculated alignments or structural protein information. Thus T-Coffee combines different algorithms into a single consistency-based alignment. The best scoring pairs of the respective sequences are used to progressively construct the overall alignment.

2.1.2 Phylogenetic Tree Reconstruction

Once related sequences are determined, alignments of homologous sequences can be used to infer phylogenies. These phylogenies help to identify *e.g.* which sequences evolved together or are ancestors of other sequences. Here, some often used algorithms for phylogenetic tree reconstruction are introduced.

2.1.2.1 Neighbor-Joining

The neighbor-joining algorithm (Saitou and Nei, 1987) is based on a distance matrix for a set of taxa. Often the required distances are calculated based on a multiple sequence alignment. Then, these distances are used to pair closest relatives and a new matrix is created that contains the combined distance of these pairs to the remaining taxa. This process is repeated until all taxa are represented in the tree. Generally, distance-based algorithms are able to calculate phylogenetic relationships very fast but do not allow retracing ancestral states at internal nodes because the sequence information is lost by calculating distances. In this study the neighbor-joining function implemented in the R-package APE has been used (Paradis *et al.*, 2004).

2.1.2.2 FastME

FastME is supposed to be an improvement over Neighbor-Joining by iteratively rearranging and improving the obtained initial tree topology (Lefort *et al.*, 2015). The distances that are used to calculate the initial tree is based on a multiple sequence alignment and various algorithms can be used to optimize these distances. Most importantly, FastME requires an evolutionary model to be specified for calculating the distances. The rearrangement of tree topology is either be done by Nearest Neigbor Interchange (NNI; Jiang *et al.*, 2000) or Subtree Pruning and Regrafting (SPR; Bordewich and Semple, 2005) and is repeated until the optimal tree based on balanced minimum evolution (BME; Desper and Gascuel, 2004) is found.

2.1.2.3 PhyML

PhyML (Guindon and Gascuel, 2003; Guindon *et al.*, 2009, 2010) is an implementation of maximum likelihood (ML) as suggested by Felsenstein (1981). It uses the maximum likelihood estimate of an evolutionary rate based on an evolutionary model to find the best fitting topology to that model. This is usually done by calculating an initial tree with on distance-based methods and then evaluating the likelihood on how well the topology fits the model. Then, parts of the tree are switched and the likelihood is estimated again. Usually these switches are based on Nearest Neigbor Interchange (NNI; Jiang *et al.*, 2000) or Subtree Pruning and Regrafting (SPR; Bordewich and Semple, 2005). If the likelihood

of the new tree is higher, this tree is used for further iterations. This process continues until the optimal tree according to ML is found. Maximum Likelihood phylogenies are thought to be the most accurate tree inference methods available today. A general assumption is that the probability for inferring the real topology increases with the amount of given data. However this is only true if the appropriate model is chosen.

2.1.2.4 SplitsTree

Phylogenetic tree inference algorithms assume a dichotomous species evolution and neglect horizontal gene transfer that is a known phenomenon in segmented viruses like influenza- and reoviruses, where recombinations of different strains occur that can lead to very contagious and pathogenic strains. A phylogenetic network is able to highlight nodes, where a clear, dichotomous topology is difficult to resolve or wrong to assume. For this reason, SplitsTree (Huson and Bryant, 2006) was used to show the conflict in the data that has been used to infer the phylogenies. It uses an alignment and creates an additional split (represented as a branch) for each position in the alignment where a dichotomous split is not congruent with the rest of the data.

2.1.3 Auxiliary Tools

Here, additional software that mostly deals with evaluation of alignments and phylogenies is described. Some help to facilitate visualization and interpretation of the obtained results by other methods.

2.1.3.1 BOOSTER

In the context of large and divergent datasets, bootstrap support for maximum likelihood phylogenies based on classic bootstrapping by Felsenstein (Felsenstein, 1981) is underestimated especially for deep branches. Booster is an implementation of 'transfer bootstraps' that corrects for these underestimations (Lemoine *et al.*, 2018).

2.1.3.2 efetch

efetch (Sayers, 2010) allows the automated retrieval of various datasets using http(s)-requests from the NCBI database. In this study, it has been extensively used for downloading sequence and taxonomy data from NCBI (NCBICoordinators, 2016) based on accession numbers from the respective databases.

2.1.3.3 Exonerate

Exonerate is a heuristic sequence comparison framework (Slater and Birney, 2005). It is part of the EMBOSS (the European Molecular Biology Open Software Suite) package that contains an extensive library of tools for dealing with molecular data (Rice *et al.*, 2000). In this study, especially FASTATRANSLATE was used to translate nucleotide data into amino acids.

2.1.3.4 FASconCAT-G

FASconCAT-G is a software package that allows different automated manipulations of multiple sequence alignments (Kück and Longo, 2014). In this study, it has been used to generate consensus sequences from given alignments.

2.1.3.5 ggtree

ggtree is an extension of the ggplot2 (Wickham, 2016) for R that allows the plotting of phylogenetic trees with various annotation methods and display modes (Yu *et al.*, 2016).

2.1.3.6 Newick Utilities

Newick Utilities are a collection of software tools for displaying and manipulating newick tree files (Junier and Zdobnov, 2010).

2.1.3.7 Pal2Nal

Pal2Nal is a software to infer a nucleotide alignment based on a given amino acid alignment (Suyama *et al.*, 2006). The original nucleotide sequence of the amino acid sequence from the alignment has to be provided to the program as well because it is not possible to retrieve the original nucleotide sequence of an amino acid due to the redundancy of the genetic code (Crick, 1968).

2.1.3.8 TrimAl

TrimAl is used for alignment masking and trimming. It has been shown that reducing columns with a very high randomization and/or gaps in alignments usually leads to better supported topologies (Capella-Gutierrez *et al.*, 2009).

2.1.3.9 tqDist

tqDist (Sand *et al.*, 2014) is used for the comparison of tree topologies based on triplets or quartets of taxa. In this study, the quartet-based comparison has been applied. The algorithm dissects a given multi taxa phylogeny into all possible quartets and compares them with all possible quartets of another multi taxa phylogeny of that contains the same taxa. This can help to identify stable topologies reconstructed *e.g.* by different alignment or phylogenetic inference algorithms.

2.2 Preliminary Work

This part is about testing the validity and applicability of the assumptions made in chapter 1.4. Transcriptomic data from 1KITE (see chapter 2.2.2) has been screened for several groups of RNA-viruses (see chapter 2.2.3). The obtained potential viral sequences were partially evaluated manually with the help of small auxiliary scripts. This procedure was necessary to identify bottlenecks and complicacies in the general methodology. Experience and knowledge gained by this process was used to improve the methods and approach as detailed in chapter 2.3.

2.2.1 Disclaimer

The material and results of the chapters 'Preliminary Work' (chapter 2.2, chapter 3.1) of this thesis have been done in very close collaboration with MSc. Sofia Paraskevopoulou and Dr. Florian Zirkel. The core of the initial prototype search script has been provided by Dipl-Biol. Malte Petersen. Results will be shown and discussed only superficially in order to show the proof of concept for the general approach. However, detailed analysis and interpretation is in preparation for publication together with MSc. Sofia Paraskevopoulou, Dr. Sandra Junglen and Prof. Dr. Christian Drosten. A manuscript titled 'Re-assessing the diversity of negative strand RNA viruses in insects' (see chapter 6.1 and the digital appendix) is already submitted and is focused on the interpretation of the findings regarding negative strand RNA viruses as displayed in Fig. 22 A and B.

2.2.2 1KITE: The 1000 Insect Transcriptome Evolution Project

The main goal of 1KITE (`http://1kite.org`) is to sample transcriptomes across all extant insect orders and families to resolve their phylogeny and answer other evolutionary questions. 1243 transcriptomes were used for this study. They were assembled and quality controlled according to Misof *et al.*, 2014. Data from this project has already been used to show that transcriptomic data can be used for virus research (Zhou *et al.*, 2018). This dataset can not onpy provide insight into insect phylogeny but also set the basis for co-evolutionary analyses with the contained viruses after they are verified and characterized. The transcriptomes consist of an average of 34609 transcripts with a mean average length of 897 nucleotides. In total, this were 42,500,986 sequences made up of 35,322,247,344 nucleotides For the sake of a better overview, we decided to summarize certain arthropod orders into groups (see Table 1); Additional information was added directly from collected sample information provided by the 1KITE Team (especially Dr. Karen Meusemann & Dr. Jeanne Wilbrandt).

Table 1: Grouped Orders.
Overview of the insect orders that have been grouped.

Group	Order
Amphiesmenoptera	Lepidoptera
	Trichoptera
Ellipura	Collembola
	Protura
Neuropterida	Megaloptera
	Neuroptera
	Raphidioptera
Polyneoptera	Blattodea
	Dermaptera
	Embioptera
	Grylloblattodea
	Isoptera
	Mantodea
	Mantophasmatodea
	Orthoptera
	Phasmatodea
	Plecoptera
	Zoraptera

2.2.3 Reference Viruses

RNA-dependent RNA-polymerase (RdRp) amino acid sequences of several groups of single stranded RNA viruses were downloaded from the NCBI database (`www.ncbi.nlm.nih.gov`, October 2014) and further used as reference viruses. These reference viruses were representatives of Arenaviridae, Bunyaviridae, Flaviviridae, Mononegavirales, *Negevirus*-like viruses, Nidovirales, Picornavirales, Orthomyxoviridae and *Togavirus*-like viruses. Taxonomical classification was based on the respective NCBI genebank entry and on the classification provided by the International Committee on the Taxonomy of Viruses (ICTV; `www.ictvonline.org`; Davison *et al.*, 2017). The sequences were sorted into the aforementioned groups and then aligned using the web-interface of T-coffee in 'expresso'-mode (`http://tcoffee.crg.cat/`; Notredame *et al.*, 2000). As the RdRp of the used reference viruses is often encoded on a polyprotein, the alignments have been manually cut to the RdRp-region. This resulted in nine different multiple sequence alignments of group-specific RdRps that were used for sequence search in the transcriptomes.

Short descriptions and typical genome organizations can be found on the following pages. If not stated otherwise, they rely on Fields *et al.*, 2007 and Davison *et al.*, 2017 . For the depiction of genome structure, the annotations are based on the respective NCBI genebank entry of the respective viruses. The term 'additional protein' is used for proteins that are either of unknown or very specific/unique function and thus not further mentioned for the sake of simplicity. Additional protein domain annotations for specific domains are derived from InterProScan. These domains are helicases, nucleases, proteases, RdRps, signal peptides, transferases, and zinc-fingers. The existance and position of those domains within the genome can give more insight about the genetic blueprint of the particular virus group.

Due to the constant efforts of the ICTV to unify virus classification and taxonomy, it is difficult to keep studies up to date with recent changes. For example, the classification of Bunyaviridae (chapter 2.2.3.2) has undergone very big changes throughout the last few years. Thus, the descriptions here mostly reflect the classification at the time of database generation in 2014. The results in this study will be based on the aforementioned classification as well.

2.2.3.1 Arenaviridae

Arenavirus, *Mammarenavirus* and *Hartmanivirus* make up the family of Arenaviridae within the single-stranded RNA negative-strand viruses. The virions are mostly spherical with a mean diameter of 110-130 nm. Their genome is bi-segmented consisting of a smaller segment (S, ca. 3.5 kb) encoding for the glycoprotein precursor (GPC) together with the nucleoprotein (NP) and a larger segment (L, ca. 7.2 kb) that contains the RdRp (see Fig. 1, Davison *et al.*, 2017). The two segments often have intra-complementary termini and thus are able to form pan-handle structures (Schlee *et al.*, 2009). These termini are conserved between the segments.

They are mostly transmitted by rodents and can cause viral hemorrhagic fever and encephalitis in humans whereas many infections happen unnoticed and are symptomatically easily mistaken as common flu-like illnesses. Well-known representatives are *Lassa virus* and *Lymphocytic choriomeningitis virus* (Fields *et al.*, 2007; Davison *et al.*, 2017).

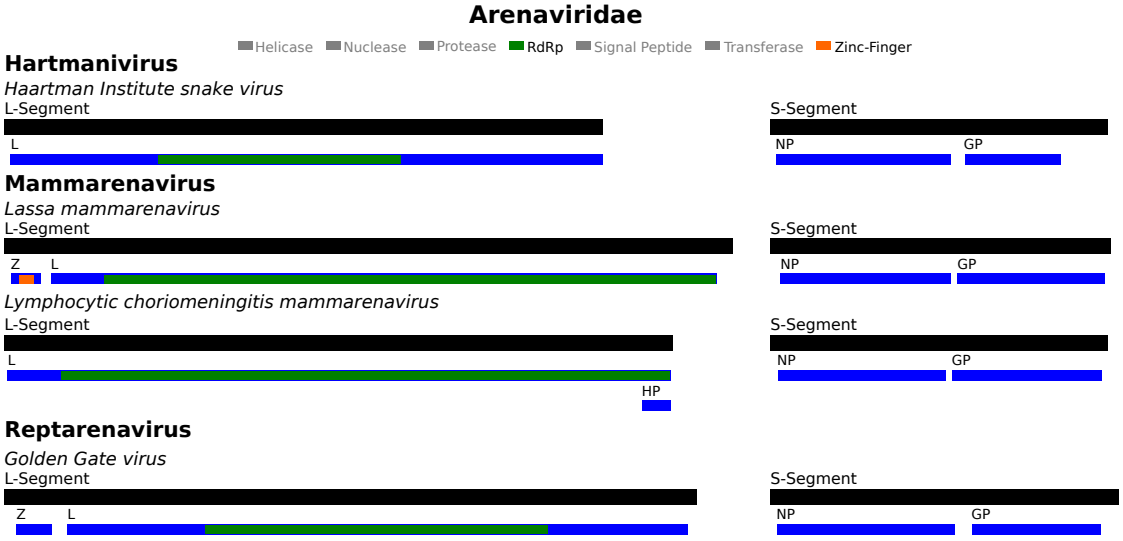

Figure 1: Genome Organization of Arenaviridae.

The genome of Arenaviridae is bi-segmented. One small (S) segment encodes the nucleo- (NP) and glycoprotein (GP) and a larger (L) segment encodes the RdRp. Additional genes are encoded by the Z and hypothetical proteins (HP).

2.2.3.2 Bunyaviridae

At the time of starting this study, the family Bunyaviridae belongs to the single-stranded RNA negative-strand viruses and consisted of five known genera: *Hantavirus*, *Nairovirus*, *Orthobunyavirus*, *Phlebovirus* and *Tospovirus*.

Recently the Bunyaviridae have been accepted as an order called Bunyavirales with the families Arenaviridae (chapter 2.2.3.1) , Cruliviridae, Feraviridae, Fimoviridae, Hantaviridae, Mypoviridae, Nairoviridae, Peribunyaviridae, Phasmaviridae, Phenuiviridae, and Wupedeviridae (Davison *et al.*, 2017) Virus particles are spherical and enveloped with a diameter of ca. 90 to 100 nm (Fields *et al.*, 2007). Their genomes are tri-segmented with a small (S) segment (S, ca. 0.9 kb to 2.9 kb) that contains the nucleoprotein (NP), a medium (M) segment (M, ca. 3.2 kb to 4.8 kb) that contains the glycoprotein (GP) and a large (L) segment (L, ca. 6.4 kb to 12.2 kb) that contains the RdRp (see Fig. 2). As for Arenaviridae, segments often have conserved intra-complementary termini (Schlee *et al.*, 2009) Many Bunyaviridae cause arthropod-borne diseases that can evoke flu-like symptoms, hemorrhagic fever or encephalitis. .

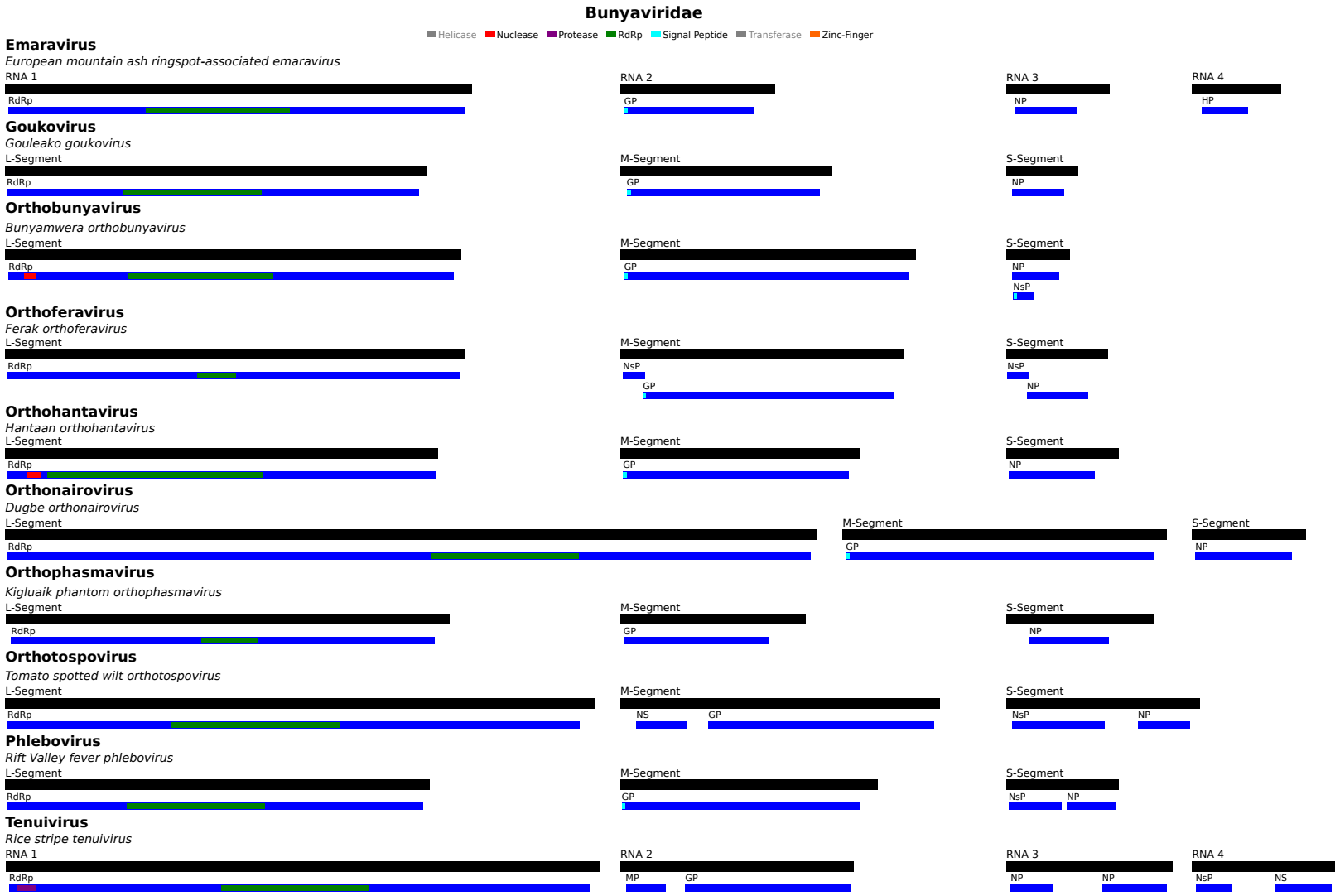

Figure 2: Genome Organization of Bunyaviridae.

Bunyaviridae have a tri-segmented genome where the S-segment encodes the nucleoprotein (NP) on one or two ORFs, the M-segment the glycoprotein (GP) and a larger L-segment that carries a polyprotein where the RdRp is located. Additional proteins for the displayed representatives are other non-structural proteins (NsP), matrix proteins (MP), and hypothetical proteins (HP).

2.2.3.3 Flaviviridae

Flaviviridae are a family containing four genera: *Flavivirus*, *Hepacivirus*, *Pegivirus* and *Pestivirus*. Virions are enveloped, with icosahedral and spherical shapes and ca. 4060 nm in diameter. They belong to single-stranded RNA positive-strand viruses and their genome is encoded on a single RNA molecule with a length of ca. 9 kb to 12 kb. This strand encodes a single Polyprotein that contains all structural proteins, membrane roteins and the RdRp (see Fig. 3). A lot of Flaviviridae are transmitted by insects, especially ticks and mosquitoes causing severe diseases like Yellow Fever, Dengue Fever, West Nile Fever, Hepatitis C and Pestivirus (Fields *et al.*, 2007; Davison *et al.*, 2017).

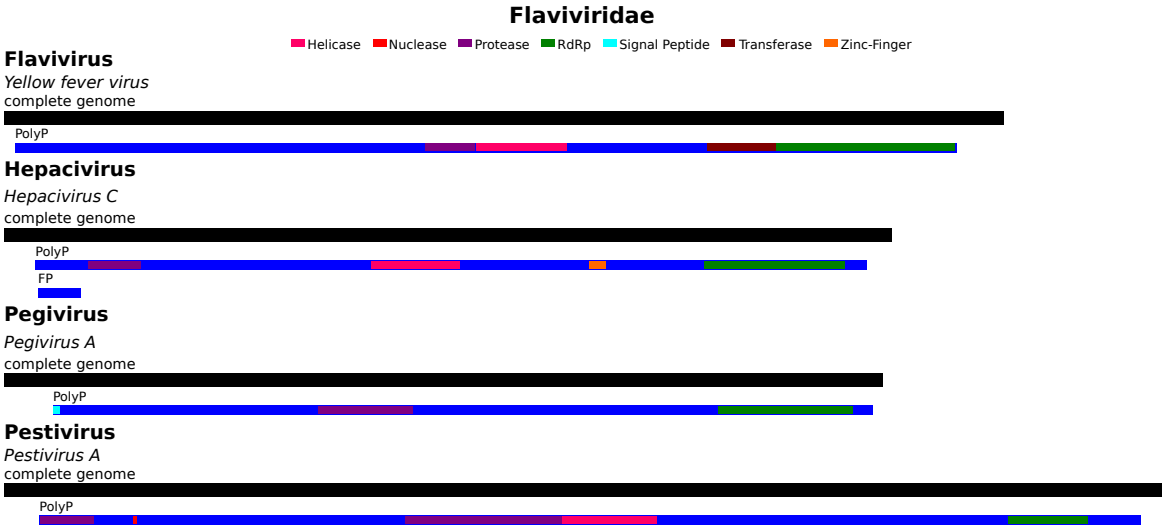

Figure 3: Genome Organization of Flaviviridae.

Flavivirus genomes consist of one large polyprotein (PolyP) that are cleaved into non-structural and structural genes. However, small accessory proteins like the F-protein (FP) of Hepatitis C are known as well and thought to be involved in morphogenesis or replication (Xu *et al.*, 2003).

2.2.3.4 Mononegavirales

Mononegavirales are an order consisting of the families Bornaviridae, Filoviridae, Mymonaviridae, Nyamiviridae, Paramyxoviridae, Pneumoviridae, Rhabdoviridae and Sunviridae. Their virion morphologies are diverse yet often are filamentous in shape with a diameter of about 50 nm. These filaments can *e.g.* form U-, 6- or circular-shaped structures. They all have a single single-stranded RNA negative-strand making up their genome. The genome sizes range from ca. 9 kb to 19 kb with multiple ORFs (mostly 5 or 6, see Fig. 4). This order contains many well known viruses with high pathogenic potentials like *Rabies virus*, *Measles virus* and *Ebola virus* (Fields *et al.*, 2007; Davison *et al.*, 2017).

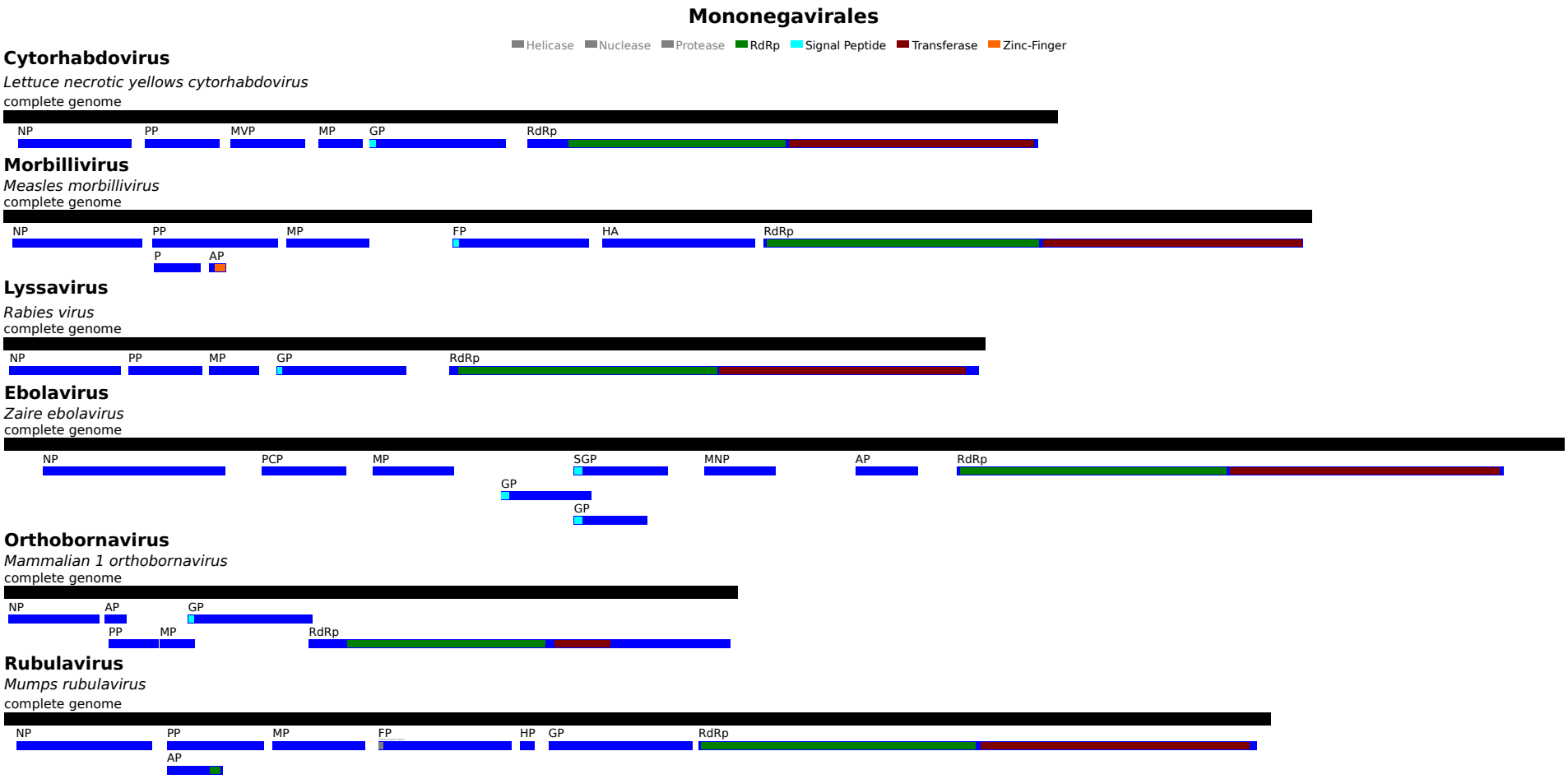

Figure 4: Genome Organization of Mononegavirales.

Mononegavirales often have a nucleoprotein (NP) at the beginning of the genome and the RdRp at the end. In between, smaller proteins like phosphoproteins (PP), glycoproteins (GP), matrix proteins (MP), movement proteins (MVP), fusion proteins (FP), spike glycoproteins (SGP), minor nucleoproteins (MNP), haemagglutinins (HA), other non-structural proteins (NsP), hypothetical proteins (HP), and other additional proteins (AP) can be found.

2.2.3.5 *Negevirus*-like viruses

Negevirus is a proposed new taxon for insect specific single-stranded RNA negative-strand viruses with a genome of about 12 kb. Their virions are spherical with diameters of ca. 50 nm. The genome encodes up to three Polyproteins (see Fig. 5). The danger for human health needs yet to be examined (Vasilakis *et al.*, 2013). In recent years, several *Negevirus*-like viruses have been discovered and mainly assigned to the genera *Nelorpivirus* and *Sandewavirus* (Nunes *et al.*, 2017). However, these genera have not yet been officially accepted by the ICTV (Fields *et al.*, 2007; Davison *et al.*, 2017).

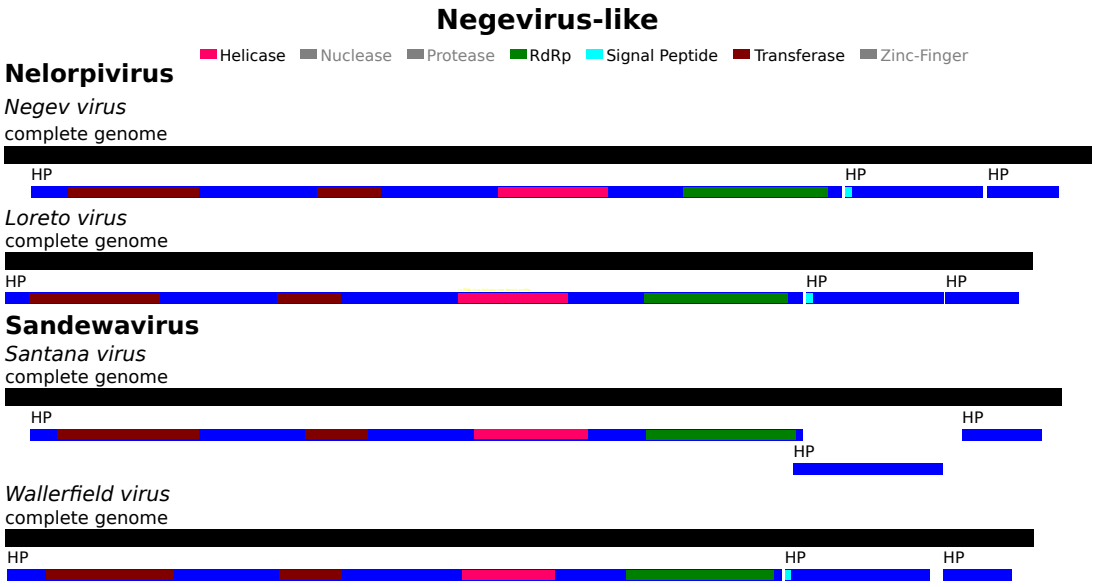

Figure 5: Genome Organization of *Negevirus*-like viruses.

Most *Negeviruses*-like viruses that have been identified so far contain a large hypothetical (poly)protein (HP) at the start of the genome which contains genes for transferases, helicases and the RdRp. This ORF followed by two other hypothetical ORFs with yet unknown functions.

2.2.3.6 Nidovirales

Nidovirales are comprised of the families Arterioviridae, Coronaviridae, Mesoniviridae and Roniviridae. They are single-stranded RNA positive-strand viruses with genome sizes of 13 kb to 31 kb consisting of multiple ORFs (6-14, see Fig. 6). The virus particles are often helical or icosahedral, have an envelope and are up to 200 nm in length. Only animal infecting viruses are known for the Arterioviridae, such as the *Equine arteritis virus* and the *Simian haemorrhagic fever virus* that often lead to the death of the animals. Most *Coronaviruses* infect mammals and birds. In humans, they usually cause harmless flu-like symptoms, however there are more dangerous species like the *Severe Acute Respiratory Syndrome virus* and the *Middle East Respiratory Syndrome virus* (Fields *et al.*, 2007; Davison *et al.*, 2017).

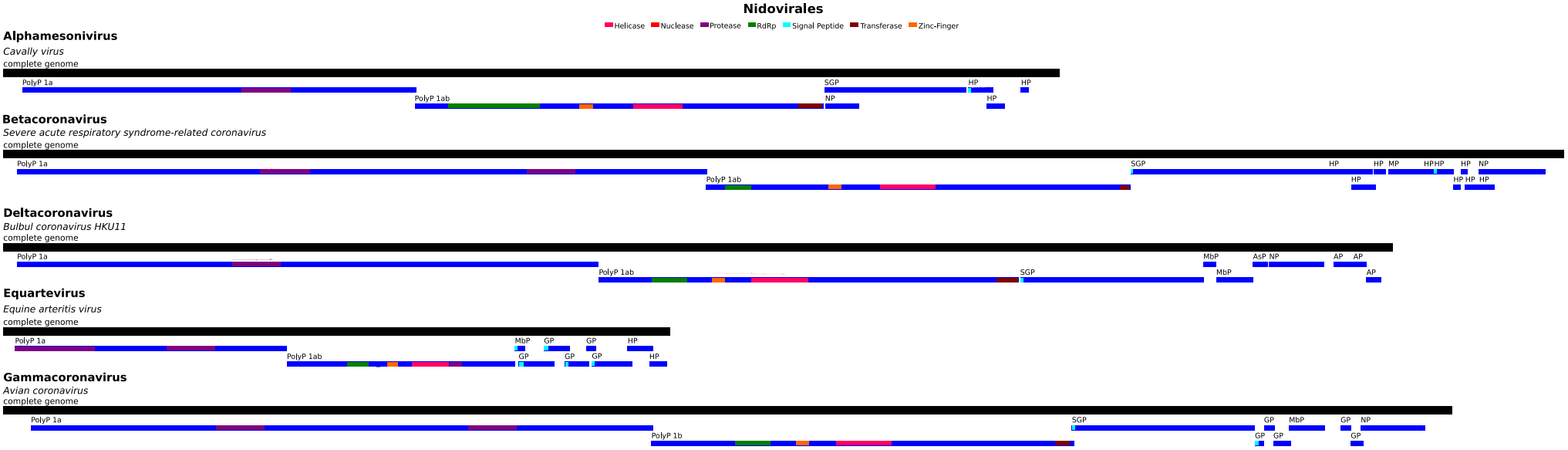

Figure 6: Genome Organization of Nidovirales.

The typical genome of Nidovirales starts with two larger ORFs that are based on a frameshift. This frameshift results in Polyprotein 1a and Polyprotein 1ab where 1ab contains the RdRp catalytic domain. This is followed by several smaller ORFs encoding specific proteins are distributed along the genome. Their order is partially conserved, often starting with the spike glycoprotein (SGP) and the nucleoprotein (NP). The remaining ORFs contain glycoproteins (GP) membrane-bound proteins (MbP), additional proteins (AP) and hypothetical proteins (HP).

2.2.3.7 Picornavirales

Picornavirales are a large order made up of the single-stranded RNA positiv-strand virus families Dicistroviridae, Iflaviridae, Marnaviridae, Picornaviridae, Polycipiviridae and Secoviridae. Their virions are of icosahedral symmetry and have a diameter of about 25 to 30 nm. The total length of the genomes vary from 2 kb to 11 kb. They have either one or two ORFs that encode polyproteins and some genera are bi-segmented (see Fig. 7). However, the RdRp is well conserved across this large order. They infect humans, animals as well as plants. Some genera are seem to be restricted to certain plant and insect species. Well known diseases caused by Picornavirales are Polio, Hepatitis A and Foot-and-mouth disease. They can also cause sicknesses like encephalitis, encephalomyocarditis, hemorrhagic fever and other flu-like symptoms (Fields *et al.*, 2007; Davison *et al.*, 2017).

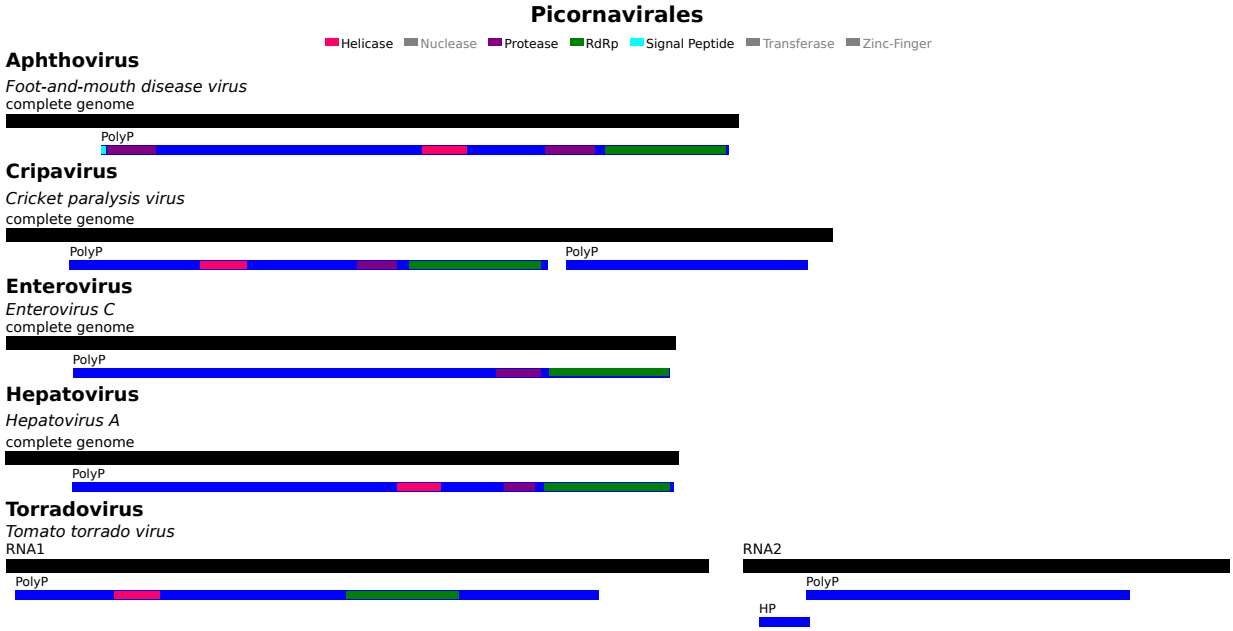

Figure 7: Genome Organization of Picornavirales.

The classical genome of Picornaviridae consists of one ORF that encodes a polyprotein (PolyP) which encodes proteases, helicases and the RdRp. However, hypothetical proteins (HP) are also predicted for some species.

2.2.3.8 Orthomyxoviridae

Orthomyxoviridae consist of the genera *Influenza A*, *Influenza B*, *Influenza C*, *Thogotovirus* and *Quaranjavirus*. Virus particles are helical and enveloped. They are single-stranded RNA negative-strand viruses with a multi-segmented genome in a range from 10 kb to 15 kb. The number of segments varies between the genera (6-8, see Fig. 8). Because of this high number of segments, there is a high chance for re-assortments by exchange of segments between multiple strains and thus to cause strains with a high threat level for human health like the *Influenza A* strain H5N1 (Zhou *et al.*, 1999; Holmes *et al.*, 2005; Dinh *et al.*, 2006; Girard *et al.*, 2010).

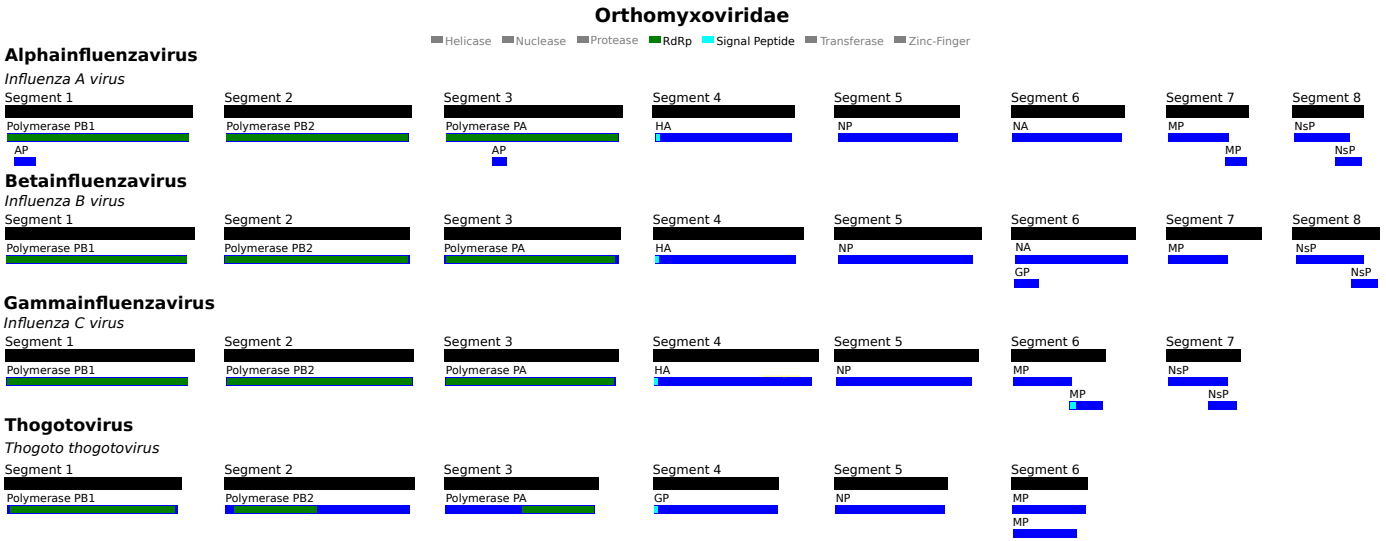

Figure 8: Genome Organization of Orthomyxoviridae.

Orthomyxoviridae have their polymerase subunits spread over three segments (PB1, PB2 and PA). Hemagglutinin (HA) and neuraminidase (NA) are on two separate segments and especially important for the classification of Influenza strains since the combination of those two enzymes determine the infection potential of the strain (Dinh *et al.*, 2006). Other segments encode glycoproteins (GP), nucleoproteins (NP), matrix proteins (MP), non-structural proteins (NsP) and additional proteins (AP).

2.2.3.9 Togaviridae

Togaviridae are a family consisting of *Alphavirus* and *Rubivirus*. Their virions are icosahedral and enveloped. They belong to the single-stranded RNA positive-strand viruses and have a genome of ca. 9.7 kb to 12 kb. Two ORFs can be found on the genome, first a polyprotein with non-structural genes followed by a polyprotein with structural genes (see Fig. 9). For humans, *Rubella virus*, *Ross River virus* and *Sindbis virus* are the most known members of Togaviridae. The latter two are arthropod borne diseases that mostly cause arthralgias and rashes. Similar symptoms are caused by the relatively recent *Chikungunya virus* that caused an epidemic on the isles around La Réunion and India in 2005-2006 (Fields *et al.*, 2007; Davison *et al.*, 2017).

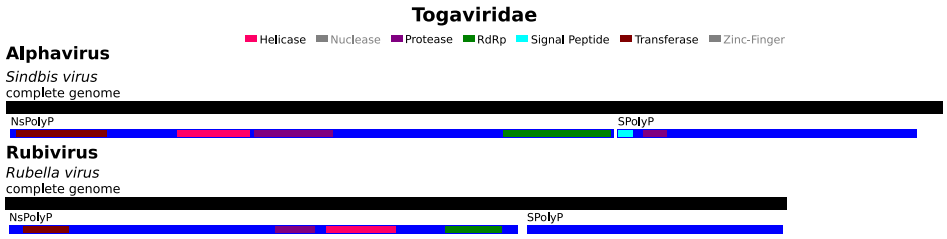

Figure 9: Genome Organization of Togaviridae.

Togaviruses usually consist of one larger ORF that encodes a polyprotein containing non-structural (NsPolyP) proteins. This ORF ist followed by a smaller ORF that contains a structural (SPolyP) polyprotein.

2.2.4 Sequence Search and Phylogenetic Tree Reconstruction

Profile hidden markov models have been created for each sequence alignment of the specific virus groups using HMMBUILD (HMMER3 v. 3.1b2). Sequence search has been automated using a custom PERL script that acted as a wrapper for EXONERATE (v. 2.2.0) and HMMER3 (v. 3.1b2). The script first translated contigs of the transcriptomes from nucleotides to amino acid for all six reading frames using FASTATRANSLATE (EXONERATE). Then all translated contigs have been searched for matches to the previously built pHMMs using HMMSEARCH with an e-value threshold of 10^{-4}. Results were summarized and checked for redundancy into a single table and basic statistics have been derived from that table using R (v. 3.2.0). The matching sequences were then checked (BLASTP; BLAST+ v. 2.2.28; e-value threshold 10^{-5}) against a virus database based on the non-redundant protein database from NCBI (05.05.2015) to identify false posotives and find the best matching reference. Full taxonomy entries were retrieved from NCBI via EFETCH based on the accession number of the match using a custom PERL script. These matches and the best matching references were then aligned to the respective original template alignments by MAFFT (v. 7.123; E-INS-i algorithm; –add option). TrimAl (v. 1.2) has been used to remove columns with a high gap content and sequences with a low sequence/resolution overlap. Phylogenetic tree reconstruction was done using PhyML (v. 3.1). 1000 bootstrap replicates under the Blosum62 substitution model were constructed, while the proportion of invariant sites and the alpha parameter of the gamma distribution were estimated.

2.2.5 Genome Organization

The open reading frames for each original virus-matching nucleotide sequence from the transcriptomes were extracted using a custom PERL script. Since some sequences were likely sequenced only fragmentarily, any sense (*i.e.* non-stop) codon was regarded as a potential start-codon. However, if a Met-codon was present it was regarded as the real start codon of that ORF. All ORFs of a length above 200 amino acids were then again compared with the virus protein database (BLASTP; e-value threshold 10^{-5}; see chapter 2.2.4) to further identify and characterize the obtained potential viral sequences in order to gain knowledge of their genome structure and thus verify the validity of the initial search results. The genome organizations have been summarized and visualized by a custom R script to enable comparison with the genome structure of known viruses. Sequences of ORFs that yielded a BLAST match or were thought to show functionality based on other references were further analyzed with InterProScan to identify protein domains and derive more functionality. Results were summarized manually for some exemplary genome organization visualizations.

2.3 TRAVIS

This part is about implementing improvements on the methodology described in chapter 2.2 and the automation of the whole process. In addition to the transcriptomes from 1KITE (see chapter 2.2.2), simulations have been made to evaluate the efficiency of the pipeline. Instead of only focusing on the RdRp-coding segments of single stranded RNA viruses as in chapter 2.2, all segments of members from the family Reoviridae (see chapter 2.3.1) have been chosen as target viruses.

2.3.1 Reoviridae

Reoviridae are a family of double-stranded RNA viruses with icosahedral virus particles of about 60 to 85 nm in diameter and have no envelope. Their known representatives infect nearly all possible host organisms including vertebrates, arthropodes, plants and fungi (Baker *et al.*, 1999; Fields *et al.*, 2007; Davison *et al.*, 2017). The family of Reoviridae currently comprises two subfamilies. The subfamily Sedoreovirinae consists of six genera: *Cardoreovirus*, *Mimoreovirus*, *Orbivirus*, *Phytoreovirus*, *Rotavirus*, and *Seadornavirus*. The subfamily Spinareovirinae consists of nine genera: *Aquareovirus*, *Coltivirus*, *Cypovirus*, *Dinovernavirus*, *Fijivirus*, *Idnoreovirus*, *Mycoreovirus*, *Orthoreovirus*, and *Oryzavirus* (Davison *et al.*, 2017).

While the RNA sequences can be very divergent between two members of the family, the genome organizations of Reoviridae are mostly conserved (Bányai *et al.*, 2014). They are often comprised of 10-12 short monocistronic segments (see Fig. 10). Monocistronic means that there is only one large open reading frame containing a single gene that usually is spanning nearly the whole segment (Fields *et al.*, 2007; Davison *et al.*, 2017).

Severe illnesses to domestic animals caused by *e.g.* the *Bluetongue virus* or *Equine Encephalitis virus* are known to be transmitted by *Culicoides sp.* and thus are classified as arboviruses (Attoui *et al.*, 2009). New members of the family Reoviridae are regularly found in various organisms and characterization is now often based on sequence similarity searches (Attoui *et al.*, 2001; Duncan *et al.*, 2004; Attoui *et al.*, 2005, 2006b,a; Moriyasu *et al.*, 2007; Anthony *et al.*, 2009; Attoui *et al.*, 2009; Belaganahalli *et al.*, 2012; Silva *et al.*, 2013; Belaganahalli *et al.*, 2013, 2014; Shen *et al.*, 2015; Rosani and Gerdol, 2017; Taniguchi *et al.*, 2017). Since their genome is segmented, it is possible to interchange segments from one virus to the other if there is a co-infection (Calisher and Mertens, 1998; Small *et al.*, 2007; Fields *et al.*, 2007; Bányai *et al.*, 2011).

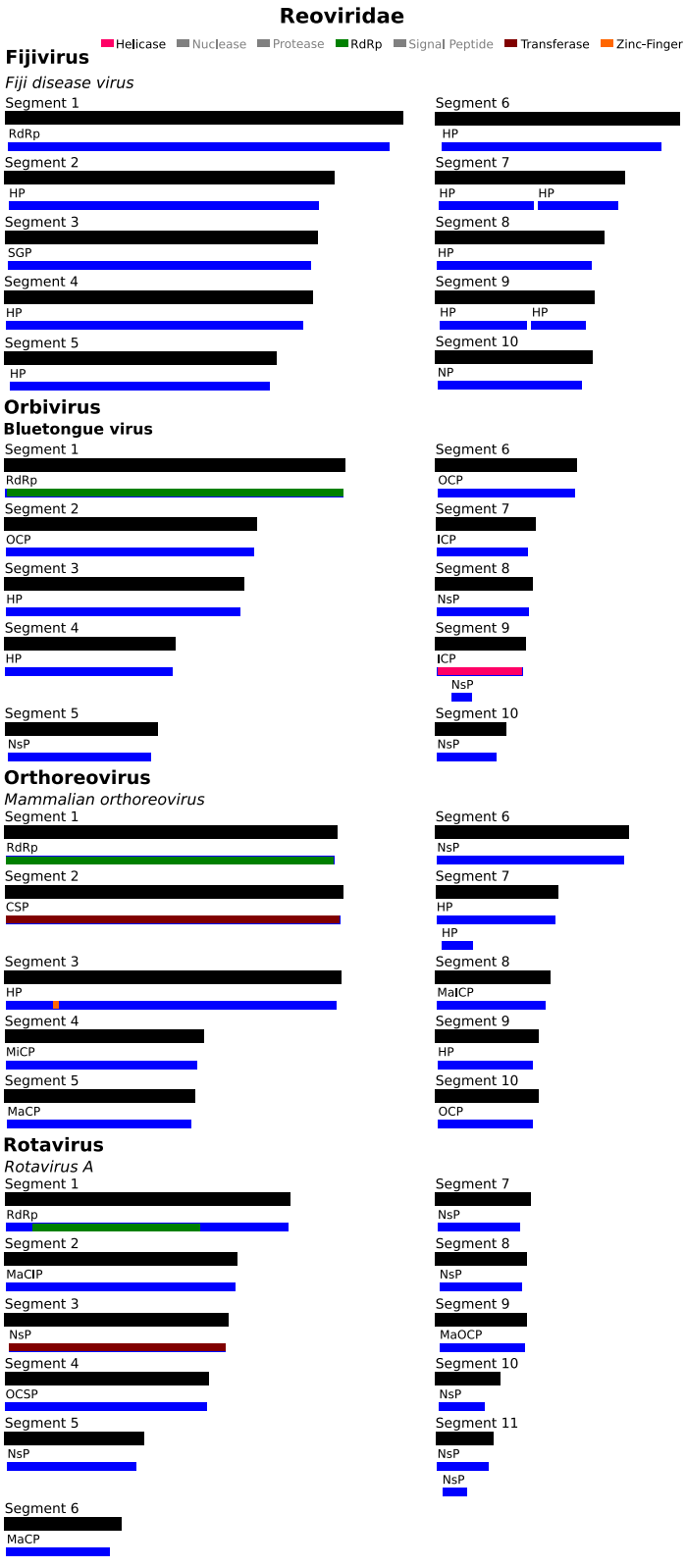

Figure 10: Genome Organization of Reoviridae.

The genome is distributed over 10 to 11 segments that usually encode a single protein. However some segments carry multiple proteins. Among the expressed proteins there are nucleoproteins (NP), spike glycoproteins (SGP), outer capsid spike proteins (OCSP), outer capsid proteins (OCP), major outer capsid proteins (MaOCP),inner capsid proteins (ICP), core spike proteins (CSP), minor inner capsid proteins (MiCP), major inner capsid proteins (MaICP) and non-structural proteins (NsP).

2.3.2 TRAVIS Pipeline Structure

Based on the assumptions made in chapter 1.4, an improved concept for the pipeline is proposed in this chapter. Afterwards the implementation of this concept is described.

2.3.2.1 Theoretical Concept

Viruses rely on their hosts replication logistics for proliferation (Modrow *et al.*, 2010; Fields *et al.*, 2007). Despite the host range is very diverse, it can be assumed that the production of viral proteins is based on the standard genetic code (Koonin and Novozhilov, 2009). To compensate for the high mutation rate of viruses, the genes of viruses are often compared at amino acid level. The amino acid sequences are more likely to be conserved due to the redundancy of the genetic code (Crick, 1968). In case of the mutation of a single nucleotide, the probability that the coded amino acid changes is reduced especially if it is on the third position of a codon (Koonin and Novozhilov, 2009). This leads to a higher chance of detecting sequence similarities for viruses that are already very divergent. Especially the highly conserved domains with specific functions should be easier to identify. Thus the whole sequence search and comparison is conducted at amino acid level. All annotated proteins of the targeted viruses make up the reference library.

This principle can also be applied to the sample library. If each single sequence from the whole sample library had to be individually compared with each single sequence from the reference library, it would take more time than using a pre-filtered sample library. If a certain virus group is targeted, the length of the proteins can be estimated and the sample sequences can be filtered for ORFs of a certain length. Thus the search space is reduced substantially which is important for large sequences that would otherwise suffer from very long calculation times (Altschul *et al.*, 1990).

Another improvement in speed can be done by looking for one or several marker genes based on a small database before initiation of the calculation for the complete reference library. The RdRp is a suitable gene for that because it is more or less conserved among the genera and is necessary for all RNA viruses (Modrow *et al.*, 2010; Fields *et al.*, 2007). Most importantly, as an RdRp is not part of any known genome of prokaryotes or eukaryotes, it is a unique marker gene for RNA viruses. Because of its necessary function in virus proliferation, it most likely has to be expressed within a host. Only if an RdRp-like structure has been found in a sample, similarities to other genes in the reference library were searched for in the respective samples.

Until now, TRAVIS supports using four different search algorithms: BLASTP, HMMESARCH, JACKHMMER and MMSEQS to use their specific strengths and balance their respective weaknesses. BLASTP (BLAST+; see chapter 2.1.1) is implemented because it is

supposed to be fast, reliable and can state similarities between two distinct sequences. The weakness of BLAST in general is that it only evaluates the similarity for each position of an alignment of two sequences independently of the surrounding positions. In contrast to that, HMMSEARCH (HMMER3; see chapter 2.1.1), consideres the probability of a certain character state that follows the probabilities of the preceding characters based on pHMMs. This, theoretically, allows higher sensitivity and thus should be able to detect more distant similarities at the expense of higher calculation time. Yet there are two major drawbacks. First, to create a pHMM, a multiple sequence alignment of proteins is needed. Therefore it is not feasible to search with a single sequence as reference. In order to be able to use HMMSEARCH for the non-marker gene sequences, these sequences have to be sorted and aligned properly. For this, MMSEQS CLUSTER (MMSeqs2; see chapter 2.1.1) is being used for clustering the non-marker genes into diverse clusters. In the case of Reoviridae this is especially helpful because the annotation of segments and proteins is very inconsistent. Thus a sequence-based grouping delivers a more consistent result than relying on the annotation. Second, a match to a pHMM does not indicate a certain similarity of two distinct sequences. So the closest relative to the match within the alignment, the pHMM is based on, cannot directly be implied. To balance these drawbacks, there is another algorithm called JACKHMMER available in HMMER3. It is an implementation of a similar algorithm as in HMMSEARCH which allows for direct comparisons of two single sequences. MMSEQS SEARCH (MMSeqs2) is a new algorithm that is designed for comparison of very large protein databases. It is a k-mer based approach that takes into account the position and the succession of the k-mers when compairing sequences. MMSEQS SEARCH is also supposed to be fast and reliable and designed to handle large datasets.

The matches in the sample library based on the reference library are considered to be 'suspicious sequences'. This means that they share properties with the sequences in the reference library but are not *per se* classified as viral. However, the number of suspicious sequences should be low enough that reciprocal search with them against the non-redundant protein database (NR; NCBI) is viable in a reasonable amount of time. For that, all ORFs based on a suspicious sequence are compared with BLASTP versus the NR. This step can then find additional matches among all publicly available annotated sequences. In the case of *e.g.* a false positive, it can be expected that the reciprocal BLAST can find a higher scoring match for the suspicious sequence (see chapter 2.3.4).

In the last step, all potential relations of all suspicious sequences to known references are analyzed by a one versus one sequence comparison. The whole sequence structure of the suspicious sequences are plotted and annotated using their corresponding references. For each suspicious ORF that matched a reference, a color scheme is applied to the respective reference based on a direct BLASTP-comparison. Thus the matching areas are visualized

and can then be evaluated by their color patterns. This is considered as an improvement over the plain numerical statistics because it can be evaluated easier and faster by humans. Tables and fasta-formatted files are delivered as well for further analysis.

2.3.2.2 Implementation

The main purpose for TRAVIS is to scan samples directed towards a certain virus group. TRAVIS will read user provided libraries and configuration files in comma separated value format (CSV). It is possible to add all valuable information that can be important for downstream analyses tailored to the project into the libraries. This information could be taxonomy, host-ranges, symptoms etc. The specified reference sequences will be downloaded from NCBI and then systematic searches at amino acid level will be performed. The output will contain multiple text files, tables and visualizations.

The user has to specify a 'main' gene that all of the viruses in the library have in common and is more or less conserved. This gene could be *e.g.* a polymerase or a capsid gene. All other genes will be regarded as 'company' genes. By default, TRAVIS will search first for the 'main' genes first and only starts searching for 'company' genes in samples that are positive for the 'main' gene. It acts as a marker gene and running 'company' searches only on 'main' positive samples will reduce the whole search time.

TRAVIS is separated into three parts that are executed subsequently: TRAVIS Henchman, TRAVIS Core, and TRAVIS Scavenger (see Fig. 11). For more details on usage and configuration, see the documentation in chapter 6.2.

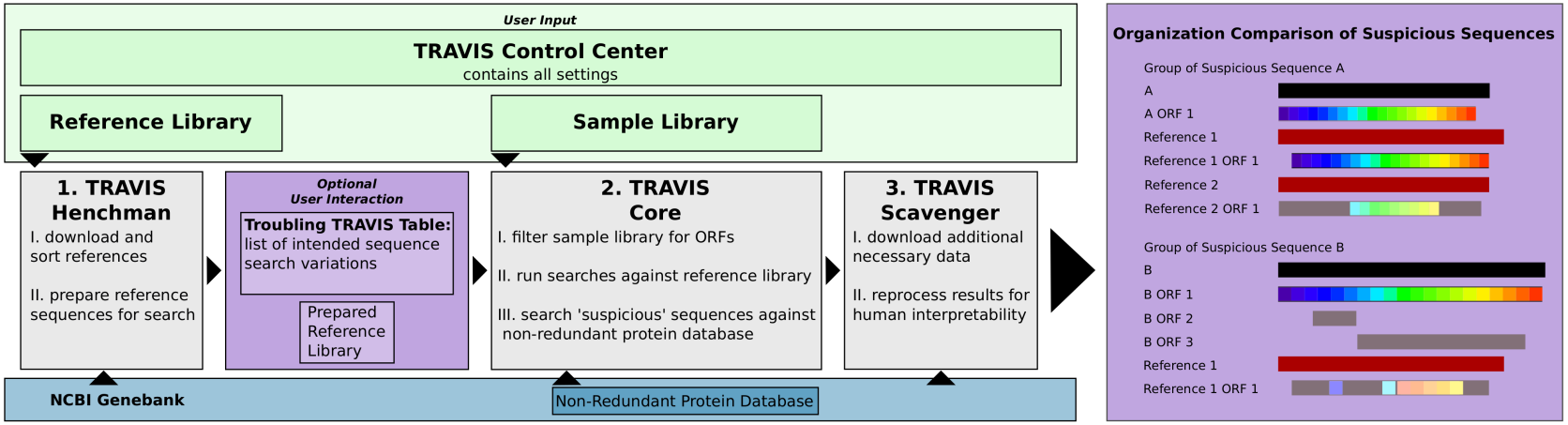

Figure 11: General Pipeline Structure.

TRAVIS consists of three parts that are executed subsequently. The user has to provide three input files (green boxes). TRAVIS is able to retrieve data from the NCBI Genebank (blue boxes). The main output files (violet boxes) are designed for human interpretation and further automated downstream analyses alike.

2.3.2.2.1 1. TRAVIS Henchman

Here, the sample library is parsed for all information and all specified sequences are downloaded from NCBI. First the 'main' sequences are sorted according to their assignment in the reference library and subsequently aligned by MAFFT. The 'main' sequences can additionally be split into groups such as taxonomic levels based on columns in the reference library table. Then all the company sequences are clustered by MMSeqs2 and annotated by unique sequence definitions provided by NCBI. All clusters are also aligned by MAFFT and are further treated as 'company clusters'. Sequences that could not be included into a cluster will be treated as 'company unclustered' and will be collected in a separate file (see Fig. 12). The main output of this part is the 'Troubling TRAVIS Table' (TTT), that contains a list of different search variations which are suggested. This is an opportunity for the user to optionally interact before the main calculations start. This interaction could be switching searches on/off or manually checking the automatically created alignments. The quality of the alignments determines the quality of the pHMMs, that HMMER3 will generate and use for sequence comparison.

```
1: procedure TRAVIS HENCHMAN(ReferenceLibrary)
2:     Initialize MainDatabase                    ▷ Will store references that are tagged as 'main' gene
3:     Initialize CompanyDatabase                 ▷ Will store references that are tagged as 'company' gene
4:     for Reference in ReferenceLibrary do       ▷ Download and sort references
5:         Download Reference from NCBI           ▷ Including annotations
6:         if ReferenceTag == 'main' then
7:             Add Reference to MainDatabase
8:         else
9:             Add Reference to CompanyDatabase
10:        end if
11:    end for
12:    for Group in SplitReferences do    ▷ Sort 'main' genes according to specified groups (family etc.)
13:        CreateFasta Group
14:        AlignFasta Group
15:    end for
16:    Cluster CompanyDatabase into ClusteredCompanyDatabase
17:    for Cluster in ClusteredCompanyDatabase do
18:        CreateFasta Cluster
19:        AlignFasta Cluster
20:    end for
21:    CreateFasta UnclusteredCompanyDatabase     ▷ Stores all unique 'company' genes
22:    Create TroublingTRAVISTable
23: end procedure
```

Figure 12: TRAVIS Henchman Algorithm.

2.3.2.2.2 2. TRAVIS Core

Here, the sample files are prepared for the searches. The ORFs are extracted from the samples and then filtered by a user specified length. This is a crucial filtering step for a directed virus search. Since the maximum genome size for the viruses of interest can be estimated, it is possible to reduce the search space by only selecting ORFs of certain lengths. All samples can be searched for all references using the supported search tools. The supported search tools up to now are BLASTP (BLAST+), HMMSEARCH (HMMER3), JACKHMMER (HMMER3) and MMSEQS (MMSeqs2). It is important to note that HMMSEARCH can only be used based on alignments whereas JACKHMMER is an implementation of similar algorithms that can use single sequences for a search. Thus HMMSEARCH cannot be run on sequences that have no other relative within the reference library. First, the search for 'main' genes is done via each specified search tool. Then the 'company' genes will be used for searching the samples that are 'main positive'. However, the software allows the user to bypass the 'main positive' setting and search for the 'company' genes also in samples where the 'main' gene has not been found.

This yields a list of 'suspicious' sequences of potential viral origin within the samples. After tracing back the original nucleotide sequence of the 'suspicious' sequences, all ORFs belonging to these sequences are compared to the non-redundant protein database (NR) from NCBI via BLASTP to identify additional related sequences that were not in the user provided virus database. This can help to identify false positives more clearly. It is also possible to exchange the non-redundant protein database by provide another customizable BLAST-database.

```
1: procedure TRAVIS CORE(SampleLibrary, TroublingTRAVISTable)
2:     Read TroublingTRAVISTable
3:     for Sample in SampleLibrary do                          ▷ Process each sample completely
4:         Initialize SuspiciousSequences          ▷ Will store sample sequences matching references
5:         ExtractORFs Sample into SampleORFs                  ▷ Within the given parameters
6:         Initialize GroupedReferences      ▷ Will store all References for non-hmmsearch searches
7:         for Reference in MainGenes do                       ▷ Based on TroublingTRAVISTable
8:             for SearchTool in Reference do
9:                 if SearchTool == 'hmmsearch' then    ▷ Run hmmsearch immediately on alignment
10:                    Run SearchTool Reference vs SampleORFs
11:                    AddMatches to SuspiciousSequences
12:                    WriteLog                                ▷ Match summary printed to Log
13:                else                ▷ Other searches will be performed on grouped references
14:                    Add Reference to GroupedReferences
15:                end if
16:            end for
17:        end for
18:        for SearchTool in Non−hmmsearchTools do
19:            Run SearchTool GroupedReferences vs SampleORFs
20:            AddMatches to SuspiciousSequences
21:            WriteLog                                        ▷ Match summary printed to Log
22:        end for
23:        if SuspiciousSequences is not empty then            ▷ If 'main' genes have been found
24:            Clear GroupedReferences                         ▷ Remove already searched references
25:            for Reference in CompanyGenes do                ▷ Based on TroublingTRAVISTable
26:                for SearchTool in Reference do
27:                    if SearchTool == 'hmmsearch' then ▷ Run hmmsearch immediately on alignment
28:                        Run SearchTool Reference vs SampleORFs
29:                        AddMatches to SuspiciousSequences
30:                        WriteLog                            ▷ Match summary printed to Log
31:                    else            ▷ Other searches will be performed on grouped references
32:                        Add Reference to GroupedReferences
33:                    end if
34:                end for
35:                for SearchTool in Non−hmmsearchTools do
36:                    Run SearchTool GroupedReferences vs SampleORFs
37:                    AddMatches to SuspiciousSequences
38:                    WriteLog                            ▷ Match summary printed to Log
39:                end for
40:            end for
41:        end if
42:        Run BLASTP Non−redundantProteinDatabase vs SuspiciousSequences
43:        WriteLog                                            ▷ Match summary printed to Log
44:    end for
45: end procedure
```

Figure 13: TRAVIS Core Algorithm.

2.3.2.2.3 3. TRAVIS Scavenger

Here, all the generated result data is parsed and summarized. Additional annotations and sequences are downloaded from NCBI via EFETCH and ORFs of the references and 'suspicious' sequences are directly compared. The visualization of the sequence organization facilitates the comparison of potential new viruses to the references. The position and length of ORFs in combination with their potential annotation can help telling true positives apart from false positives.

1: **procedure** TRAVIS SCAVENGER(*TRAVISCoreLog*)
2: **Read** *Log*
3: **for** *Sample* **in** *SampleLibrary* **do** ▷ Process each sample completely
4: **Sort** *References* **by** *SuspiciousSequences*
5: **Download** *Reference* **from** NCBI ▷ Additional Information, Origin
6: **Annotate** *SuspiciousSequences* ▷ Based on References
7: **Run** *BLASTP Reference* **vs** *SuspiciousSequences* ▷ Pairwise
8: **Plot** *PairwiseComparisons*
9: **end for**
10: **end procedure**

Figure 14: TRAVIS Scavenger Algorithm.

2.3.3 Data Preparation

This chapter describes the process of setting up the data and references for running TRAVIS.

2.3.3.1 Generation of the Reference Library

The 2017 release of Virus Taxonomy by the ICTV has been used for setting up the reference library for Reovirdae (`https://talk.ictvonline.org/taxonomy/`; Davison *et al.*, 2017). Based on this list, the NCBI database has been manually searched for the respective full genomes, if available (`https://www.ncbi.nlm.nih.gov/`; NCBICoordinators, 2016). Taxonomical information like subfamily and genus has also been added to the reference library for easier evaluation of the results. The RdRp has been used as the 'main' gene (see chapter 2.3.2.1). Sets of complete genomes were obtained from the corresponding assembly report on NCBI, if available. Additional information about the viruses was taken from the respective publications based on the genebank entry.

2.3.3.2 Generation of the Sample Library

The sample library for the search for Reoviridae consists of two parts. The first part are semi-simulated infected transcriptomes where a real, virus-free transcriptome has been infected with mutant of a real virus *in silico*. These have been generated to evaluate the potential efficiency of TRAVIS. The second part consists of real transcriptomes from the 1KITE-project. This was to test whether TRAVIS is able to handle real word data.

2.3.3.2.1 Semi-simulated Infected Transcriptomes

Semi-simulated infected transcriptomes were added to the sample library for benchmark tests. Since there are endless possibilities and limited computing resources, only one scenario was randomized 100 times. One transcriptome (*Gyrinus marinus*, published in Misof *et al.*, 2014) was chosen randomly and a BLASTP-search against the viral refseq library from NCBI (downloaded at 02 Nov. 2017) was conducted for this sample. All sequences that yielded hits were removed from the sample in order to prevent misleading results for this simulation.

1000 contigs from the virus-free sample were chosen randomly to create a semi-simulated virus-free transcriptome. All 10 segments of *Rotavirus A* from the reference library were used to simulate different mutations of a virus that were used to 'infect' the semi-simulated transcriptome. Each segment was mutated in 10%-increments from 10% to 90% distance to the original sequence. Mutation took place randomly at nucleotide level while no InDels were produced and thus keeping the ORF structure intact. If a nucleotide was supposed to change, a check on all affected codons in each frame was performed. If a stop-codon would have been introduced at a codon that was a sense-codon before, another site for mutation was chosen randomly. For each mutation step, these mutated viral sequences

were combined the 1000 drawn sequences from the semi-simulated virus-free transcriptome to make up a semi-simulated virus-infected transcriptome. The original virus sequences were also introduced into the semi-simulated virus-free transcriptome. This process has been repeated 100 times (see Fig. 15).

The use of real sequences ensures more meaningful benchmark results in the context of real world data compared to completely simulated sequences. A comparison with the real samples should provide an estimate on how efficient the pipeline is able to retrieve highly divergent sequences.

```
1: procedure SimulateTranscriptome (TemplateTranscriptome, TemplateVirus)
2:     Initialize NonViralSequences        ▷ Will store sample sequences that do not match any virus
3:     Run BLASTP TemplateTranscriptome vs ViralRefSeq
4:     AddNonMatching to NonViralSequences
5:     Load TemplateVirus                  ▷ Contains all segments of the template virus
6:     for 1 in 100 do                     ▷ Generate 100 random simulated transcriptomes
7:         Initialize SimulatedTranscriptome
8:         Initialize InfectedTranscriptome
9:         Draw 1000 random sequences from NonViralSequences into SimulatedTranscriptome
10:        Join SimulatedTranscriptome with TemplateVirus into InfectedTranscriptome ▷ Original virus
11:        for i in 10 to 90 by 10 do      ▷ Mutate original virus in percentage stepwise
12:            Mutate TemplateVirus by i% into MutatedVirus
13:            Join SimulatedTranscriptome with MutatedVirus into InfectedTranscriptome ▷ i% distance to Original virus
14:        end for
15:    end for
16: end procedure
```

Figure 15: Semi-simulated Infected Transcriptome Generation.

2.3.3.2.2 1KITE Transcriptomes

The transcriptomes from the 1KITE-project were prepared as described in chapter 2.2.2.

2.3.3.3 TRAVIS Control Center Settings

Reovirus genomes consist of short segments with the longest proteins of about 1500 amino acids, therefore it is reasonable to neglect longer ORFs that exceed this limit. The maximum ORF length for evaluation was set at 3000 amino acids. The minimum ORF length was set to 50 because very short ORFs often have no or unknown functions and thus are probably not valuable for the intended interpretation. All searches for 'company' genes have been set to 'main positive'. Thus, only samples where an RdRp-like sequence has been found were considered.

The search parameters were set to default with only limiting the maximum of displayed matches to the best 10 and using an e-value threshold of 10^{-6} to allow very distant hits.

MMseqs2 (v. 5437c6334d659119089cd8758a63838c29753048) was used for clustering the reference sequences with the call parameters '-c 0.01 -v 0 –cluster-mode 0 -s 7.5 –mask 0'. MAFFT (v. 7.302) was used for aligning the reference clusters with the call parameters '–maxiterate 1000 –genafpair –adjustdirection –quiet –reorder'.

For the sequence searches, HMMSEARCH and JACKHMMER from HMMER3 (v. 3.1b2), BLASTP from BLAST+ (v. 2.6.0) and MMSEQS from MMSeqs2 were used on all references.

No manual adjustments were made in between running the three parts of TRAVIS except for adjusting folder paths and the number of usable CPU cores in TCC because TRAVIS Core has been run on a high performance computing cluster on 12 cores (Intel® Xeon® @2.67 GHz) with 106GB memory, whereas TRAVIS Henchman and TRAVIS Scavenger were run on a Desktop computer on 4 cores (Intel® Core™i3-2120 CPU @3.30 GHz) with 16GB memory.

The alignments of the references during the run were not manually checked. Additionally, the calculation time and number of identified suspicious sequence for each search tool were summarized using basic descriptive statistics in R. This has been done in order to evaluate the overall automation process and efficiency of the particular tools and algorithms.

2.3.4 False Positives vs. True Positives

Several properties of the suspicious sequences were considered when evaluating the results. Generally, following criteria had to be fulfilled in order to classify a sequence as true positive:

- The nucleotide sequence had to be of similar length compared to the references. If a true segment has been identified, it should not be much longer than a reference. However, smaller sequences might just be fragments of the virus.
- The ORF structure had to be similar to the reference. Since Reoviridae have mostly monocistronic segments, often only one long ORF was expected to yield matches against the references.
- The matching regions of the suspicious ORFs should not yield significantly better matches to well annotated non-viral sequences. Especially in cases where the hit could be based on a ubiquitously expressed protein domain, it is expected to yield good matches on non-viral sequences. Most importantly, if the sequences are supposed to be part of the host genome, they are most likely false positives.
- If several fragments of a certain viral ORF have been identified, they should match different regions of that ORF. That means *e.g.* three fragments that cover the span of a whole viral segment, where one fragment matches start, middle and end respectively. This would indicate that the virus in the sample could only be sequenced and/or assembled partially.
- If different segments were found, they should show similarities to the segments of the same virus. However, better matches to other viruses cannot be excluded *per se* because of potential re-assortment of segments.

The true positives were collected into a table and annotated with the best matching virus segment by NCBI-accession number for further analysis and comparison. In this case, the best matches were not only determined by sequence identity but evaluated also in context with the other matches within the respective sample. Sequences that could not reliably classified but results still indicated that they might be of viral origin, were labeled as 'questionable'. The number of true positives for each search tool have been set into context with the total number of identified sequences using basic descriptive statistics. This allowed the direct comparison of the used search tools in terms of false positive rate and missed true positive rates.

2.3.5 Genome Organization

The genome organization contributes a lot to the classification of Reoviridae whereas the pure sequence similarity plays a minor role (see chapter 2.3.1; Upadhyaya *et al.*, 1998; Graham *et al.*, 2006; Deng *et al.*, 2012). Since the assembly of the transcriptomes was targeted towards the host and not to extract viruses in the first place, the settings were most likely not ideal for viral sequences. It is a general problem to assemble viral sequences simply due to their high inner-species variation (Eriksson *et al.*, 2008; Yang *et al.*, 2012). These problems reduce the probability of a fully assembled virus within the transcriptomes. However, it was expected to retrieve a large proportion of fragmentarily assembled viral genomes and a method to estimate the size as well as the whole genome organization had to be developed. It is important to note that the approach described in this chapter is highly experimental and not yet part of the pipeline but it is a first simple attempt to make the sequence evaluation more meaningful and comparable between the samples.

Several properties of the potential viral sequences can be derived from the interpretation of the output of TRAVIS that can be used for genome estimation. First, the closest known relative. If the non-redundant protein database for the reciprocal BLAST is up to date, it is possible to find the latest publicly available closest related virus. Second, the position of the match between the suspicious sequence and its closest known relative. Based on these two properties, the completeness of the genome or at least segment of the potential new virus can be reckoned by following the concept of reference mapping. For example, if a transcript has a length of 1000 bp and matches well starting from position 1000 of a virus with a length of 3000 bp, the new virus is probably missing 1000 bp at the beginning as well as at the end of the sequence. Of course this principle can also be applied to *e.g.* three different fragments that match different regions of the same reference virus. If one fragment matches the start, the second in the middle and the third at the end of the reference, it is likely to have a nearly full segment where the connective regions of the new virus have either not been properly sequenced or assembled.

However, mapping or aligning the suspicious sequences to the reference viruses is very difficult and error-prone at nucleotide level if the sequences are very distant to each other. Since the identification and verification of the suspicious sequences is already based on the well alignable region of the particular ORFs, similar methods should be able to make reference mapping possible based on the respective amino acid sequences. To achieve that, the suspicious ORFs were aligned with the corresponding ORF of the reference by MAFFT on amino acid level. Pal2Nal was then used to infer the original nucleotide sequences of the respective amino acid sequence and thus a complete nucleotide alignment has been created. The suspicious sequences were then used to calculate a consensus sequence with FASconCAT-G to obtain the complete estimated sequence including gaps also to indicate

the missing trails. Additionally, a consensus sequence was calculated for the amino acid alignment to also have an estimate about the protein (see Fig. 16 and Fig. 17).

To make the generated sequences comparable and give an additional objective measure for the obtained consensus sequences, the Gapless Forced Alignment Score (GFAS) was introduced. In its essence, it is an identity expressed as percentage of two given sequences. In contrast to the more sophisticated BLAST, GFAS scores the complete sequences based on a pairwise alignment that strongly penalizes gaps. GFAS thus yields lower scores and does not take into account InDels or ambiguities compared to BLAST. This alignment is created by using MAFFT with high gap penalty costs. The number of positions in the alignment, where both sequences had an identical character state, were counted and divided by the number of positions where both sequences do have character states except gaps.

To test the explanatory power of GFAS, simulations have been made. For that, one million pairs of random sequences of lengths between 1 and 10000 amino acids have been created. The GFAS of each pair has been calculated and the median was 4% GFAS whith an upper quartile of 5% GFAS. These statistics in combination with the density estimate (see Fig. 18) imply that most likely GFAS-identities above 5% probably indicate non-randomness.

1: **procedure** GENOME ESTIMATION(*SuspiciousSequences*,*Reference*)
2: **Align** *AminoAcidSequences* **into** *AminoAcidAlignment*
3: **GenerateConsensus** *AminoAcidAlignment* ▷ FASconCAT-G
4: **CalculateGaplessForcedAlignmentScore** *AminoAcidConsensus* **vs** *Reference*
5: **ReverseTranslate** *AminoAcidAlignment* **into** *NucleotideAlignment* ▷ Pal2Nal
6: **GenerateConsensus** *NucleotideAlignment* ▷ FASconCAT-G
7: **end procedure**

Figure 16: Genome Estimation Algorithm.

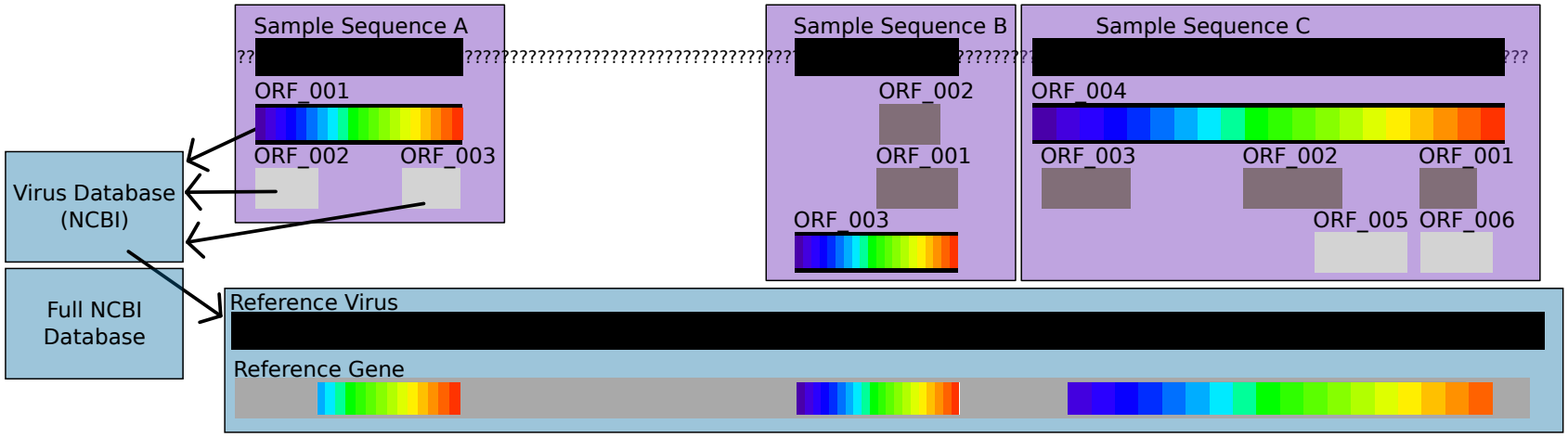

Figure 17: Core Concept of the Genome Estimation.

Depicted is an example for three suspicious sequences 'Sample Sequence A, B and C' that are supposed to be closest related to the 'Reference Virus'. It is a summarized plot for the different sequences that are part of the output of TRAVIS Scavenger. Each sample sequence matches different regions of the reference virus. Since the matching regions in this case are unambiguous the missing parts of the potential new virus can be estimated based on the reference virus. The missing parts are represented as question marks.

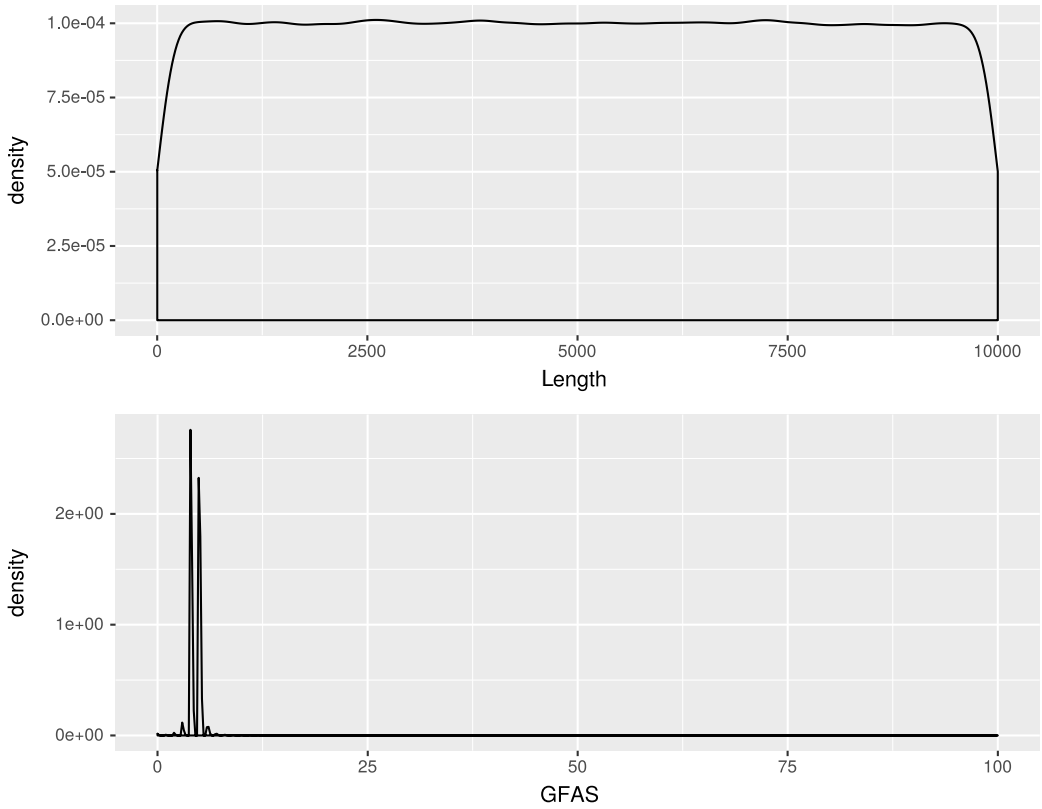

Figure 18: Simulation of GFAS-identities for Randomized Sequences.

Density estimates of GFAS-identities for one million randomly drawn amino acid sequences of up to 10000 amino acids in length. The density of the lengths of simulated was distributed in such a way, that nearly all potential lengths were covered (above). GFAS-identities peaked at 4% - 5% suggesting a deviation of sequence similarity from random chance above 5% GFAS-identity (below).

2.3.6 Inference of Phylogeny

Originally, Reoviridae have been classified by comparing the patterns of gel electrophoresis. Today, the RNA-dependent RNA polymerase is considered to be the only gene that allows reliable and meaningful phylogenetic reconstruction and classification across the family of Reoviridae on molecular level (Attoui *et al.*, 2002; Distéfano *et al.*, 2003). Other segments can be used for reconstructing phylogenies within species (von Bonsdorff and Maunula, 1998). Additionally, it has been shown that several segments of *Epizootic Haemorrhagic Disease virus* (EHDV, Reoviridae) support similar geographical origins on sequence level while other segments from the same sample and virus hint towards another geographical origin (Anthony *et al.*, 2009). *Barley yellow dwarf virus* (type species of Luteoviridae) has been chosen as the outgroup based on an InterProScan result of the RdRp of *Rotavirus C* (NC_007547), where a *Luteovirus*-like polymerase domain has been detected (see Fig. 19).

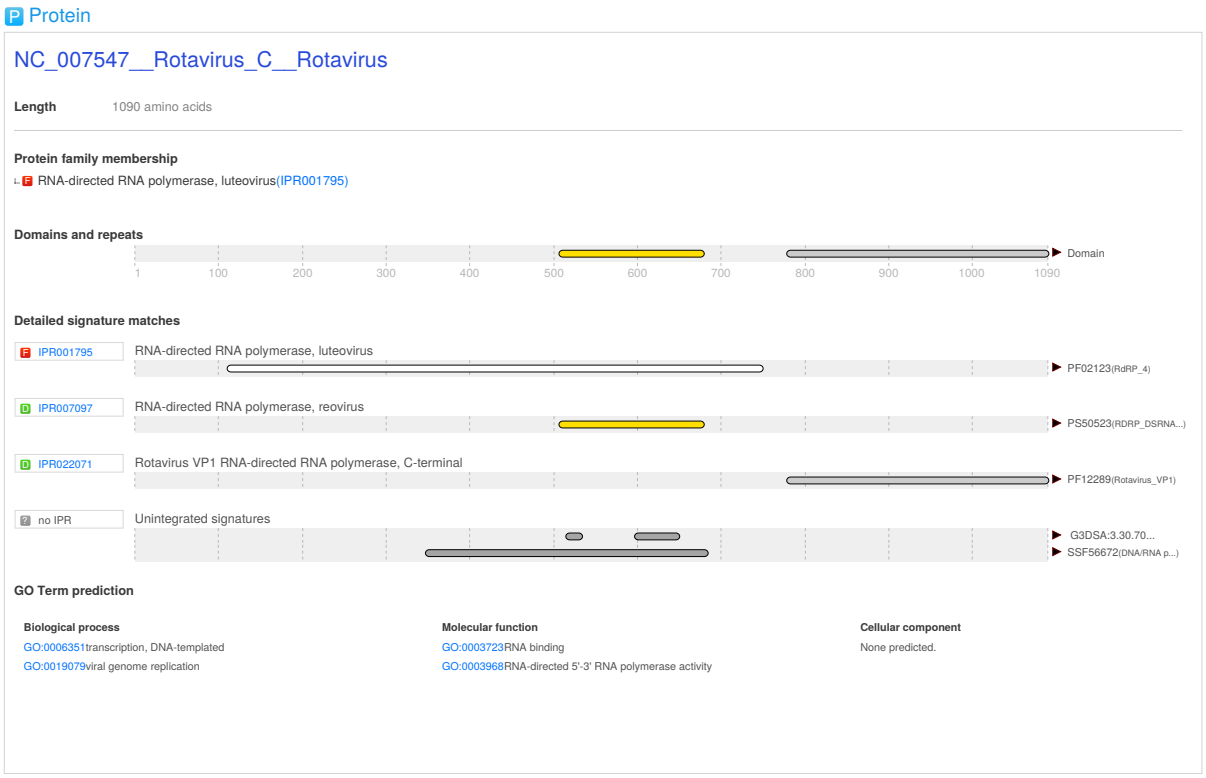

Figure 19: InterProScan of the RdRp of *Rotavirus C* (NC_007547).
InterProScan detected a *Luteovirus*-like polymerase domain in addition to the expected *Reovirus*-like polymerase domain as it has been the case for most of the other Reoviridae-references. In other cases no, or only small non-Reoviridae-specific domains were detected.

It also has to be considered that it is difficult to infer a phylogeny for many taxa that have relatively short sequences. However, in such cases *e.g.* Neighbor-Joining (NJ) methods can outperform Maximum Likelihood (ML) approaches but still result in similar topologies (Takahashi and Nei, 2000). An additional problem with segmented viruses is that it is hard

to reconstruct proper phylogenies because of the re-assortment of segments that can happen during co-infections with different viruses or strains. Since the 1KITE transcriptomes are expected to contain very distantly related new reoviral sequences, the inferred phylogenies are predicted to be very unstable. Thus it is necessary to compare different variations of phylogenetic reconstruction. For this purpose a NJ method (APE-package, R), an improved NJ method (FastME) and a ML method (PhyML) have been used (see chapter 2.1.2). Blosum62 and WAG substitution models have been set for FastME and PhyML to see the influence of substitution model on the topology. RdRp sequences of all reference viruses from the reference library, all true positive viruses from the transcriptomes and their best matches based on the TRAVIS Scavenger plots were included into an alignment for phylogenetic reconstruction. The initial alignment of the RdRps on amino acid level has been calculated using MAFFT (E-INS-i). TrimAl was then used to trim columns stepwise increasing the gap-threshold from 10% to 90% in 5%-steps in order to see the influence of gap-trimming on the topology. Each of the trimming steps resulted in an alignment that has been the base for phylogenetic reconstruction using the aforementioned methods with 1000 bootstrap replicates each. The overall bootstrap support was the criterion for choosing the best supported trees among the resulting trees for each method. Since the sequences within the alignments were expected to be very diverse and preliminary tests indicated that the topologies calculated for such diversity would be very unstable, a threshold of 60% of bootstraps was considered as 'confidence'-level indicating that more than half of the calculated trees for a specific method were showing the respective topologies. Additionally, an alignment of the RdRps based on the hydrophobicity has been calculated and as well treated in the same way as the pure amino acid alignment. The best supported trees for each method were plotted with using Newick Utilities.

The topologies for each method were compared pairwise using quartet distances calculated with tqDist. The percentage of identical topologies for all resolved quartets was used as an indicator on how consistent the topologies between the methods were. Additionally, all branch lengths for branches with bootstrap supports lower than 90% were set to zero and thus considered unresolved. The consistency in topologies were again calculated with tqDist in order to estimate the influence of nodes with low support. The obtained similarity estimates were summarized using basic descriptive statistics. In order to show the conflict in resolution, the original alignments were also used to generate ConvexHull-NeighborNets via SplitsTree.

Based on the best NJ-phylogeny (APE, R), taxa that form monophyletic clades of at least three taxa that could be found in the other phylogenies based on the pure amino acid alignment were grouped and color-coded for the best supported tree (*i.e.* the variation with the highest median bootstrap values) of each phylogenetic reconstruction

method. Schematic block-like summaries of those grouped topologies were made to obtain interpretable diagrams. Additionally, a scaled variation of these schemata were made for the sake of readability. Together with the results from tqDist, this grouping can help to identify the stable proportions of the calculated phylogenies. These groups were also applied to the trees based on the the hydrophibicity alignment as well as the SplitsTree networks.

To summarize the phylogeny of all potential new Reoviruses, 'transfer' bootstraps for the best supported PhyML tree were calculated using BOOSTER and plotted via ggtree.

3 Results

3.1 Preliminary Work

The preliminary work showed that the transcriptomes from the 1KITE-project were indeed containing previously unknown sequences of potential viral origin. General summaries and tentative phylogenies were calculated to show the potential of transcriptomoc data combined with profile Hidden Markov Models.

3.1.1 Sequence Search and Phylogenetic Tree Reconstruction

All available information about the samples and potential taxonomy of the viral sequences have been summarized. Based on this summary, several sub-summaries have been made to get an overview of the obtained potential viral sequences. In total, there were 2406 potential viral sequences distributed over 757 of the transcriptomes across all grouped orders (see Fig. 21, Fig. 20 and Table 2). There were significant differences in the proportion of infected transcriptomes between the grouped orders (Pearson's Chi-squared test; x-squared = 69.495, df = 20; p-value = 2.202e-07). A post-hoc test for identifying the detailed significant differences was done with the FIFER-package for R (see Table 3; Fife, 2017). According to these tests, only Amphiesmenoptera and Polyneoptera stand out. While all other orders show less clean than infected samples, Amphiesmenoptera have more clean than infected samples and in Polyneoptera there are nearly as many clean as infected samples (see Fig. 20 and Table 3).

2367 of the potential viral sequences originated from non-bloodfeeding while only 39 were found in bloodfeeding arthropods. This supports the assumption, that most known viruses in arthropods are likely from blood-feeding arthropods (Arboviruses) because there is a bigger medical and therefor historical interest in research. Despite the pHMMs were designed for specific virus groups, several contigs have been identified as viral based on multiple pHMMs (see Table 4). While some sequences could only be identified as viral by specific pHMMs, especially the Flaviviridae, Nege-like, Toga-like and Picorna-like viruses showed more overlap than the other groups. The fact that there is overlap between several distinct viruses supports a potential relationship.

The obtained sequences have an average length of 2999 bp with a minimum of 198 bp and a maximum of 20930 bp. 1478 sequences were long enough to confidently be included in multiple sequence alignments and derive tentative phylogenies (see Fig. 22, larger high resolution variations with sequence IDs can be found in chapter 1 of the digital appendix). Branches of reference viruses that are associated with arthropods were colored orange. Red branches indicate the potential viruses from the 1KITE transcriptomes and black branches reference viruses, that are associated with non-arthropod hosts. Known

groups of viruses have been labeled and marked with a gray overlay. Additionally, sequences that form clades have been assigned roman numerals. Blue dots indicate that the full coding sequence is known, red dots indicate a full genome. For a better overview and resolution, sequences were grouped into A: non-segmented RNA viruses (-), B: segmented RNA viruses (-), C: Flavivirus-like superfamily (+), D: Picornavirus-like viruses (+), E: Togavirus-like superfamily (+) and F: Nidovirales-like viruses. There are many sequences form the transcriptomes that form clades with only other known arthropod-associated viruses (A: II, IV, VIII, XIII; B: I, III, IV, VI, VIII, IX, X, XI, XV, XVI; C: IV, VII, IX; D: I, IV, XV; E: I) and some clades with only non-arthropod-associated viruses (D: V, IX, X, XIII, XVI, XVII. XVIII; E: VII, XVI, XIX; F: I, III). However, another large portion forms clades only with other sequences from the transcriptomes or are just single sequences on very long branches (A: V, VII, IX, X, XI, XII; B: II, V, XVII; C: II, III, V, VI, X, XI, XII, XIII; D: III, VI, VII, XII; E: II, III, IV, V, VI, VIII, X, XI, XII, XIII, XIV, XV, XVII, XVIII, XXI; F: V). These phylogenies show that was possible to extract viruses from the 1KITE transcriptomes that are very distantly related to known viruses. Based on the reference genera, these viruses potentially form new genera and families. Detailed analysis of the relationships are currently under investigation as stated in chapter 2.2.1.

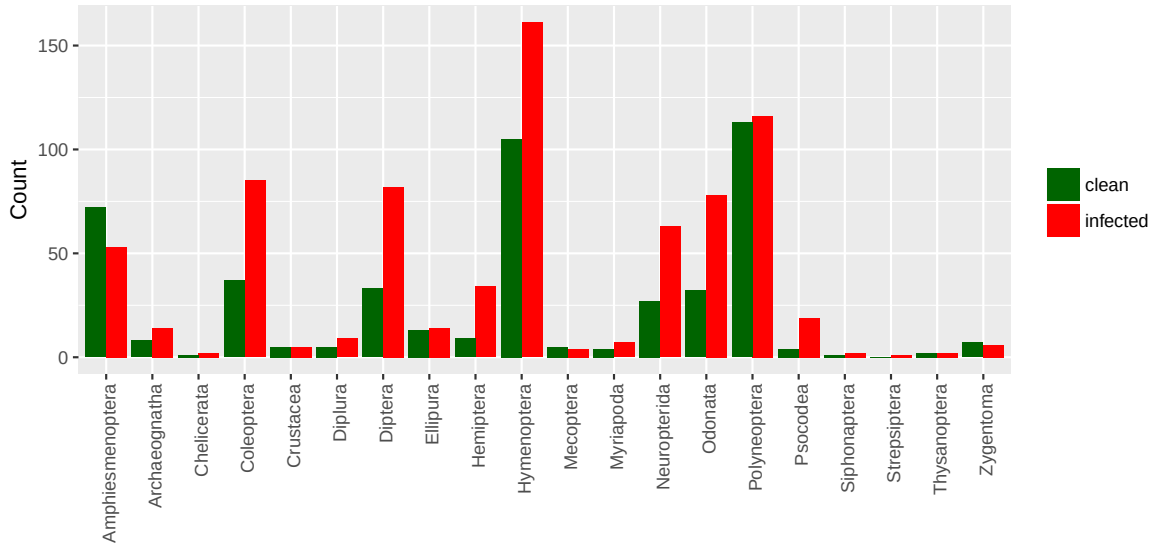

Figure 20: Infection Status.

Displayed is the relation of infected to clean number of transcriptomes per order.

Table 2: Viral Load by Order.

Number of clean and infected transcriptomes and potential viral sequences by grouped host orders.

Grouped Order	Clean	Infected	Potential Viral Contigs
Amphiesmenoptera	72	53	105
Archaeognatha	8	14	20
Chelicerata	1	2	12
Coleoptera	37	85	285
Crustacea	5	5	10
Diplura	5	9	49
Diptera	33	82	310
Ellipura	13	14	47
Hemiptera	9	34	127
Hymenoptera	105	161	437
Mecoptera	5	4	7
Myriapoda	4	7	38
Neuropterida	27	63	207
Odonata	32	78	275
Polyneoptera	113	116	357
Psocodea	4	19	98
Siphonaptera	1	2	2
Strepsiptera	0	1	1
Thysanoptera	2	2	4
Zygentoma	7	6	15

Table 3: Post-Hoc Test of Viral Load by Order.

Number of clean and infected transcriptomes and potential viral sequences by grouped host orders.

Compared Grouped Orders	Raw P-value	Adjusted P-value
Amphiesmenoptera vs. Coleoptera	0.0000	0.0015
Amphiesmenoptera vs. Diptera	0.0000	0.0014
Amphiesmenoptera vs. Hemiptera	0.0000	0.0019
Amphiesmenoptera vs. Hymenoptera	0.0010	0.0146
Amphiesmenoptera vs. Neuropterida	0.0001	0.0039
Amphiesmenoptera vs. Odonata	0.0000	0.0014
Amphiesmenoptera vs. Psocodea	0.0005	0.0099
Coleoptera vs. Polyneoptera	0.0007	0.0122
Diptera vs. Polyneoptera	0.0003	0.0099
Hemiptera vs. Polyneoptera	0.0007	0.0122
Neuropterida vs. Polyneoptera	0.0018	0.0204
Odonata vs. Polyneoptera	0.0004	0.0099
Polyneoptera vs. Psocodea	0.0038	0.0415

Table 4: pHMM Result Overlap.

Listed are the number of different contigs from the transcriptomes that have been identified as viral by a certain pHMM combination. A=Arenaviridae, B=Bunyaviridae, O=Orthomyxoviridae, F=Flavviridaei, Ni=Nidovirales, P=Picorna-like, Ne=Nege-like, T=Toga-like, M=Mononegavirales-like

pHMM Combination	Identified Contigs
AB	4
B	178
BO	55
F	72
FM	75
FMNePT	4
FMNiP	11
FMP	166
FMPT	2
FNe	2
FP	2
FT	4
M	233
Ne	159
NeP	14
NePT	22
NeT	322
Ni	9
NiP	122
O	60
P	887
T	5

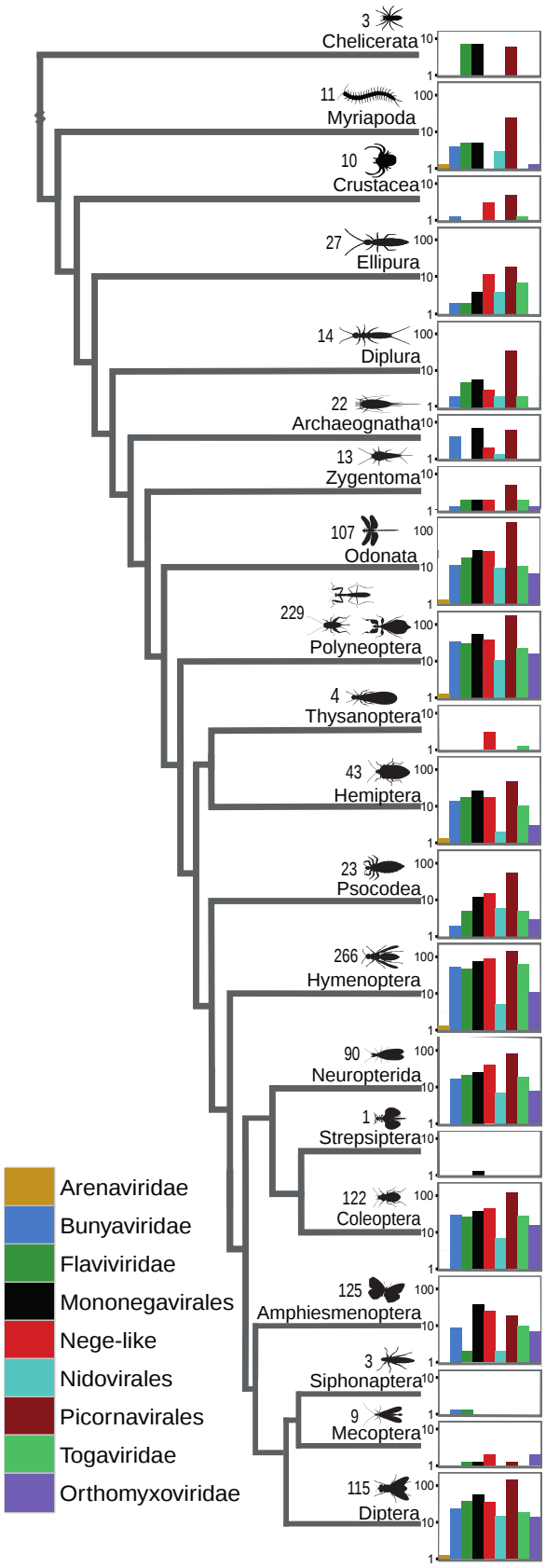

Figure 21: Virus Distribution.

Amount of contigs identified by the different pHMMs across the grouped arthropod orders. The number in front of the arthropod icons indicate the number of scanned transcriptomes.

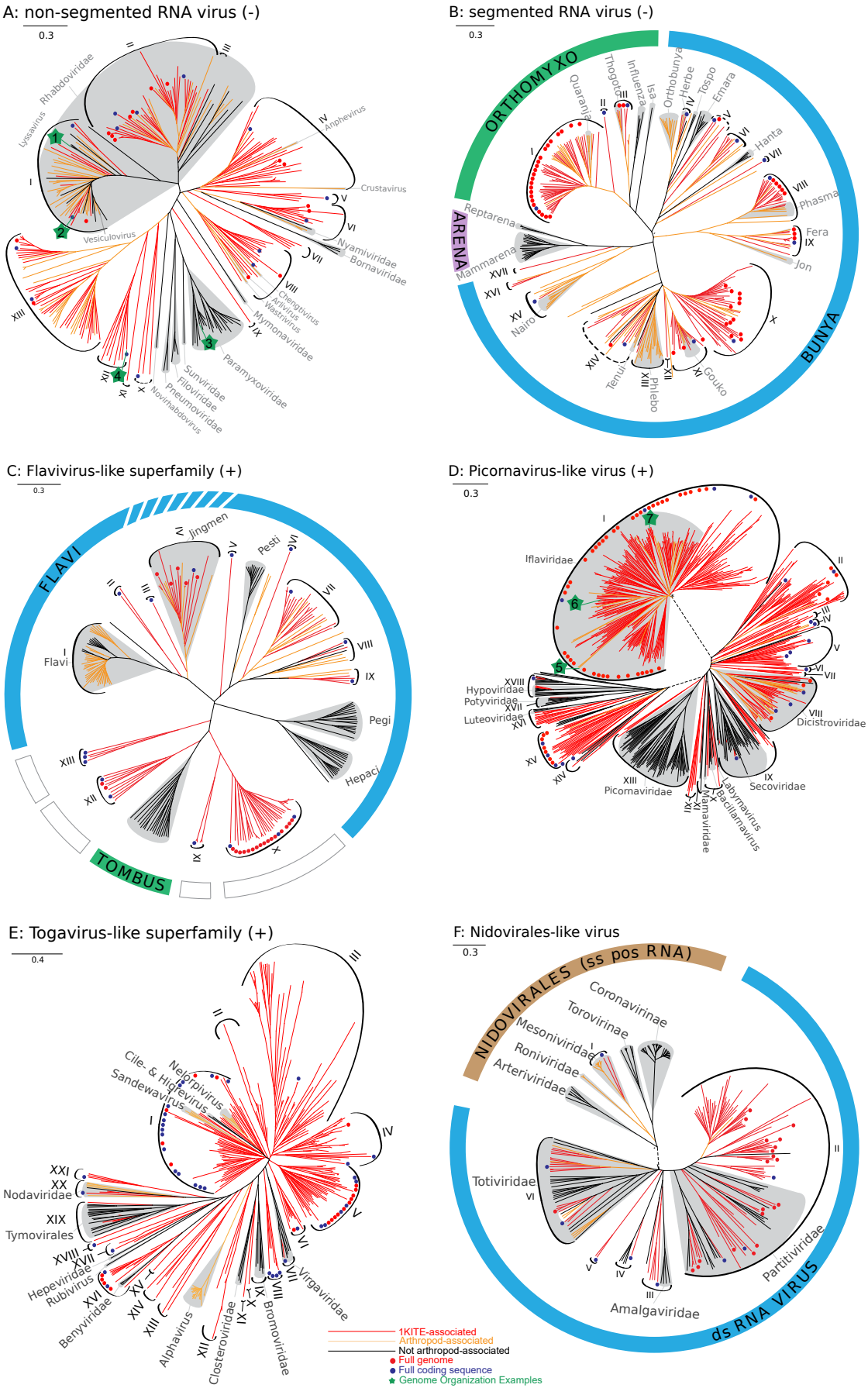

Figure 22: Tentative Phylogenetic Trees.

Reconstructed phylogenies for 1478 sequences in context with their expected closest known relatives. Exemplary genome structures are shown in chapter 3.1.2. Figure: Dr. Florian Zirkel

3.1.2 Genome Organization

The preliminary plots allowed closer examination of the potential viral sequences and enabled to verify most findings (see Fig. 23). Verification was possible especially for sequences which show relatedness to virus groups that have a more or less conserved ORF patterns. However, for virus groups that have most of their genes on one continuous ORF (polyprotein), protein domain structure was more convincing. The preliminary plots have been used as a template to generate more detailed plots for some selected sequences based on additional InterProScan annotations (see Fig. 24). The size of the sequences and the positions of the identified protein domains in comparison to known reference viruses gave more insight to the affiliation towards a specific virus group. Here, four new viruses and three reference viruses were chosen as example for describing the genome organization evaluation process (see Fig. 23 and Fig. 24). Sequence names starting with '1KV' are originating from the transcriptomes, all other viruses are references from NCBI, starting with their respective genebank accession number. Some functionality was neglected in these genome depictions for the sake of clarity.

The first four sequences belong to the non-segmented RNA viruses.
1KV_mono_000167 shows a similar genome structure to NC_001542_Rabies_virus. They are about 12000bp long and have five ORFs with lengths of over 200 amino acids. There are four smaller ORFs in the first half of the sequence with similar lengths and one large ORF on the second half of the sequence. The first ORFs of both structures were identified as nucleocapsid proteins, the third ORF as matrix protein and the fourth ORF as glycoprotein. The longer fifth ORF carries the polymerase functionality. Based on the InterProScan results, it was possible to derive protein domain structure on these polymerase ORFs. In the beginning, the actual replicase domain is placed, followed by an mRNA cap formation domain. A methyltransferase domain follows towards the end of the ORF. A similar structurization of the last ORF can be found in NC_002200_Mumps_virus. Its sequence is about 3000 bp longer and contains two more ORFs compared to the previous viruses. Yet the first ORF still contains the nucleocapsid gene. 1KV_mono_000076 is again about 12000 bp long. Its last ORF is similar to the previous ones, but lacking the methyltransferase domain. However, there are only two larger ORFs instead of four. Despite there was no blast match available in the first ORF, it could have been identified as a putative glycoprotein by InterProScan. The second ORF is supposed to carry the nucleocapsid. So, compared to the other viruses in this group, the nucleocapsid and the glycoprotein seem to have switched positions.

The next three sequences belong to the Picornavirales-like viruses and have all genes encoded within one ORF on a polyprotein. Their length is about 9000 bp and the blast matches identify the ORFs only as polyprotein. Here, the domain structure detected by

InterProScan is very valuable for deriving functionality and thus verifying the viral origin. NC_003781_Infectious_flacherie_virus and 1KV_picorna_000579 share a nearly identical structure. The first three domains contribute to the nucleocapsid, a helicase domain is found in the middle and a peptidase followed by the replicase in the end of the polyprotein. In contrast to that, the structure of 1KV_picorna_000119 is modified. While the detected domains are the same as in the previous viruses, the three nucleocapsid domains are positioned at the end of the polyprotein. The other domains, *i.e.* helicase, peptidase, and replicase, are in the same composition as in NC_003781_Infectious_flacherie_virus and 1KV_picorna_000579. This case demonstrates that genome structure may change by rearranging a whole polyprotein while keeping the functionality intact.

By comparing the genome structures of the found potential viral sequences to known references, it was possible to estimate the completeness of the genomes. In total, 285 of the potential viral sequences have been estimated to contain the full coding sequence (CDS) and 2121 a partial CDS compaired to known reference viruses based on their length (see chapter 3.1.1 and Fig. 22) and genome structure.

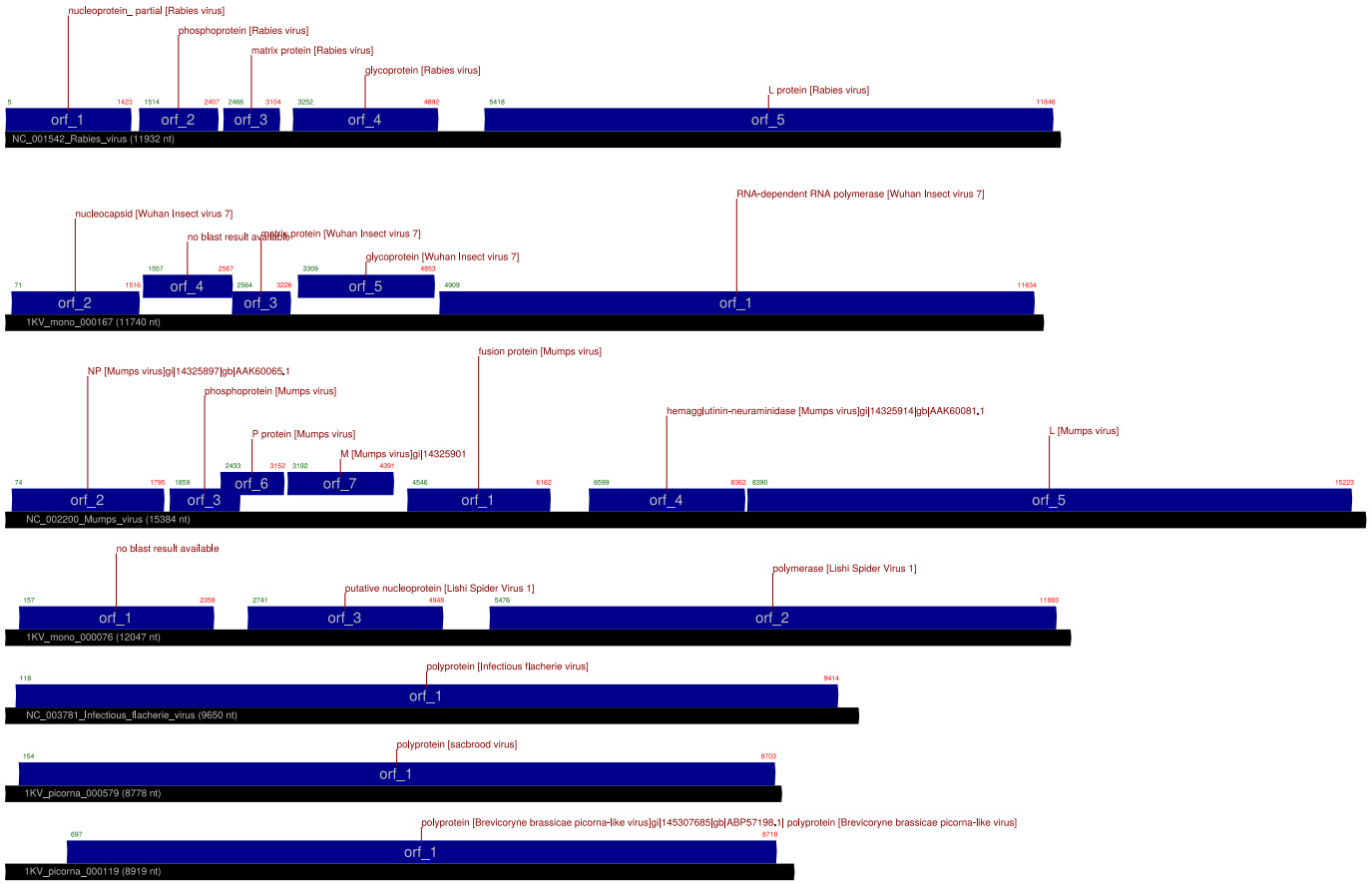

Figure 23: Preliminary Plots of the Genome Organization.

Automatically generated genome structure plot annotated with the best scoring blast matches. Note that ORF numbering is based on the internal handling of data within the plot script and has no certain importance. Green numbers indicate the start and red numbers the end position of the respective ORF.

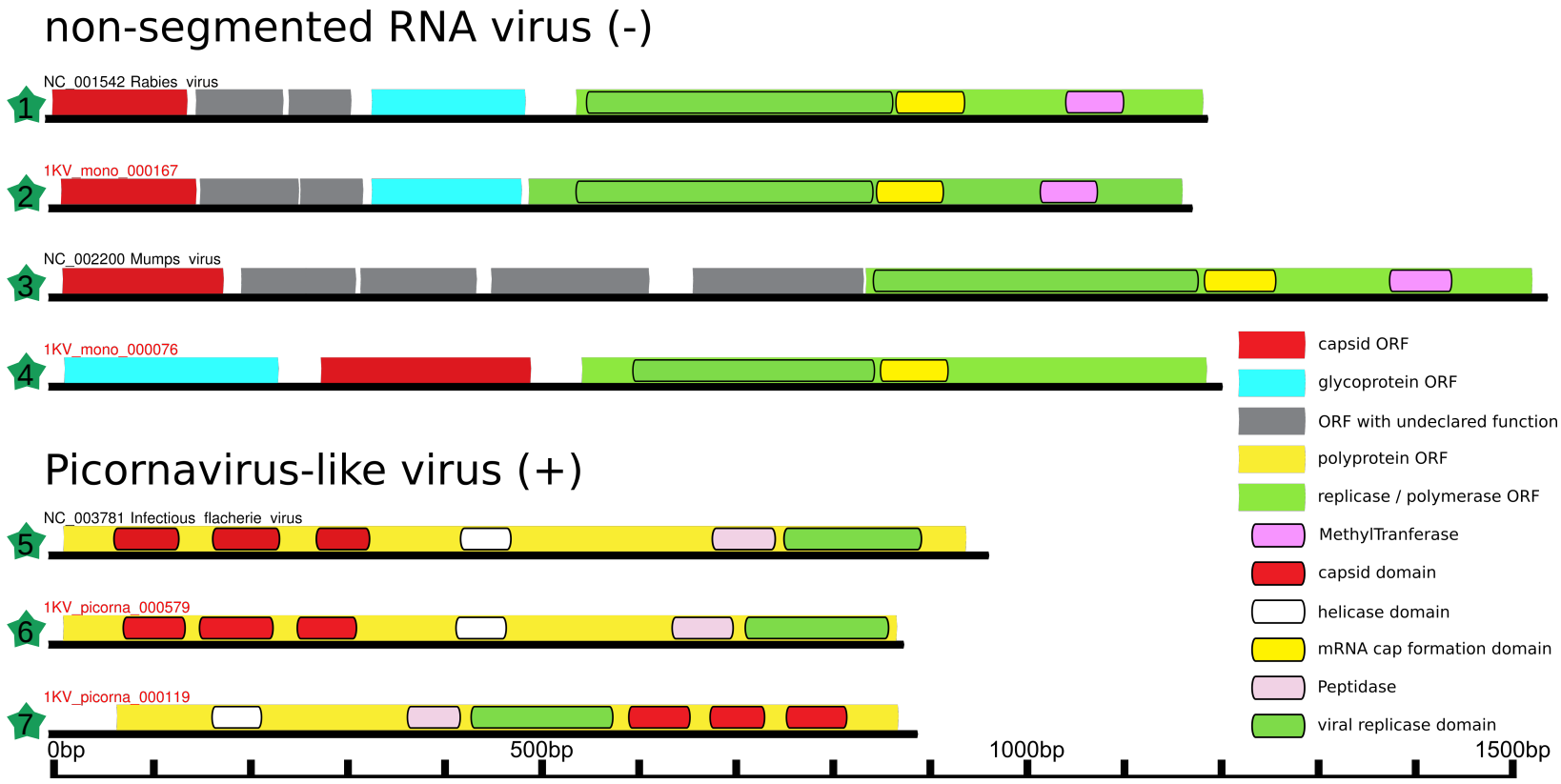

Figure 24: Detailed Genome Organization.

Manually checked and modified genome organization based on Fig. 23 and the respective InterProScan protein domain information. ORFs have been color coded by their functionality and additional symbols have been introduced for displaying protein domains.

3.2 TRAVIS

3.2.1 Simulations

It was possible to retrieve most of the randomly mutated *Rotavirus A* segments even if they were up to 80% mutated (see Fig. 25). This worked well for all segments. Sequences that were mutated 90% could not have been identified in any case. Some members of Reoviridae are known to have sequence similarities down to approximately 10-20% amino acid identity compared to other members of the family (Attoui *et al.*, 2006a). Thus, the retrieval rate of TRAVIS for sequences that are up to 80% mutated at nucleotide level does not seem sufficient. However, the mutation for the simulated transcriptomes was randomly assuming no rate heterogeneity. Since there are conserved regions in the segments of real Reoviridae, one can assume that this 80% maximum mutation rate for being detectable probably applies to the conserved domains and not the whole sequence.

However, HMMSEARCH seemed to perform poorly compared to the other algorithms although it is claimed to be able to detect very distant homologies. There are three main explanations for this. First, it was not possible to create alignments for all ORFs automatically based on the used settings. Therefore, no pHMM could have been built for these respective segments. Second, the alignments that have been created were not checked and reduced to the conserved motifs leaving many areas with little to no phylogenetic signal that could have mislead the algorithm. Third, the mutations happened randomly and most likely destroyed the conserved domains. This eliminated the signatures, HMMSEARCH has been designed for. In contrast to that, JACKHMMER found the segments reliably. MMseqs2 was able to identify the correct sequences only up to 70% mutation rate.

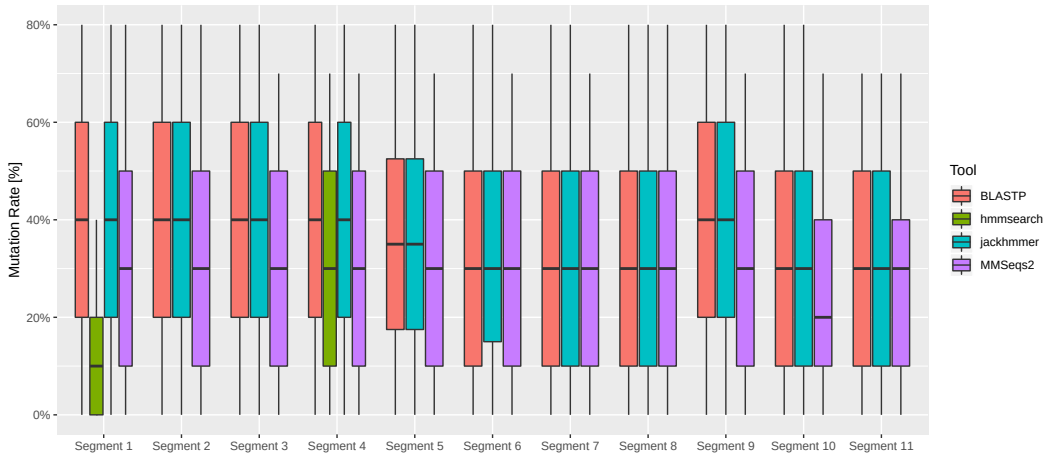

Figure 25: Segment Retrieval Efficiency.

Percentage of mutation rate that could have been correctly identified as viral categorized by the used search tool. Segments could have been identified based on the used reference library if they were up to 80% mutated.

3.2.2 1KITE Transcriptomes

2665 contigs were flagged as suspicious in total, where only 357 were considered to be true positives based on the criteria stated in chapter 2.3.4. As expected, the amount of detected potential viral sequences as well as the calculation time differed highly between the used search tools (see Table 5, Table 6 and Fig. 26). Since HMMSEARCH could only be run on alignments, the overall detection rate for HMMER3 was estimated by combining the results of HMMSEARCH and JACKHMMER. Considering the missed true positives that could be identified by the other methods, HMMER3 scores best by missing the least true positives compared to the other search tools. This comes at the costs of having the highest rate of false positives among the search tools that have been used for identifying the suspicious sequences based on the Reoviridae reference library.

MMSeqs2 delivered the least amount of false postives at a considerable faster speed than BLAST, but also missed true positives the most. This is likely a candidate for more conservative searches.

The reciprocal BLAST of the suspicious sequences versus the non-redundant protein database (NR) showed that it was possible to find matches within that database for 2521 (95%) of the total 2665 sequences. Thus the chance to have better matches for the false positives than based on the Reoviridae reference library was higher and identification of the false positives easier. Overall, the combination of different tools allows the identification of more true positives and confirming these with other algorithms.

While the initial searches based purely on the small Reoviridae reference library was fairly short, the reciprocal BLAST of the obtained suspicious sequences versus the non-redundant protein database from NCBI (NR) took at least 27.5 times longer for the maximum time of BLASTP compared to BLASTP_vs_NR (see Table 6 and Fig. 26). Especially HMMSEARCH and JACKHMMER were surprisingly fast although they are generally considered as slower than the other used algorithms (Madera and Gough, 2002; Johnson *et al.*, 2010; Steinegger and Söding, 2017). Taken into account the minimal missed true positives and the fastest search times, a pure HMMER3-based run of TRAVIS can probably detect most of the potential viruses and save a large proportion of the time. However, the reciprocal BLAST versus the NR still would still need to be done to identify the large number of false positives.

Since the overall sensitivity of the search tools has been set very high, it was expected that many false positives will be found. However the total rate of false positive exceeded the expectations. Table 7 contains a list of all the transcriptomes that only contained false positives.

Table 5: Comparison of the Number of Suspicious Sequences by Search Tool.

Results HMMSEARCH and JACKHMMER have been combined to estimate the total efficiency of HMMER3. The overall false positive rates are very high with at least 68% for MMSeqs2.

Search Tool	Total Detected	Total Missed	True Detected	True Missed	False Positive Rate
hmmsearch	502	2163	140	217	73%
jackhmmer	2212	553	319	38	86%
BLASTP	1105	1560	317	40	72%
MMSeqs2	838	1827	274	83	68%
HMMER	2489	176	336	21	87%
BLASTP vs NR	2521	144	320	37	88%

Table 6: Computation Times for all Search Tools per Sample.

Computation times for each search tool of the initial searches based on the initial Reoviridae-reference library including the reciprocal BLAST of the suspicious sequences versus the non-redundant protein database.

Search Tool	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
hmmsearch [min]	0	0	0	0.0065	0.01	0.11
jackhmmer [min]	0.26	0.46	0.56	0.94	0.70	8.90
MMSeqs2 [min]	0.45	1.43	1.96	2.11	2.60	6.53
BLASTP [min]	0.86	3.82	4.92	5.17	6.21	14.86
BLASTP vs NR [min]	5.56	58.40	88.58	114.84	158.06	413.03

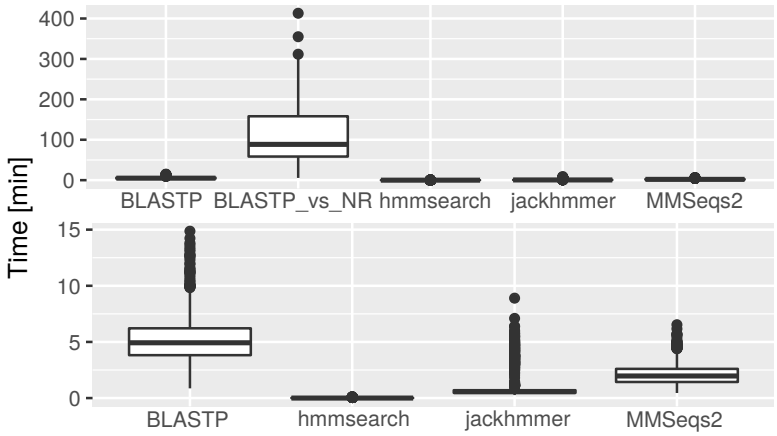

Figure 26: Computation Times for all Search Tools per Sample.

Above: Boxplots of the computation times [min] for each search tool of the initial searches based on the initial Reoviridae-reference library including the reciprocal BLAST of the suspicious sequences versus the non-redundant protein database.

Below: Boxplots of the computation times [min] for each search tool of the initial searches based on the initial Reoviridae-reference library.

Some segments seemed to cause false positives surprisingly often: segment 1 of *Avian orthoreovirus* (NC_015132), segment 1 of *Nelson bay reovirus* (AF218360), segment 8 of *Kadipiro virus* (NC_004208), segment 11 of *Liao ning virus* (NC_007746), segment 3 of *Grass carp reovirus* (KU254568), segment 6 of *Dendrolimus punctatus cypovirus 22* (NC_025850) and segment 10 of *Dendrolimus punctatus cypovirus 22* (NC_025838). Those segments were combined into a small database and have been checked with the false positives via BLASTP. 2296 (99%) of the false positives yielded matches to these fallacious references (see Table 8). In order to investigate that matter more closely, the sequences were checked for fallacious domains. It is known that several Reoviridae proteins show sequence similarities to genes that can also be found ubiquitously in organisms. These proteins contain *e.g.* RNA-binding sites, zinc-fingers, or coiled-coil helices (Attoui *et al.*, 2006a,b). Some example false positives from the transcriptomes have been chosen to illustrate the potential of these fallacious domains to cause false positives in Fig. 28,Fig. 27, Fig. 29, Fig. 30 and Fig. 31. InterProScan of the fallacious proteins revealed that the matching region to the suspicious sequence was consistent with the position of the detected ubiquitous domains. Exported graphics have been manually modified, reduced to the most important aspects, and adjusted to match the respective ORFs and allow direct comparison. The fallacious sequences used in the reference library contained ubiquitously expressed protein domains like coiled-coil helices (see Fig. 27), double-stranded RNA binding motifs (see Fig. 28), zinc fingers (see Fig. 29), and (Transmembrane-)signalling peptides (see Fig. 30). However, there also were occurrences of detected proteins that were more difficult to evaluate.

Table 7: Assemblies That Contained Only False Positives.

All identified contigs in these assemblies were most likely false positives.

INSbusTBQRAAPEI-82	INShauTADRAAPEI-95	INShauTANRAAPEI-95
INShauTAQRABPEI-11	INSnfrTAWRAAPEI-11	INSnfrTBBRAAPEI-16
INSnfrTBGRAAPEI-93	INSfrgTAWRAAPEI-43	INSjdsTALRAAPEI-39
INSjdsTAORAAPEI-44	INSjdsTAPRAAPEI-45	INStmbTATRAAPEI-9
INStmbTAQRAAPEI-94	INStmbTAXRAAPEI-16	INStmbTBJRAAPEI-36
INStmbTBORAAPEI-46	INStmbTBPRAAPEI-20	INSytvTAHRAAPEI-17
INSytvTASRAAPEI-45	INSytvTBARAAPEI-94	INSytvTBMRAAPEI-45
INSytvTBURAAPEI-79	INSytvTBHRAAPEI-14	INSytvTBYRAAPEI-22
INShkeTABRAAPEI-95	INSswpTATRAAPEI-13	INSswpTBLRAAPEI-41
INShkeTAHRAAPEI-94	INShkeTAKRAAPEI-36	INShkeTAMRAAPEI-39
INShkeTBERAAPEI-75	INShkeTBSRAAPEI-13	INSeqtTBDRAAPEI-84
INSeqtTAMRAAPEI-95	INSeqtTBMRAAPEI-9	INSeqtTBCRACPEI-79
INSeqtTBQRAAPEI-84	INSeqtTBRRAAPEI-87	INSeqtTBWRAAPEI-94
INSeqtTCYRAAPEI-46	INSeqtTDARAAPEI-56	INSeqtTDGRAAPEI-84
INSeqtTDPRAAPEI-11	INSeqtTAJRAAPEI-35	INSlupTBHRAAPEI-21
INSqiqTAHRAAPEI-18	INSlupTAWRAAPEI-9	INSntgTABRAAPEI-216
INSntgTAMRAAPEI-203	INSqiqTBPRAAPEI-94	INSqiqTCRRAAPEI-71
INSqiqTDBRABPEI-118	INSobdTBFRAAPEI-109	INSobdTDARAAPEI-57
INSobdTDBRAAPEI-61	INSobdTDSRAAPEI-17	INSobdTEFRAAPEI-41
INSerlTBORAAPEI-62	INSqzbTABRAAPEI-210	INSerlTAXRAAPEI-21
INSerlTBYRAAPEI-16	INSkzdTALRAAPEI-32	INSkzdTAORAAPEI-35
INSerlTCJRAAPEI-35	INSofmTAJRAAPEI-56	INSofmTAKRAAPEI-57
INSofmTAWRAAPEI-109	INSofmTCZRAAPEI-83	INSqiqTBHRAAPEI-71
INSerlTBIRAAPEI-43	INSofmTBJRAAPEI-61	INSofmTBSRAAPEI-93
INSpmbTAHRAAPEI-206	INSofmTCGRAAPEI-30	INSofmTCMRAAPEI-37
RINSinlTBYRAAPEI-43	RINSinlTAERACPEI-57	RINSinlTBWRAAPEI-37
RINSinlTDARAAPEI-71	RINSinlTCRRAAPEI-37	RINSwvkTAERAAPEI-22
RINSwvkTAHRAAPEI-22	RINSjamTABRADPEI-15	ANIsrmTAAURAAPEI-222
WHANIsrmTMAXRAAPEI-74	INSnfrTAQRAAPEI-37	RINSinlTBXRAAPEI-41
INShauTADRAAPEI-95	WHANIsrmTMBXRAAPEI-30	
WHANIsrmTMALRAAPEI-22	WHANIsrmTMDERAAPEI-115	

Table 8: BLAST Statistics for the Best Matches of the False Positives Versus the Fallacious Reference Sequences.

Given are standard statistics for assessing the reliability of a BLAST match.

Parameter	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Length	5	32	61	76	79	632
Identity [%]	15.15	29.58	33.33	34.80	37.93	100.00
E-value	0	0	0.000031	0.607660	0.325000	10
Bitscore	14.60	21.60	33.50	42.38	44.70	355.00

These were *e.g.* Poly(-ADP-ribose) glycohydrolases (PARG, see Fig. 31). The matches to PARG are often other well matching RNAs from other (transcriptomic) shotgun sequencing projects that have no other potential viral relation.

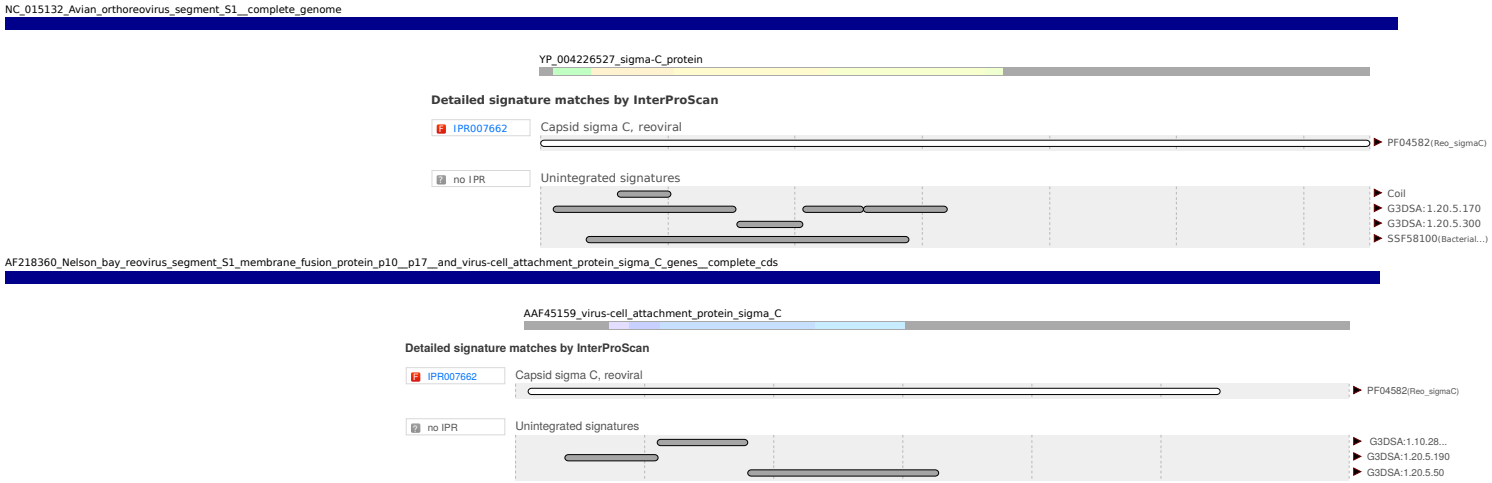

Figure 27: Coiled-coil Helices as a Source of False Positives.

The sigma c protein of *Avian orthoreovirus*, segment 1 (NC_015132) matched ORF 22 of contig s16_L_18_0_a_52_6_I_3823 from RINSinITCRRAAPEI-37 with 19% and *Nelson bay reovirus*, segment 1 (AF218360) with 22% identity. However, the contig was about double the size of the virus segment and ORF 22 covered the whole span of the sequence. It was matching the end of several myosin heavy chains at about 97% identity. The potential cause for this false positive match is supposed to be the coiled-coil helix domain that is similar to the coiled-coil helices in the matching region of the myosin heavy chain proteins. Although InterProScan did not report the coiled-coil helix domain for *Nelson bay reovirus*, it is expected to still contain a similar domain below the detection threshold because of the similarity to the respective protein of *Avian orthoreovirus*.

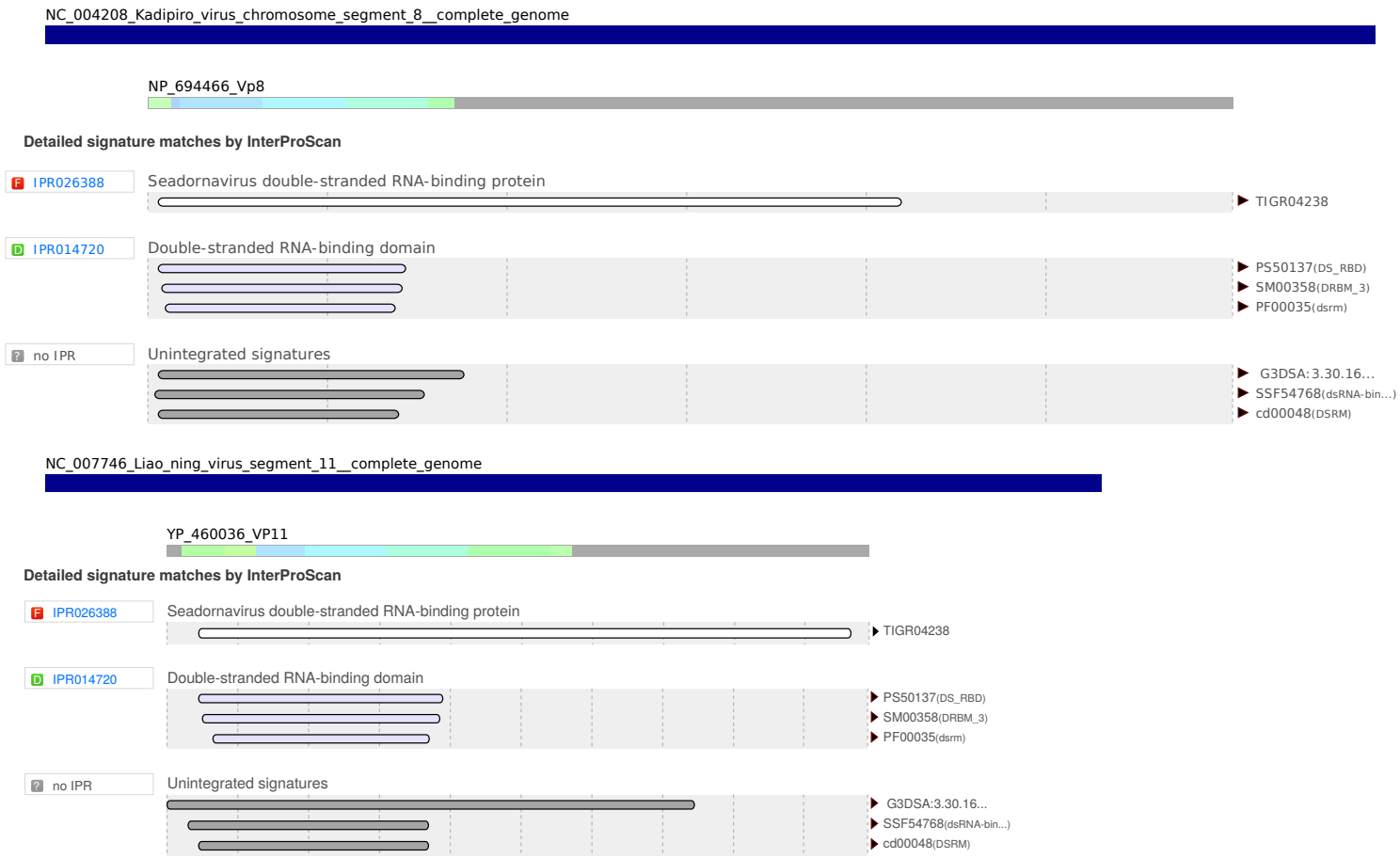

Figure 28: Double-stranded RNA Binding Motifs as a Source of False Positives.

Segment 8 of *Kadipiro virus* (NC_004208) and segment 11 of *Liao ning virus* (NC_007746) matched ORF 12 of contig s8354_L_28176_0_a_46_1_l_3087 from RINSinITCRRAAPEI-37 with about 31% identity. However, the contig was about three times the size of the virus segments. It was matching several interferon-inducible double-stranded RNA-dependent protein kinase activators at about 80% identity over nearly the whole length and the ORF was about as long as these references. The potential cause for this false positive match is supposed to be the double-stranded RNA binding motif that can be found in a variety of proteins. This similarity has already been pointed out when the viruses were published (Attoui *et al.*, 2000, 2006b).

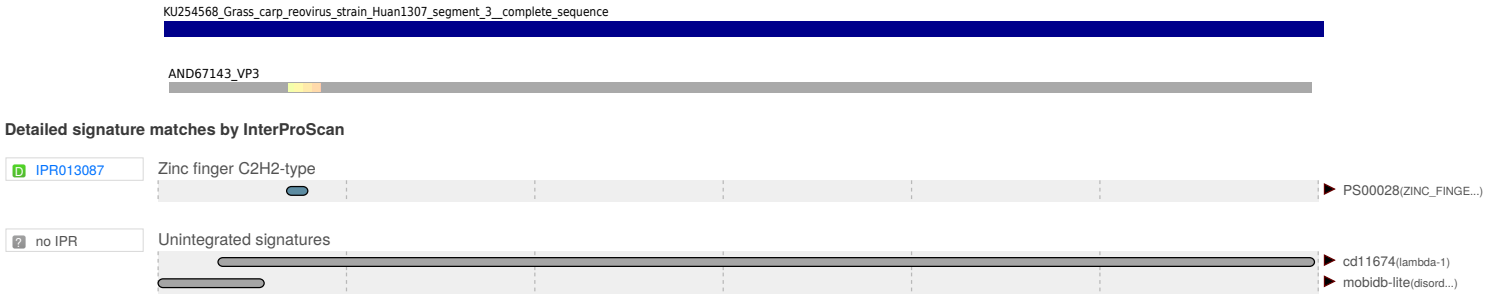

Figure 29: Zinc-Fingers as a Source of False Positives.

Segment 3 of *Grass carp reovirus* (KU254568) matches a small part of ORF 6 of contig s5707_L_14415_0_a_3_0_l_902 from INSnfrTBERAAPEI-19 with 33% identity. The contig was only a small fragment compared to *Grass carp reovirus*. It had several short ORFs whereas a long ongoing ORF would have been expected. A zinc finger domain has been detected by InterProScan that distinctly is in the area of the match from INSnfrTBERAAPEI-19.

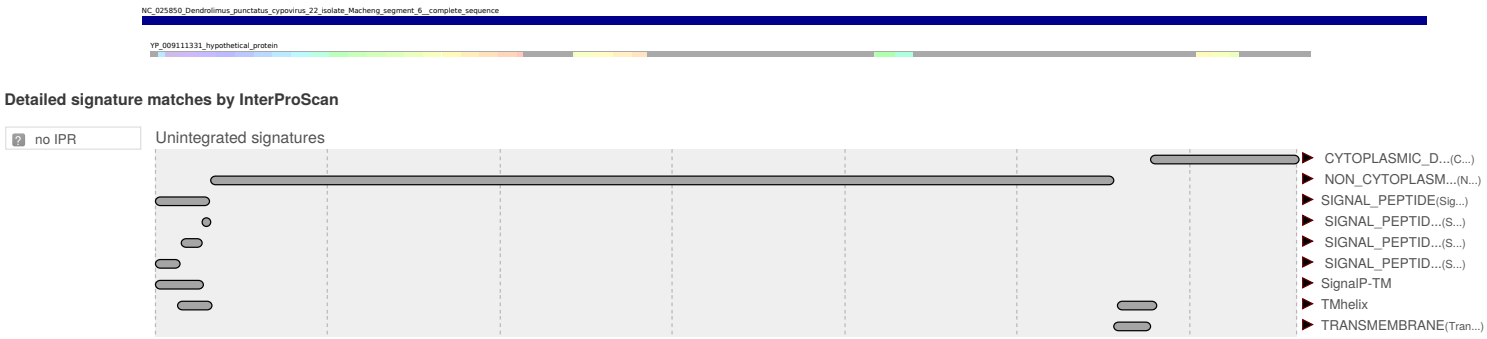

Figure 30: (Transmembrane-)Signalling Peptides as a Source of False Positives.

The start of segment 6 of *Dendrolimus punctatus cypovirus 22* (NC_025850) matches ORF 1 of contig s3450_L_4003_0_a_7_2_l_759 from INSytvTBYRAAPEI-22 with 26% identity. There are several predicted Transmembrane signaling protein domains predicted on segment 6 of *Dendrolimus punctatus cypovirus 22* and a large non-cytoplasmic signal peptide. In general, this segment yielded matches to many non or barely characterized proteins from other (transcriptomic) shotgun assemblies.

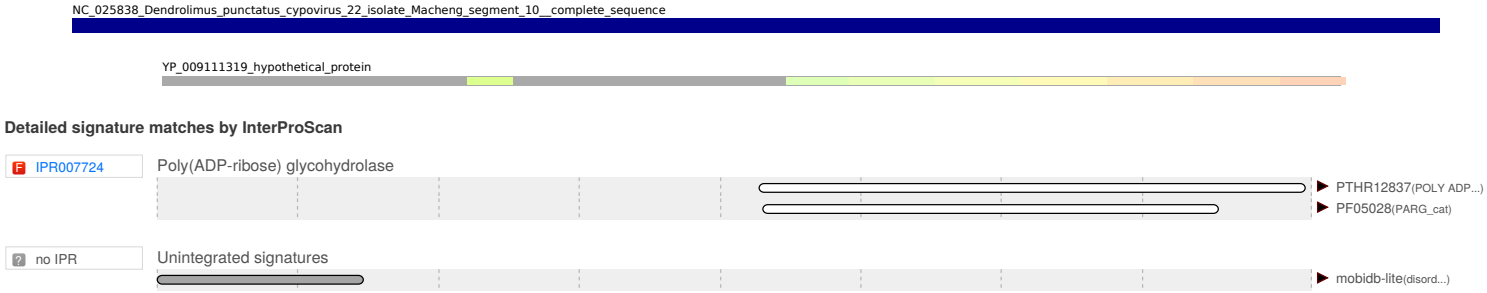

Figure 31: Poly-(ADP-ribose) Glycohydrolase (PARG) as a Source of False Positives.

Segment 10 of *Dendrolimus punctatus cypovirus 22* (NC_025838) matches the end of ORF 1 of contig C114088_a_6_0_l_1776 from RINSinlTCRRAAPEI-37 with about 30% identity. However, the contig length was similar to the viral segment as well as to the other references that were matching at about 50-65% identity along the whole ORF. These references were annotated as PARG but originated from other (transcriptomic) shotgun assemblies.

As for the distribution of potential viral sequences among the samples, it can be stated that suspicious sequences were detected across nearly all arthropod orders (see Table 9). However, some of the orders did not contain any true positives eventually.

2653 of all suspicious sequences were from non-blood-feeding hosts. Despite still containing the false positives, this is supporting the hypothesis that most of the already identified arthropod-associated viruses are in relation to blood-feeding species.

Table 9: Suspicious Sequences and True Positives by Order

Suspicious sequences were detected across nearly all orders. However the number of true positives is much lower and dropped to zero/NA for some orders.

Order	Suspicious Sequences	True Positives
Archaeognatha	66	3
Blattodea	56	4
Coleoptera	149	5
Collembola	36	9
Dermaptera	17	7
Diplura	53	NA
Diptera	221	30
Embioptera	16	NA
Grylloblattodea	10	NA
Hemiptera	194	71
Hymenoptera	980	129
Isoptera	18	NA
Lepidoptera	83	23
Mantodea	80	5
Megaloptera	19	NA
Neuroptera	238	18
Odonata	110	14
Orthoptera	28	1
Phasmatodea	39	6
Plecoptera	43	15
Psocodea	29	1
Raphidioptera	60	8
Siphonaptera	9	NA
Trichoptera	41	NA
Zygentoma	70	NA

3.2.2.1 Details of the True Positives

69 of the 1228 transcriptomes were containing potential viral sequences that were supposed to be true positives. In 35 of these transcriptomes, potentially full segments based on the genome mapping (see chapter 2.3.5) have been detected. On average, 3.8 full segments of a complete *Reovirus*-like set were contained in these samples (median: 3, see Fig. 32).

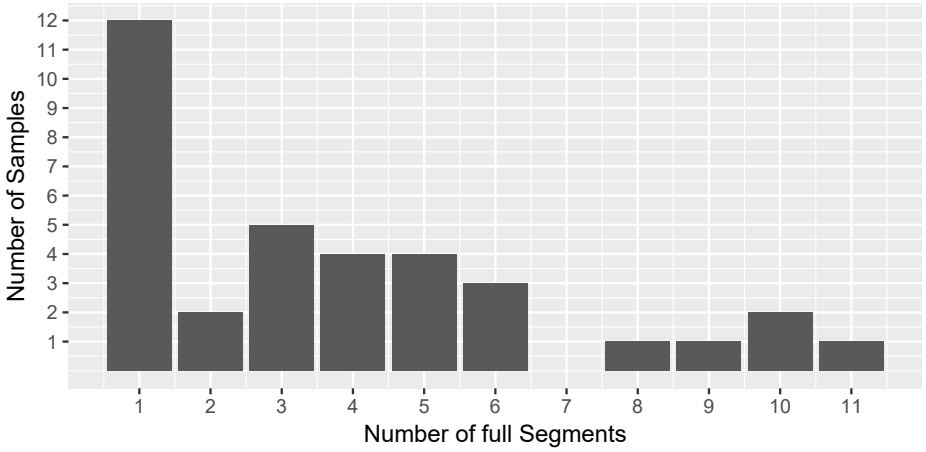

Figure 32: Number of Nearly Full Segments Found per Transcriptome.
In most transcriptomes, only one full segment could be found. However, samples containing complete sets of 9-11 segments were represented as well.

The following subsection contains detailed results of the true positives for some representative true positives. This includes meta-data of the sample, a table with general information about the true positives including the supposed closest known relative and estimation about the completeness of the genome. Additionally, the genome structure with predicted function of the identified ORFs is given according to chapter 2.3.5. In the illustrations of the genomes, the estimated nucleotide sequence is represented by a black bar and the hypothetical proteins by blue bars. The gray areas indicate the actual assembled parts from the transcriptomes. Tables and graphs for the other true positive transcriptomes can be found in the digital appendix (chapter 2). The result patterns in terms of assembly success and completeness of the genomes are similar to the selected representatives in this chapter.

INSfrgTACRAAPEI-21 (chapter 3.2.2.1.1) contained three full segments of a virus similar to *Cimodo virus* including the RdRp segment. True positives like from this transcriptome were easy to identify because the segments were fully assembled and the matching regions showed more than 30% BLAST identity and up to 21% GFAS identity.

INSjdsTBGRAAPEI-62 (chapter 3.2.2.1.2) contained four near full segments. Two fragments showed highest similarity to *Southern rice black-streaked dwarf virus* where the other segments were more similar to other viruses. In general, the sequences had below 30% BLAST similarity and since the potential closest relatives were different, this could

either be a case of very high divergence, an occurred re-assortment or a combination of both. Additionally, the GFAS identity is at about 7-8%. These sequences are probably at the edge of detectable yet verifiable distance.

INSytvTAERAAPEI-14 (chapter 3.2.2.1.3) contained only small fragments of four different segments. Three of these fragments were related to the RdRp of *Rice ragged stunt virus* and were matching at three different regions of the same segment (see Fig. 35). Two other fragments could also be assigned to other segments of *Rice ragged stunt virus* and another fragment to *Hubei reo-like virus 6*. In these cases the genome estimation showed that large proportions of the sequences are missing (see Fig. 35). The BLAST identity of the matching regions ranged from 22% to 35% and GFAS identity from 12-21%.

INSytvTBTRAAPEI-75 (chapter 3.2.2.1.4) contained a full segment with an RdRp similar to *Hubei reo-like virus 14* and a segment similar to segment 6 of *Dendrolimus punctatus cypovirus 22*. The matches to *Dendrolimus punctatus cypovirus 22* are difficult to assess because its segment 6 is a likely fallacious sequence as stated in chapter 3.2.2 (see Fig. 30). C76466_a_12_0_l_2779 additionally shows similarity to a hypothetical protein from several whole genome shotgun sequencing contigs with no other annotated ORFs or functions (*e.g. Habropoda laboriosa*). Since these hypothetical proteins have no other known functions but were detected by TRAVIS, a potential viral origin cannot be completely excluded. However, the BLAST identity ranged from 29-54% where GFAS was 16-18%.

INSytvTCBRAAPEI-33 (chapter 3.2.2.1.5) contained fragments of several segments similar to *Kadipiro virus* including about half of the RdRp segment. Despite *Kadipiro virus* is a potential fallacious reference, these segments are considered to be true positives since multiple different segments have been identified. In total, 11 segments could be at least partially detected, a number typical for a whole genome of a *Reovirus*. Additionally, the BLAST identity ranges from 22-46% and GFAS identity from 10-19%. Due to the many small fragments it might be speculated that sequencing occurred at the time of a declining infection or the RNA in the sample generally already started to decay. The median length of the contigs per transcriptome is 852.4 bp, the upper quartile 1051.9 bp and the maximum 1904.2 bp. So this sample with about 1221.3 bp per contig on average has generally larger contigs than most of the other transcriptomes. This leads to the assumption that the short lengths of the obtained potential viral sequences is more likely due to a declining infection than the overall degradation of RNA within the sample.

INShkeTATRAAPEI-56 (chapter 3.2.2.1.6) contained a near full genome of a virus similar to *Dendrolimus punctatus cypovirus* (Zhao *et al.*, 2003a,b) with partially over 90% BLAST and GFAS identity. *Dendrolimus punctatus* is a moth belonging to Lasiocampidae and INShkeTATRAAPEI-56 is the transcriptome of *Bicyclus anynana*, a butterfly from the family Nymphalidae. Since both families belong to the order of Lepidoptera, it can be speculated

that these two viruses have co-evolved. However, 14 different segments have been predicted based on the results of TRAVIS, more than the other known Reoviridae.

INSfrgTBCRAAPEI-57 (chapter 3.2.2.1.7) contained nearly the full genome of eleven segments of *Nilarpavata lugens reovirus* (NLRV; Nakashima *et al.*, 2018). The identified ORFs share an amino acid identity of mostly over 97% for BLAST as well as for GFAS. This virus is a known plant pathogen transmitted by *Nilarpavata lugens*, the same species as the scanned transcriptome originates from. It is remarkable that despite the usual high mutation rate for viruses, the obtained sequences show such a high similarity. Since the whole genome of *Nilarpavata lugens reovirus* was in the initial search database, it was easily retrievable with all used search tools. Sequence 6 is a good example for the well working algorithm of genome estimation where two fragments of a potential relative could be joined (see Fig. 39).

INSpmbTABRAAPEI-227 (chapter 3.2.2.1.8) contained several full sequences highly identical to *Diaphorina citri reovirus* (Nouri *et al.*, 2015) with over 98% BLAST and GFAS identity. The transcriptome originates as well from the same species, *Diaphorina citri*. In contrast to *Nilarpavata lugens reovirus* found in INSfrgTBCRAAPEI-57, *Diaphorina citri reovirus* was not in reference library for the initial searches but it was still possible to retrieve six full and one partial segments of ten that are known. Additionally, other questionable sequences of potential viral origin have been identified. They are mostly related to known hypothetical proteins of *Diaphorina citri*.

INSqiqTALRAAPEI-30 (chapter 3.2.2.1.9) is interesting because it contained a fragmentary RdRp that is Mononegavirales-like. However, other segments that might be related to Chuviridae have also been detected. All identified viruses except *Liao ning virus* are thought to be distantly related to Mononegavirales (Tokarz *et al.*, 2014; Li *et al.*, 2015; Shi *et al.*, 2016a). Classical Mononegavirales are single stranded RNA viruses and Chuviridae are already known to have two segments. Sequence 1 and 2 support evidence for Chuviridae and Sequence 3 is likely to be related to *Liao ning virus*. With BLAST identities ranging from 20-34% and GFAS identity from 8-19%, the potential viral sequences are distant to the references. However a common origin of all RNA-viruses has already been speculated (Koonin *et al.*, 2015). In this hypothesis, Reoviridae originated after Eukaryogenesis and Mononegavirales have evolved more recently. The findings in INSqiqTALRAAPEI-30 might thus support this hypothesis.

INSofmTBWRAAPEI-126 (chapter 3.2.2.1.10) contained a full RdRp similar to the one of *Dill cryptic virus* which belongs to Partitiviridae. The BLAST identity to *Dill cryptic virus* is 59% and 32% to *Rotavirus A*. Again, this is evidence for the relationship of different RNA viruses as stated by Koonin *et al.*, 2015.

3.2.2.1.1 INSfrgTACRAAPEI-21

Table 10: Sample Information of INSfrgTACRAAPEI-21.

Filename	120215_I277_FCD0KP1ACXX_L1_INSfrgTACRAAPEI-21.free.fas
Assembly ID	INSfrgTACRAAPEI-21
Order	Hymenoptera
Order details	NA
Family	Eulophidae
Family details	NA
Species	*Diglyphus isaea*
Number of specimen	ca 200
Stage	adult
Sample location	Lab culture of unknown geographical origin
Sample date	12-May-2011
Blood-feeding	no
Suspicous sequences	20

Table 11: Suspicious Sequences in INSfrgTACRAAPEI-21.

3 of 20 sequences were true positives and 17 sequences were false positives similar to the false positives listed in 3.2.2.

Sequence ID	ORF	Match	Identity	Completeness
s2486_L_3986_2_a_50_7_I_2082	ORF_007	segment 6, *Cimodo virus* (KF880765)	30%	full
s2487_L_3986_3_a_42_3_I_4091	ORF_011	**RdRp**, *Cimodo virus* (KF880772)	41%	full
s2883_L_4857_0_a_52_0_I_3600	ORF_001	segment 2, *Cimodo virus* (NC_024916)	34%	full

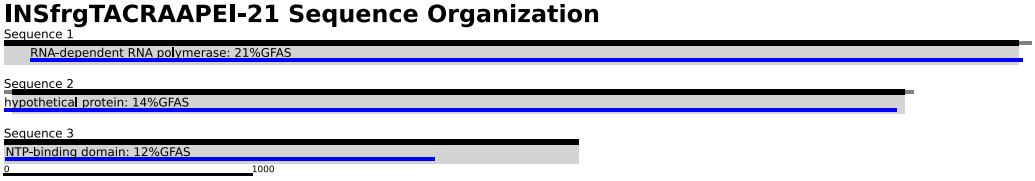

Figure 33: Sequence Organization of INSfrgTACRAAPEI-21.

3.2.2.1.2 INSjdsTBGRAAPEI-62

Table 12: Sample Information of INSjdsTBGRAAPEI-62.

Filename	120215_I277_FCD0KP1ACXX_L8_INSjdsTBGRAAPEI-62.free.fas
Assembly ID	INSjdsTBGRAAPEI-62
Order	Zygentoma
Order details	NA
Family	Lepismatidae
Family details	NA
Species	*Ctenolepisma longicaudata*
Number of specimen	8
Stage	adult
Sample location	Germany, North Rhine-Westphalia, Bonn
Sample date	2011
Blood-feeding	no
Suspicous sequences	25

Table 13: Suspicious Sequences in INSjdsTBGRAAPEI-62.

5 of 25 sequences were true positives and 20 sequences were false positives similar to the false positives listed in 3.2.2.

Sequence ID	ORF	Match	Identity	Completeness
C169885_a_3_0_I_363	ORF_001	**RdRp**, *Southern rice black-streaked dwarf virus* (NC_014714)	27%	partial (end)
C225767_a_61_0_I_1979	ORF_003	1. segment 6, *Aedes pseudoscutellaris reovirus* (NC_007671) 2. segment 5, *Inachis io cypovirus 2* (NC_023488)	24% 20%	full full
C228749_a_27_0_I_3157	ORF_013	**RdRp**, *Southern rice black-streaked dwarf virus* (NC_014714)	26%	partial (start-mid)
C228891_a_36_0_I_3316	ORF_012	segment 4, *Mal de Rio Cuarto virus* (NC_008729)	21%	full
C229267_a_61_0_I_4098	ORF_013	segment 2, *Fiji disease virus* (NC_007154)	17%	full

INSjdsTBGRAAPEI-62 Sequence Organization

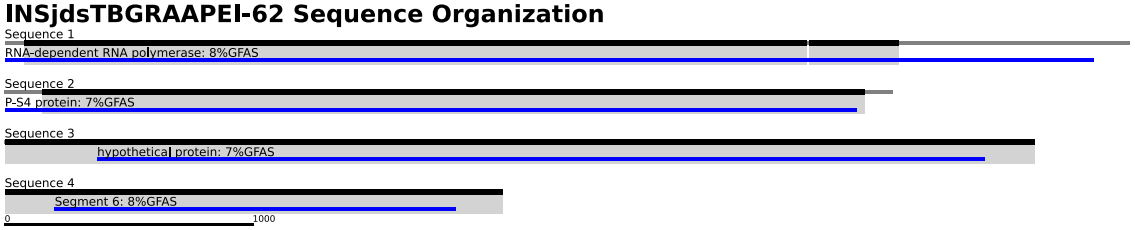

Figure 34: Sequence Organization of INSjdsTBGRAAPEI-62.

3.2.2.1.3 INSytvTAERAAPEI-14

Table 14: Sample Information of INSytvTAERAAPEI-14.

Filename	120429_I266_FCC0HG0ACXX_L7_INSytvTAERAAPEI-14.free.fas
Assembly ID	INSytvTAERAAPEI-14
Order	Hemiptera
Order details	Sternorrhyncha
Family	Psyllidae
Family details	NA
Species	*Glycaspis brimblecombei*
Number of specimen	ca. 20
Stage	missing
Sample location	Australia South Australia Adelaide River Torrens
Sample date	20-Feb-2012
Blood-feeding	no
Suspicous sequences	14

Table 15: Suspicious Sequences in INSytvTAERAAPEI-14.

6 of 14 sequences were true positives and 8 sequences were false positives similar to the false positives listed in 3.2.2.

Sequence ID	ORF	Match	Identity	Completeness
C230333_a_4_0_l_242	ORF_001	segment 2, *Rice ragged stunt virus* (NC_003750)	30%	partial (mid)
C329411_a_5_0_l_478	ORF_003	**RdRp**, *Rice ragged stunt virus* (NC_003771)	35%	partial (mid)
C338577_a_7_0_l_539	ORF_001	1. segment 8, *Raspberry latent virus* (NC_014605) 2. segment 7, *Rice ragged stunt virus* (NC_003770)	35% 27%	partial (end) partial (end)
C345732_a_3_0_l_606	ORF_003	**RdRp**, *Rice ragged stunt virus* (NC_003771)	27%	partial (start)
C352171_a_3_0_l_695	ORF_001	1. hypothetical protein, *Hubei reo-like virus 6* (KX884718) 2. segment 4, *Lymantria dispar cypovirus 14* (AF389455)	30% 22%	partial (end) partial (end)
C369021_a_4_0_l_1374	ORF_004	**RdRp**, *Rice ragged stunt virus* (NC_003771)	23%	partial (end)

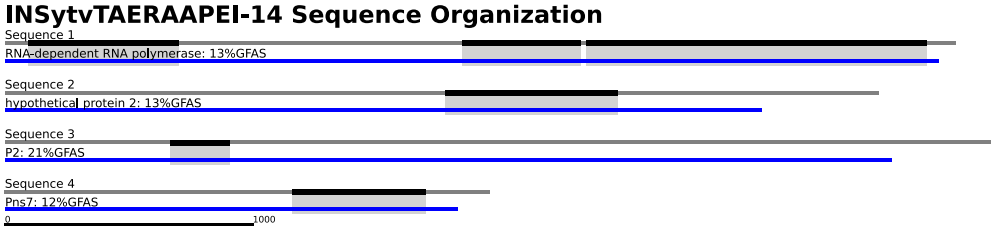

Figure 35: Sequence Organization of INSytvTAERAAPEI-14.

3.2.2.1.4 INSytvTBTRAAPEI-75

Table 16: Sample Information of INSytvTBTRAAPEI-75.

Filename	120521_I249_FCC0U4RACXX_L8_INSytvTBTRAAPEI-75.free.fas
Assembly ID	INSytvTBTRAAPEI-75
Order	Hymenoptera
Order details	NA
Family	Pompilidae
Family details	NA
Species	*Heterodontonyx sp*
Number of specimen	2
Stage	adult
Sample location	Australia, Western Australia, 118 km N Esperance
Sample date	07-Nov-2011
Blood-feeding	no
Suspicous sequences	16

Table 17: Suspicious Sequences in INSytvTBTRAAPEI-75.

2 of 16 sequences were true positives, 3 were questionable and 11 sequences were false positives similar to the false positives listed in 3.2.2. Questionable sequences are marked with (?).

Sequence ID	ORF	Match	Identity	Completeness
C76466_a_12_0_l_2779	ORF_003	1. hypothetical protein, *Habropoda laboriosa* (LHQN01027684) 2. hypothetical protein, *Dendrolimus punctatus cypovirus 22* (NC_025850)	54% 29%	full full
C79130_a_22_0_l_4026	ORF_005	**RdRp**, *Hubei reo-like virus 14* (KX884607)	38%	full
(?) s5118_L_11025_0_a_29_6_l_6233	ORF_023	1. hypothetical protein, *Cerapachys biroi* (KK108206) 2. hypothetical protein, *Dendrolimus punctatus cypovirus 22* (NC_025850)	52% 31%	full full
(?) s5242_L_11500_0_a_15_9_l_4053	ORF_003	1. hypothetical protein, *Cerapachys biroi* (KK108206) 2. hypothetical protein, *Dendrolimus punctatus cypovirus 22* (NC_025850)	50% 31%	full full
(?) s5243_L_11500_1_a_9_6_l_3723	ORF_002	1. hypothetical protein, *Cerapachys biroi* (KK108206) 2. hypothetical protein, *Dendrolimus punctatus cypovirus 22* (NC_025850)	50% 31%	full full

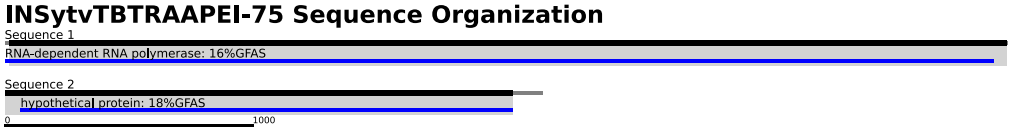

Figure 36: Sequence Organization of INSytvTBTRAAPEI-75.

3.2.2.1.5 INSytvTCBRAAPEI-33

Table 18: Sample Information of INSytvTCBRAAPEI-33.

Filename	120521_I249_FCC0U4RACXX_L8_INSytvTCBRAAPEI-33.free.fas
Assembly ID	INSytvTCBRAAPEI-33
Order	Hymenoptera
Order details	NA
Family	Vespidae
Family details	NA
Species	*Katamenes arbustorum*
Number of specimen	2
Stage	adult
Sample location	Italy, Valle de Cogne, Lillaz
Sample date	16-Jul-2011
Blood-feeding	no
Suspicous sequences	24

Table 19: Suspicious Sequences in INSytvTCBRAAPEI-33.

15 of 24 sequences were true positives, 2 questionable and 7 sequences were false positives similar to the false positives listed in 3.2.2. Questionable sequences are marked with (?).

Sequence ID	ORF	Match	Identity	Completeness
(?) C100890_a_12_0_I_2006	ORF_001	hypothetical protein, *Dendrolimus punctatus cypovirus 22* (NC_025838)	36%	full
C45645_a_3_0_I_254	ORF_001	**RdRp**, *Kadipiro virus* (NC_004210)	42%	partial (end)
C45671_a_8_0_I_254	ORF_001	segment 2, *Liao ning virus* (NC_007737)	50%	partial (end)
C55655_a_3_0_I_326	ORF_001	segment 10, *Kadipiro virus* (NC_004206)	31%	partial (start-mid)
C58033_a_12_0_I_346	ORF_001	segment 7, *Kadipiro virus* (NC_004209)	28%	partial (end)
C63000_a_4_0_I_397	ORF_001	segment 12, *Kadipiro virus* (NC_004199)	34%	partial (mid-end)
C63732_a_3_0_I_405	ORF_001	**RdRp**, *Kadipiro virus* (NC_004210)	46%	partial (mid)
C67095_a_3_0_I_447	ORF_003	segment 3, *Kadipiro virus* (NC_004213)	45%	partial (end)
C69827_a_3_0_I_484	ORF_002	segment 4, *Kadipiro virus* (NC_004214)	22%	partial (mid)
C83036_a_4_0_I_756	ORF_001	**RdRp**, *Kadipiro virus* (NC_004210)	42%	partial (mid)
(?) C84632_a_21_0_I_808	ORF_003	segment 11, *Liao ning virus* (NC_007746)	22%	partial (start)
C89564_a_17_0_I_1010	ORF_003	segment 9, *Kadipiro virus* (NC_0042076)	29%	full
C92606_a_7_0_I_1176	ORF_003	segment 2, *Kadipiro virus* (NC_004212)	27%	partial (start)
C93256_a_5_0_I_1220	ORF_001	segment 6, *Kadipiro virus* (NC_004216)	29%	partial (start-mid)
C93816_a_4_0_I_1263	ORF_004	segment 5, *Kadipiro virus* (NC_004215)	32%	partial (mid)
C97782_a_6_0_I_1600	ORF_005	**RdRp**, *Kadipiro virus* (NC_004210)	34%	partial (start)

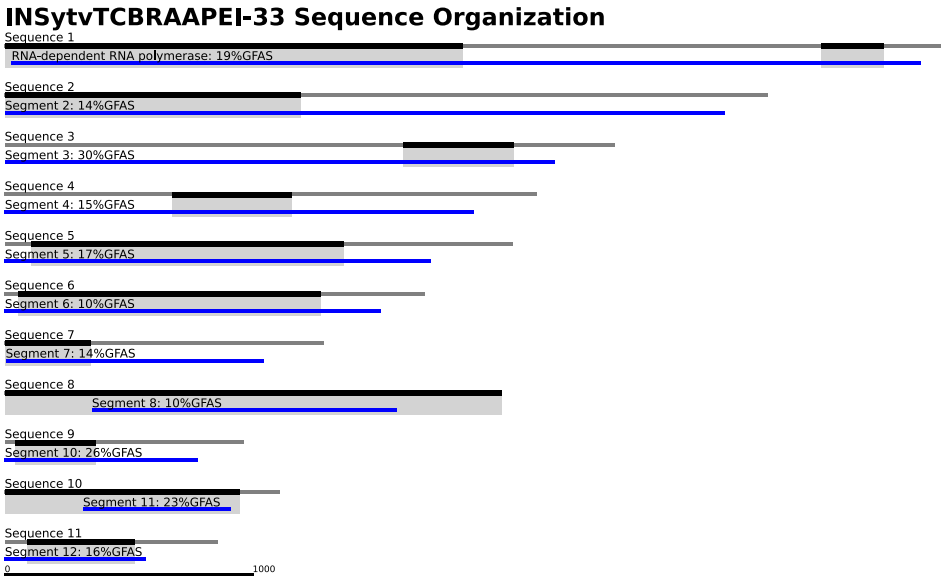

Figure 37: Sequence Organization of INSytvTCBRAAPEI-33.

3.2.2.1.6 INShkeTATRAAPEI-56

Table 20: Sample Information of INShkeTATRAAPEI-56.

Filename	120816_I269_FCC10KYACXX_L8_INShkeTATRAAPEI-56.free.fas
Assembly ID	INShkeTATRAAPEI-56
Order	Lepidoptera
Order details	NA
Family	Nymphalidae
Family details	NA
Species	*Bicyclus anynana*
Number of specimen	2
Stage	NA
Sample location	Germany Lab culture with Samples originating from Malawi, Nkhata Bay
Sample date	14-May-2012
Blood-feeding	no
Suspicous sequences	35

INShkeTATRAAPEI-56 Sequence Organization

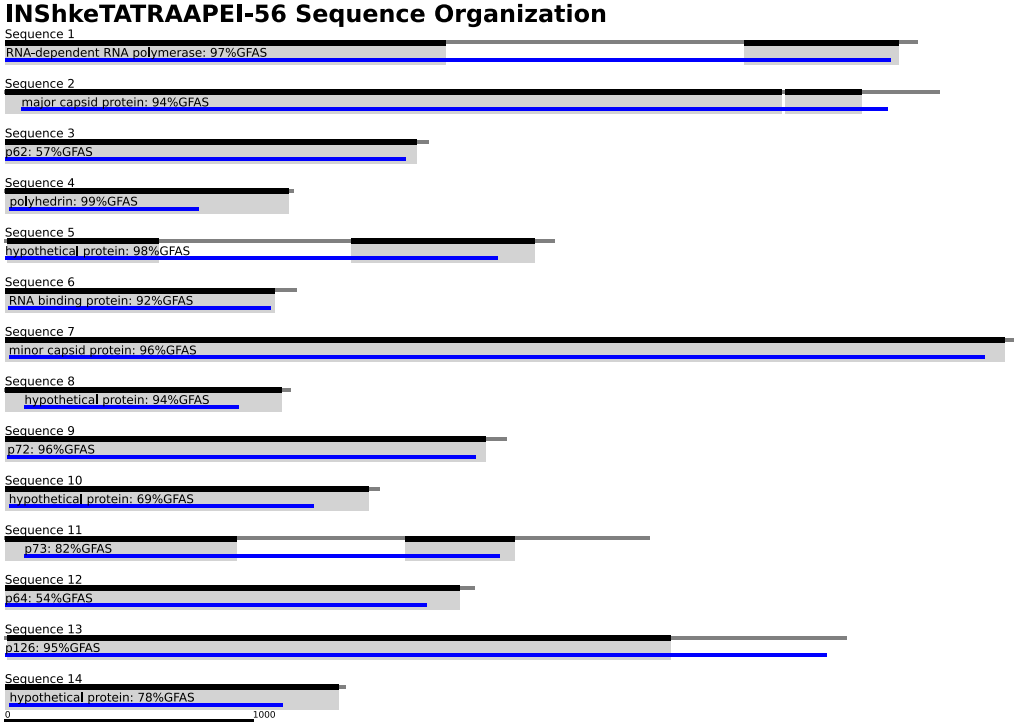

Figure 38: Sequence Organization of INShkeTATRAAPEI-56.

Table 21: Suspicious Sequences in INShkeTATRAAPEI-56.

21 of 35 sequences were true positives and 14 sequences were false positives similar to the false positives listed in 3.2.2.

Sequence ID	ORF	Match	Identity	Completeness
C160677_a_4_0_l_308	ORF_002	major capsid protein, *Dendrolimus punctatus cypovirus 22* (NC_025846)	94%	partial (end)
C183635_a_4_0_l_443	ORF_002	segment 5, *Dendrolimus punctatus cypovirus 22* (NC_025849)	95%	partial (end)
C195871_a_9_0_l_573	ORF_001	segment 5, *Dendrolimus punctatus cypovirus 22* (NC_025849)	76%	partial (start)
C198731_a_3_0_l_611	ORF_002	segment 6, *Dendrolimus punctatus cypovirus 22* (NC_025850)	99%	partial (start)
C199032_a_12_0_l_616	ORF_005	segment 10, *Dendrolimus punctatus cypovirus 22* (NC_025838)	38%	partial (mid-end)
C199445_a_3_0_l_623	ORF_002	**RdRp**, *Dendrolimus punctatus cypovirus 22* (NC_025847)	98%	partial (end)
C200405_a_3_0_l_639	ORF_002	segment 5, *Dendrolimus punctatus cypovirus 22* (NC_025849)	90%	partial (mid)
C205512_a_3_0_l_739	ORF_003	segment 6, *Dendrolimus punctatus cypovirus 22* (NC_025850)	99%	partial (end)
C215434_a_32_0_l_1086	ORF_005	segment 12, *Dendrolimus punctatus cypovirus 22* (NC_025840)	92%	full
C215988_a_27_0_l_1117	ORF_001	segment 14, *Dendrolimus punctatus cypovirus 22* (NC_025842)	94%	full
C216436_a_61_0_l_1144	ORF_002	segment 13, *Dendrolimus punctatus cypovirus 22* (NC_025841)	99%	full
C219116_a_51_0_l_1322	ORF_006	segment 11, *Dendrolimus punctatus cypovirus 22* (NC_025839)	78%	full
C219998_a_19_0_l_1398	ORF_001	segment 10, *Dendrolimus punctatus cypovirus 22* (NC_025838)	86%	full
C222412_a_16_0_l_1659	ORF_002	segment 9, *Dendrolimus punctatus cypovirus 22* (NC_025853)	92%	full
C223206_a_4_0_l_1775	ORF_004	**RdRp**, *Dendrolimus punctatus cypovirus 22* (NC_025847)	98%	partial (start-mid))
C223558_a_30_0_l_1835	ORF_003	segment 8, *Dendrolimus punctatus cypovirus 22* (NC_025852)	77%	full
C224058_a_10_0_l_1936	ORF_014	segment 7, *Dendrolimus punctatus cypovirus 22* (NC_025851)	96%	full
C226066_a_4_0_l_2676	ORF_001	segment 4, *Dendrolimus punctatus cypovirus 22* (NC_025848)	95%	partial (start-mid)
C226586_a_8_0_l_3105	ORF_006	major capsid protein, *Dendrolimus punctatus cypovirus 22* (NC_025846)	94%	partial (start-mid)
C227042_a_15_0_l_4032	ORF_003	minor capsid protein, *Dendrolimus punctatus cypovirus 22* (NC_025845)	96%	full
s1837_L_1284_0_a_10_6_l_1679	ORF_006	segment 10, *Dendrolimus punctatus cypovirus 22* (NC_0258385)	31%	full

3.2.2.1.7 INSfrgTBCRAAPEI-57

Table 22: Sample Information of INSfrgTBCRAAPEI-57.

Filename	120215_I277_FCD0KP1ACXX_L1_INSfrgTBCRAAPEI-57.free.fas
Assembly ID	INSfrgTBCRAAPEI-57
Order	Hemiptera
Order details	Auchenorrhyncha, Fulgoromorpha
Family	Delphacidae
Family details	NA
Species	*Nilaparvata lugens*
Number of specimen	ca 30
Stage	NA
Sample location	Germany lab culture with Samples from a private breeder Ralf Nauen, Bayer CropScience, Monheim, Germany
Sample date	October 2011
Blood-feeding	no
Suspicous sequences	29

Table 23: Suspicious Sequences in INSfrgTBCRAAPEI-57.

13 of 29 sequences were true positives, one questionable and 15 sequences were false positives similar to the false positives listed in 3.2.2. Questionable sequences are marked with (?).

Sequence ID	ORF	Match	Identity	Completeness
C136646_a_12_0_I_409	ORF_001	segment 6, *Nilaparvata lugens reovirus* (NC_003659)	99%	partial (start)
(?) C172497_a_34_0_I_1212	ORF_003	segment 11, *Liao ning virus* (NC_007746)	21%	full
C174953_a_23_0_I_1381	ORF_003	segment 10, *Nilaparvata lugens reovirus* (NC_003652)	99%	full
C175507_a_9_0_I_1422	ORF_004	segment 4, *Nilaparvata lugens reovirus* (NC_003657)	98%	partial (start)
C176757_a_22_0_I_1539	ORF_007	segment 9, *Nilaparvata lugens reovirus* (NC_003661)	97%	full
C176757_a_22_0_I_1539	ORF_002	segment 9, *Nilaparvata lugens reovirus* (NC_003661)	99%	full
C179933_a_11_0_I_1913	ORF_002	segment 7, *Nilaparvata lugens reovirus* (NC_003660)	99%	full
C180291_a_5_0_I_1971	ORF_007	segment 4, *Nilaparvata lugens reovirus* (NC_003657)	99%	partial (mid-end)
C182269_a_17_0_I_2426	ORF_001	segment 6, *Nilaparvata lugens reovirus* (NC_003659)	98%	full
C183817_a_8_0_I_3194	ORF_005	segment 3, *Nilaparvata lugens reovirus* (NC_003656)	99%	full
C184525_a_7_0_I_4357	ORF_013	**RdRp**, *Nilaparvata lugens reovirus* (NC_003654)	99%	full
s11081_L_33395_0_a_24_4_I_1768	ORF_001	segment 8, *Nilaparvata lugens reovirus* (NC_003653)	100%	full
s11916_L_40961_0_a_19_1_I_3705	ORF_010	segment 2, *Nilaparvata lugens reovirus* (NC_003655)	98%	full
s7224_L_11880_0_a_68_0_I_3428	ORF_007	segment 5, *Nilaparvata lugens reovirus* (NC_003658)	94%	full

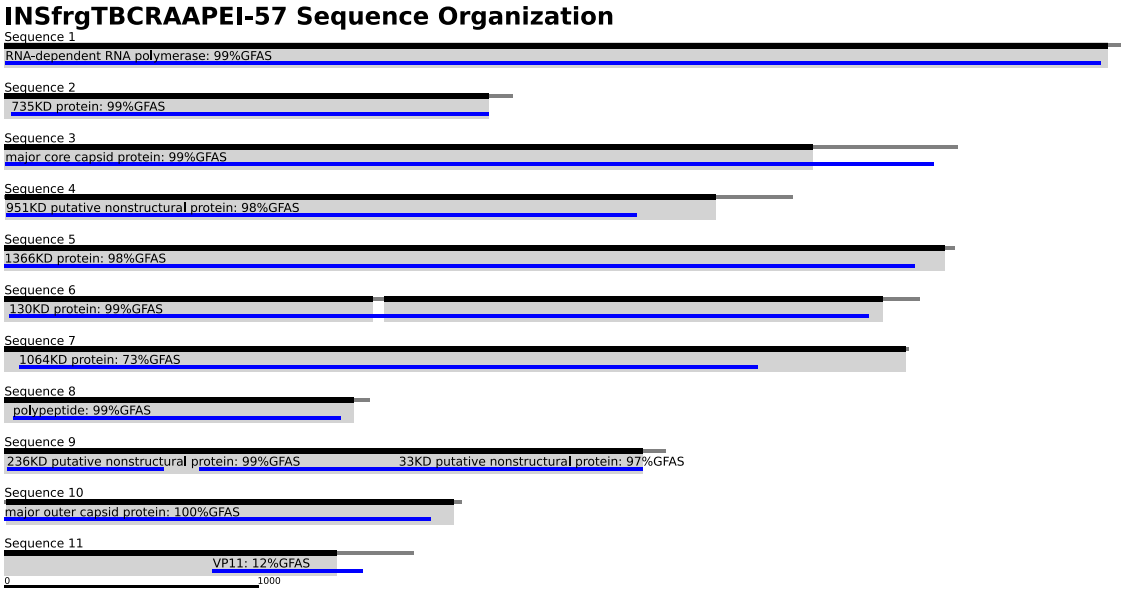

Figure 39: Sequence Organization of INSfrgTBCRAAPEI-57.

3.2.2.1.8 INSpmbTABRAAPEI-227

Table 24: Sample Information of INSpmbTABRAAPEI-227.

Filename	130901_I238_FCC2BVYACXX_L8_INSpmbTABRAAPEI-227.free.fas
Assembly ID	INSpmbTABRAAPEI-227
Order	Hemiptera
Order details	Sternorrhyncha
Family	Psyllidae
Family details	NA
Species	*Diaphorina citri*
Number of specimen	1
Stage	adult
Sample location	USA, lab culture
Sample date	Oct-2011
Blood-feeding	no
Suspicous sequences	13

Table 25: Suspicious Sequences in INSpmbTABRAAPEI-227.

8 of 13 sequences were true positives, 4 were questionable and 1 sequence was false positive similar to the false positives listed in 3.2.2. Questionable sequences are marked with (?).

Sequence ID	ORF	Match	Identity	Completeness
C195920_a_4_0_l_624	ORF_003	glycoprotein, *Hubei chuvirus-like virus 1* (NC_033328)	27%	partial (end)
(?) C204193_a_11_0_l_851	ORF_005	RISC-loading complex, *Diaphorina citri reovirus* (XM_008483089)	100%	partial (end)
C210209_a_50_0_l_1131	ORF_003	nonstructural polypeptide, *Diaphorina citri reovirus* (KT698833)	98%	full
(?) C212087_a_23_0_l_1259	ORF_008	1. sigma 1, *Mammalian Orthoreovirus* (JQ412761) 2. cingulin-like protein , *Diaphorina citri* (XM_008487952)	19% 99%	full full
C215393_a_26_0_l_1642	ORF_001	major outer capsid protein, *Diaphorina citri reovirus* (KT698831)	98%	full
C216069_a_24_0_l_1779	ORF_001	minor core structural protein, *Diaphorina citri reovirus* (KT698836)	98%	full
C217395_a_40_0_l_3251	ORF_001	inner capsid protein, *Diaphorina citri reovirus* (KT698835)	98%	full
C217401_a_36_0_l_3447	ORF_010	B-spike protein, D*iaphorina citri reovirus* (KT698832)	96%	full
C217415_a_47_0_l_3787	ORF_001	major core capsid protein, *Diaphorina citri reovirus* (KT698834)	99%	full
C217419_a_50_0_l_4334	ORF_006	**RdRp**, *Diaphorina citri reovirus* (KT698830)	99%	full
(?) s4262_L_5267_0_a_36_2_l_1100	ORF_008	1. sigma 1, *Mammalian Orthoreovirus* (JQ412761) 2. WEB family protein , *Diaphorina citri* (XM_008487952)	32% 100%	full full
(?) s9042_L_16135_0_a_29_4_l_1349	ORF_004	1. VP2, *Morris orbivirus* (KX907619) 2. hypothetical protein , *Diaphorina citri* (XM_008487952)	32% 99%	full full

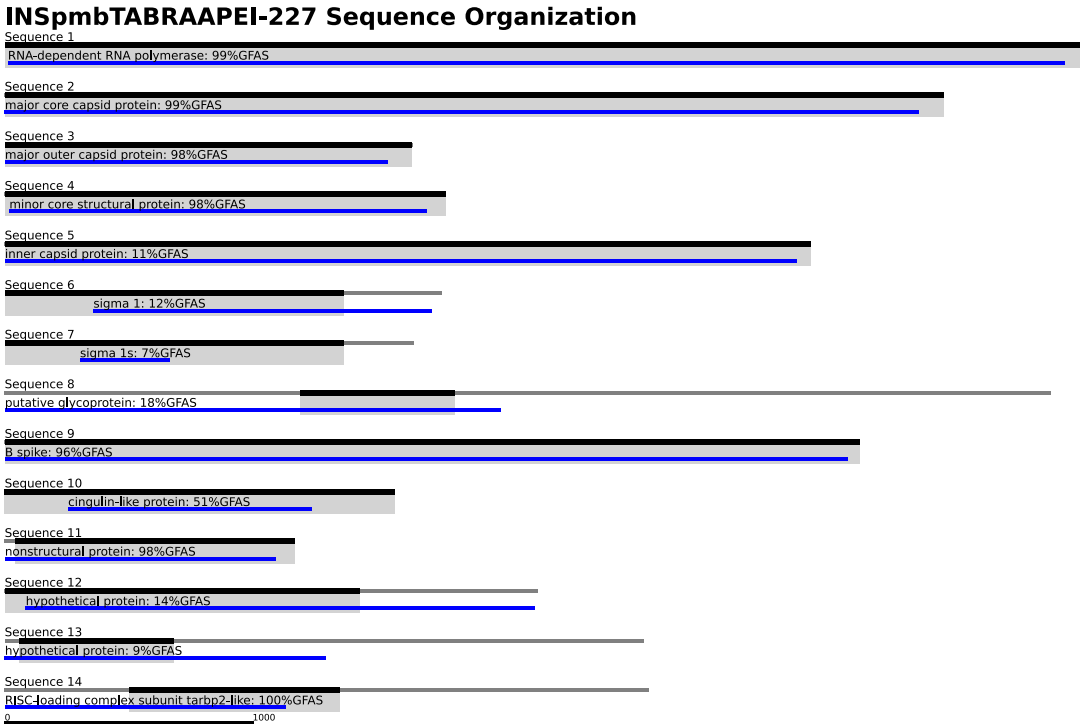

Figure 40: Sequence Organization of INSpmbTABRAAPEI-227.

3.2.2.1.9 INSqiqTALRAAPEI-30

Table 26: Sample Information of INSqiqTALRAAPEI-30.

Filename	130112_I269_FCC1M19ACXX_L2_INSqiqTALRAAPEI-30.free.fas
Assembly ID	INSqiqTALRAAPEI-30
Order	Dermaptera
Order details	NA
Family	Spongiphoridae
Family details	NA
Species	*Nesogaster amoenus*
Number of specimen	7
Stage	adult
Sample location	Malaysia, Selangor Ulu, Gombak Taman Rimba Komanwel
Sample date	04-Apr-2012
Blood-feeding	no
Suspicous sequences	17

Table 27: Suspicious Sequences in INSqiqTALRAAPEI-30.

2 of 17 sequences were true positives, 4 were questionable and 11 sequences were false positives similar to the false positives listed in 3.2.2. Questionable sequences are marked with (?).

Sequence ID	ORF	Match	Identity	Completeness
(?) C78089_a_27_0_l_563	ORF_001	segment 2, *Wuchang Cockraoch Virus 3* (NC_007746)	28%	partial (end)
(?) C86188_a_38_0_l_821	ORF_002	segment 11, *Liao ning virus* (NC_007746)	22%	full
C95883_a_13_0_l_2632	ORF_006	1. **RdRp**, *Deer tick mononegavirales-like virus* (KJ746903) 2. **RdRp**, *Hubei chuvirus-like virus 1* (NC_033327)	21% 20%	partial (end) partial (end)
(?) s2864_L_5034_0_a_14_4_l_1313	ORF_003	segment 11, *Liao ning virus* (NC_007746)	29%	full
(?) s2865_L_5034_1_a_13_4_l_1412	ORF_004	segment 11, *Liao ning virus* (NC_007746)	29%	full
s5742_L_20935_0_a_25_2_l_3158	ORF_002	glycoprotein, *Wuchang Cockraoch Virus 3* (KM817605)	34%	full
s5742_L_20935_0_a_25_2_l_3158	ORF_014	nucleoprotein, *Wuchang Cockraoch Virus 3* (KM817605)	29%	full

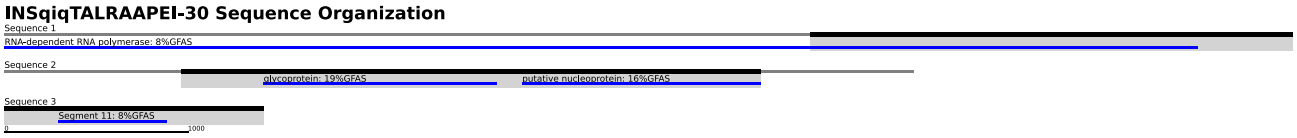

Figure 41: Sequence Organization of INSqiqTALRAAPEI-30.

3.2.2.1.10 INSofmTBWRAAPEI-126

Table 28: Sample Information of INSofmTBWRAAPEI-126.

Filename	130919_I247_FCC2V7VACXX_L2_INSofmTBWRAAPEI-126.free.fas
Assembly ID	INSofmTBWRAAPEI-126
Order	Blattodea
Order details	NA
Family	Ectobiidae
Family details	Pseudophyllodromiinae
Species	*Ellipsidion sp*
Number of specimen	3
Stage	nymph
Sample location	Australia, Queensland ,Brisbane, St Lucia
Sample date	09-Mar-2013
Blood-feeding	no
Suspicous sequences	10

Table 29: Suspicious Sequences in INSofmTBWRAAPEI-126.

1 of 10 sequences was true positive and 9 sequences were false positives similar to the false positives listed in 3.2.2.

Sequence ID	ORF	Match	Identity	Completeness
C397659_a_60_0_I_2000	ORF_014	1. **RdRp**, *Rotavirus A* (NC_011507)	32%	partial
		2. **RdRp**, *Dill cryptic virus* (NC_022614)	59%	full

INSofmTBWRAAPEI-126 Sequence Organization

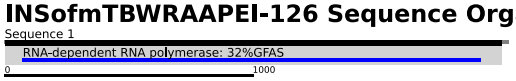

Figure 42: Sequence Organization of INSofmTBWRAAPEI-126.

3.2.3 Inference of Phylogeny

The overall bootstrap support showed a similar pattern for all reconstruction variations (see Fig. 43,Fig. 44, Fig. 45, Fig. 46, Fig. 47, Fig. 48). As intended, the total alignment length decreased with the incremental reduction of columns that have a certain percentage of gaps. However the bootstrap supports were more or less stable up to a certain gap trimming threshold, where the support decreased substantially. This threshold was at about 75-80% gap trimming irrespective of alignment variation or reconstruction method. In the case of FastME and PhyML, the used substitution models (Blosum62 and WAG) performed equally. When comparing the pure amino acid alignment phylogenies the support for the PhyML reconstruction was considerably higher with a median at about 75% where the NJ-variations were only at about 50%. On the hydrophobicity alignment, the bootstraps for PhyML and R were around 50% and for FastME about 30%. In context with the rate of bootstraps below the confidence level of 60% it can be assumed that neither of the alignment methods and the phylogenetic reconstruction algorithms were able to derive a stable, well supported phylogeny. However, it showed that the different methods show more or less consistent reconstruction success when dealing with similar datasets regardless of gaps up to a certain degree.

The topology similarities calculated with tqDist revealed that a large proportion of about 60% of the topologies were identical irrespective of alignment method, phylogenetic reconstruction algorithm and chosen substitution model. This holds true even for the collapsed trees with bootstrap support below 90% (see Fig. 49 and Fig. 50). Except for the R reconstruction based on the hydrophobicity alignment, all other phylogenetic reconstruction algorithm were able to produce identical topologies when only the gap-trimming and substitution model variations were considered.

The schematic topologies reveal that the 60% that make up the consistent parts are mostly based on the accepted genera of Reoviridae (see Fig. 51, Fig. 52, Fig. 53 and Fig. 54). While the succession of the groups is not consistent, the groups themselves remain together except in some cases for the phylogenies based on the pure amino acid alignment. However, except for *Marbled eel reovirus* and *White bream reovirus* that are supposed to be members of *Aquareovirus*, and *Aedes pseudoscutellaris reovirus* that belongs to *Dinovernavirus*, all accepted genera form distinct monophyletic clades. This is not the case for the phylogenies based on the hydrophobicity alignment where the groups tend to be more fragmented. Fig. 53 and Fig. 54 clearly show that the 'backbone' of the phylogenetic trees are the most difficult part to be correctly inferred by the phylogenetic reconstruction algorithms. Often, the inner topologies seem to be higher resolved compared to the outer topologies. Several single taxa, especially sequences originating from the transcriptomes, have no stable position and thus might be considered as 'rogue taxa' (Wilkinson, 1996).

The conflict in the data was visualized by a split network for each alignment variation (see Fig. 55 and Fig. 56). When comparing the two networks, it seems like there is less conflicting signal in the hydrophobicity alignment than in the pure amino acid alignment because the net is more dense at the basal nodes. However, it can be seen that the supposed monophyletic groups are still visible but the star-like origin of the phylogenies with a very high conflict in signal make it less likely to infer the correct topologies. This case shows that less conflict does not necessarily lead to a topology that can be resolved better. In addition, the possible re-assortment based on the segmented structure of the Reoviridae genome might be reflected in that high conflict in phylogenetic signal.

The recalculation of bootstrap support by BOOSTER (see Fig. 57) revealed very high support for most of the inner clades, that was already present in the original tree. However, the support for the backbone increased yet did not reach a high support in many cases. Overall this supports the assumption that the used algorithms for phylogenetic reconstruction are well able to group closely related taxa together yet fail to correctly reconstruct the relationships of deep branches as stated by Takahashi and Nei, 2000.

Despite uncertainties in the general topology, a number of new *Reovirus*-like sequences were identified. Most of them form clades with other known viruses that lack a proper classification up to this point but there are also some sequences that likely belong into established genera based on their consistent position within the trees. It is also possible to make assumptions about the classification of other published, not yet classified viruses. The sequences found in INSlupTASRAAPEI-89, INSlupTBKRAAPEI-31, INSpmbTABRAAPEI-227, and INSfrgTBCRAAPEI-57 are probably part of *Fijivirus*. The latter two are additionaly nearly identical to their neighboring taxa *Diaphorina citri reovirus* and *Nilaparvata lugens reovirus* and have been found in the same host species. INSeqtTCZRAAPEI-47 contributes sequences that could belong to *Seadornavirus*. In the case of INSofmTCYRAAPEI-79 and INSeqtTBNRAAPEI-11, it can be speculated that the viruses belong to *Phytoreovirus*.

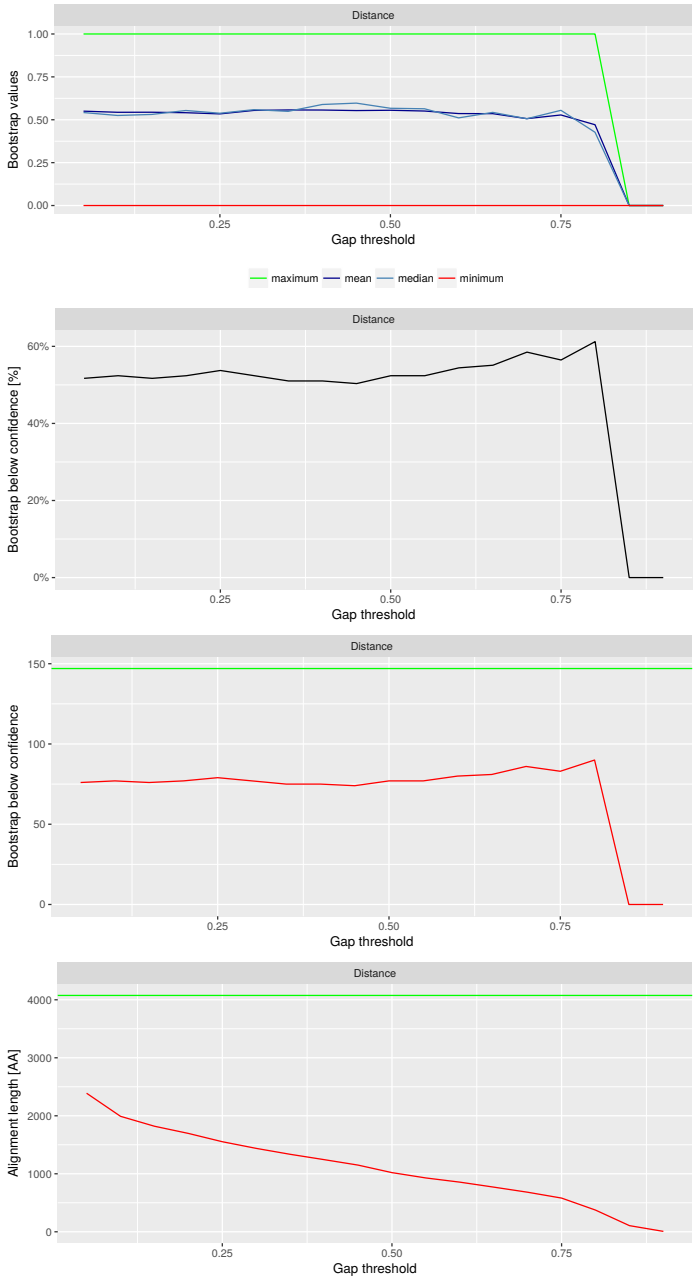

Figure 43: Bootstrap Support Based on the gap Trimming Variation of the Pure Amino Acid Alignment and the Neighbor Joining (r) Reconstruction.

The median bootstrap supports remained stable at around 50% regardless of the gap-trimming step until about 75-80% gap trimming, where the support decreased substantially. Before, the proportion of bootstraps below the confidence level was between 45% and 50%.

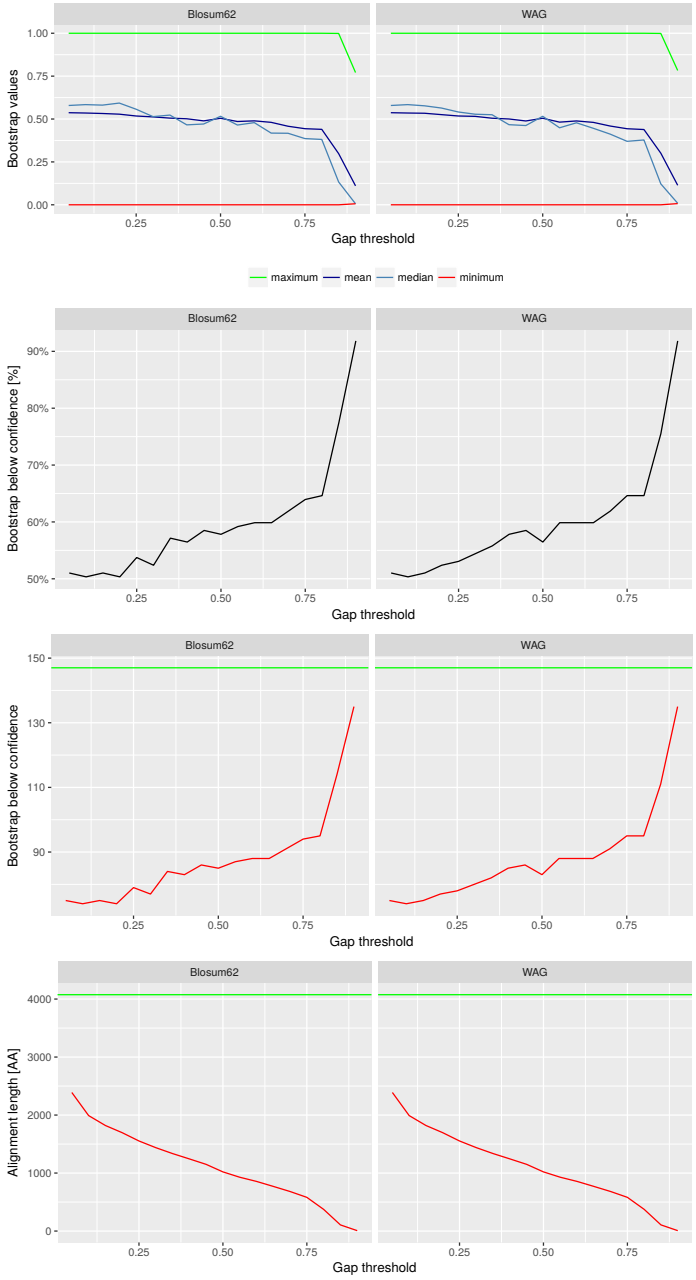

Figure 44: Bootstrap Support Based on the gap Trimming Variation of the Pure Amino Acid Alignment and the FastME Reconstruction.

The median bootstrap supports declined slightly from around 50% to 45% regardless of the gap-trimming step and substitution model until about 75-80% gap trimming, where the support decreased substantially. Before, the proportion of bootstraps below the confidence level was increased slightly from 50% to 65%.

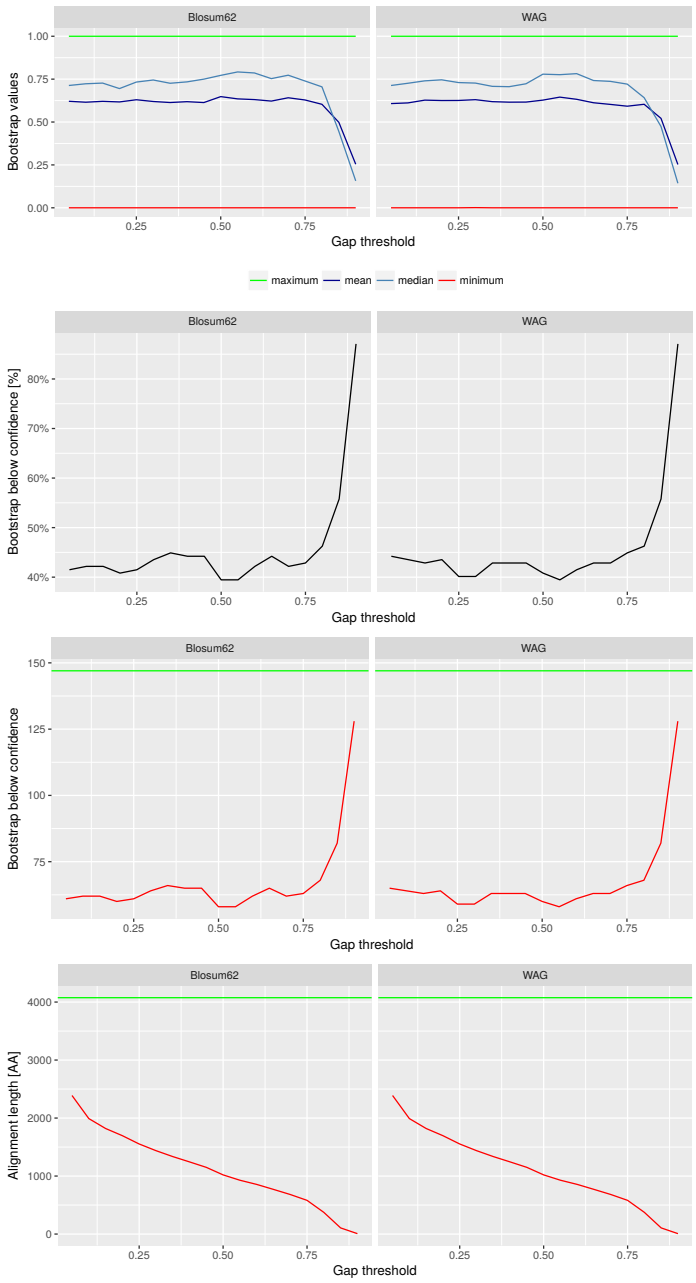

Figure 45: Bootstrap Support Based on the gap Trimming Variation of the Pure Amino Acid Alignment and the Phyml Reconstruction.

The median bootstrap supports remained stable at around 75% regardless of the gap-trimming step and substitution model until about 75-80% gap trimming, where the support decreased substantially. Before, the proportion of bootstraps below the confidence level was between 40% and 45%.

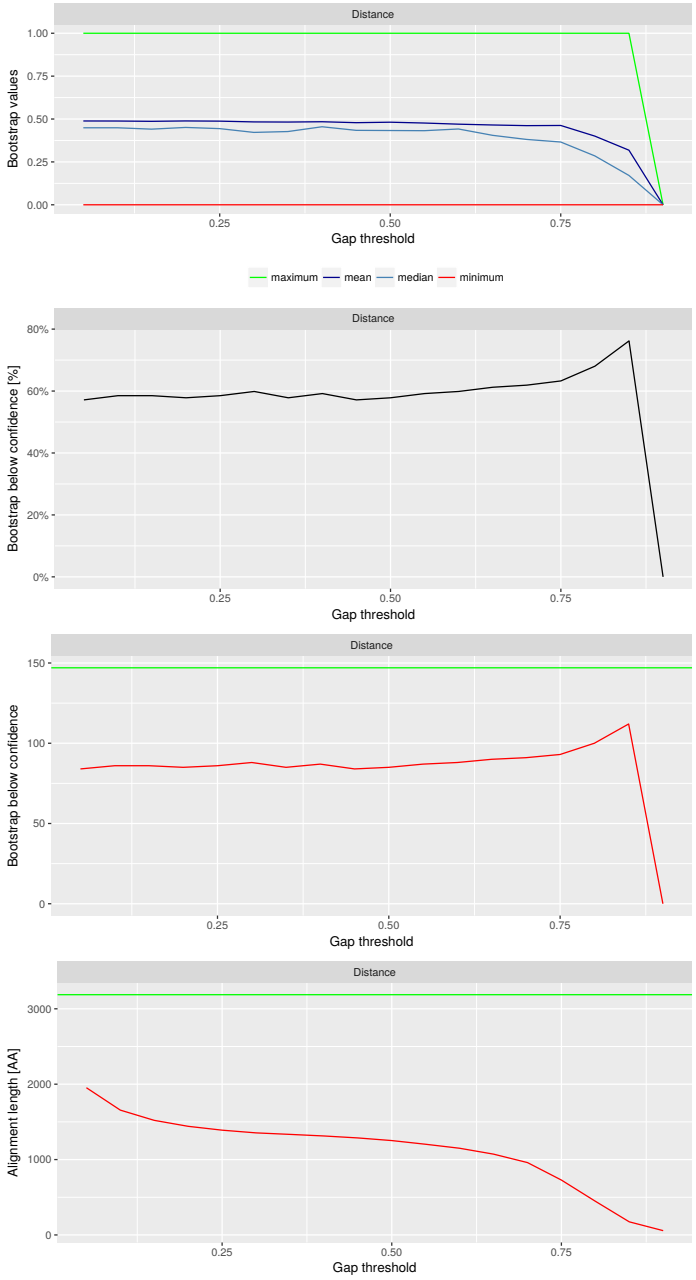

Figure 46: Bootstrap Support Based on the gap Trimming Variation of the Pure Hydrophobicity Alignment and the Neighbor Joining (R) Reconstruction.

The median bootstrap supports remained stable at around 50% regardless of the gap-trimming step until about 75-80% gap trimming, where the support decreased substantially. Before, the proportion of bootstraps below the confidence level was between 45% and 50%. The overall pattern is the same as for the pure amino acid alignment in Fig. 43.

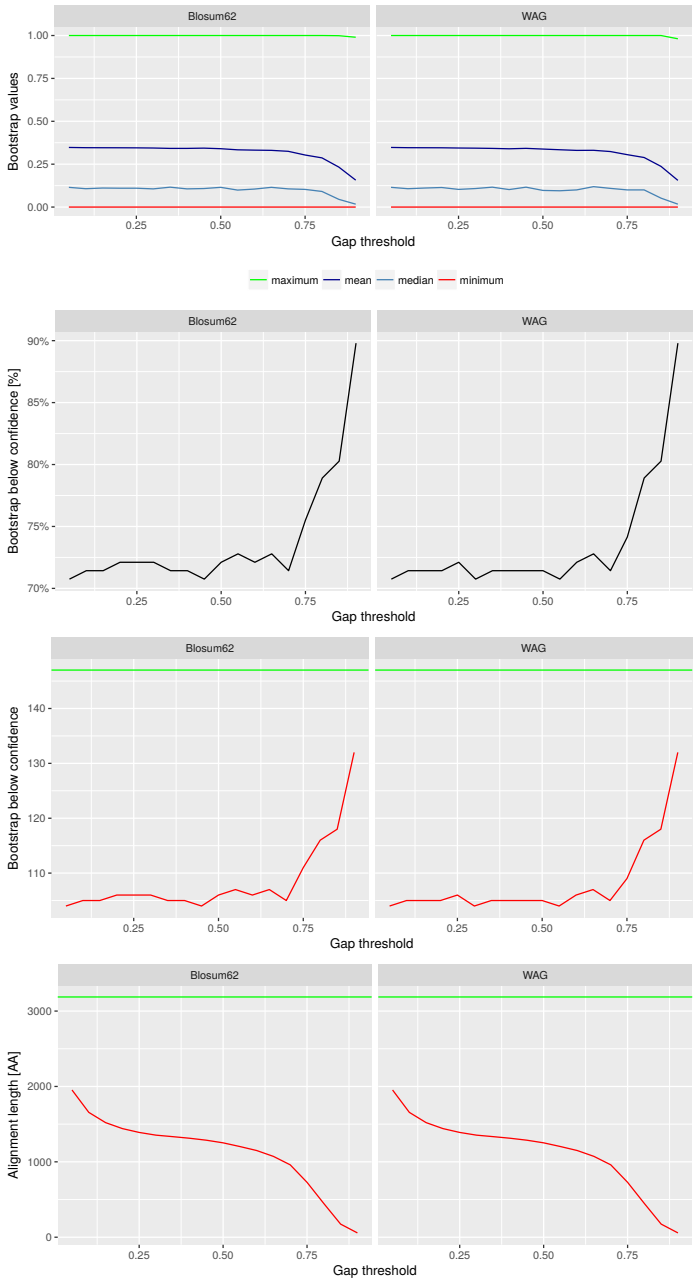

Figure 47: Bootstrap Support Based on the gap Trimming Variation of the Pure Hydrophobicity Alignment and the FastME Reconstruction.

The median bootstrap supports remained stable at around 35% regardless of the gap-trimming step and substitution model until about 75-80% gap trimming, where the support decreased substantially. Before, the proportion of bootstraps below the confidence level was was slightly above 70%. The overall pattern differed from the pure amino acid alignment shown in Fig. 44 by showing a much lower but consistent support.

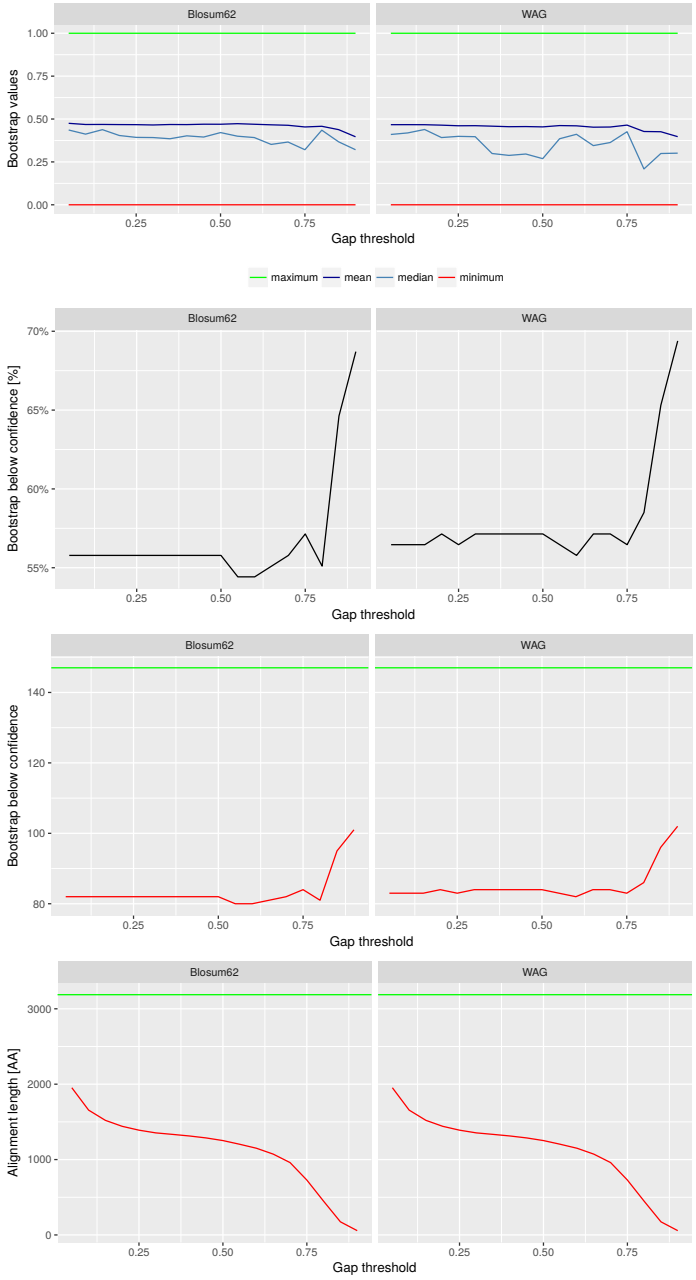

Figure 48: Bootstrap Support Based on the gap Trimming Variation of the Pure Hydrophobicity Alignment and the Phyml Reconstruction.

The median bootstrap supports remained stable at around 50% regardless of the gap-trimming step and substitution model until about 75-80% gap trimming, where the support decreased substantially. Before, the proportion of bootstraps below the confidence level was around 55%. In contrast to the pure amino acid alignment, the support was much lower than in Fig. 45.

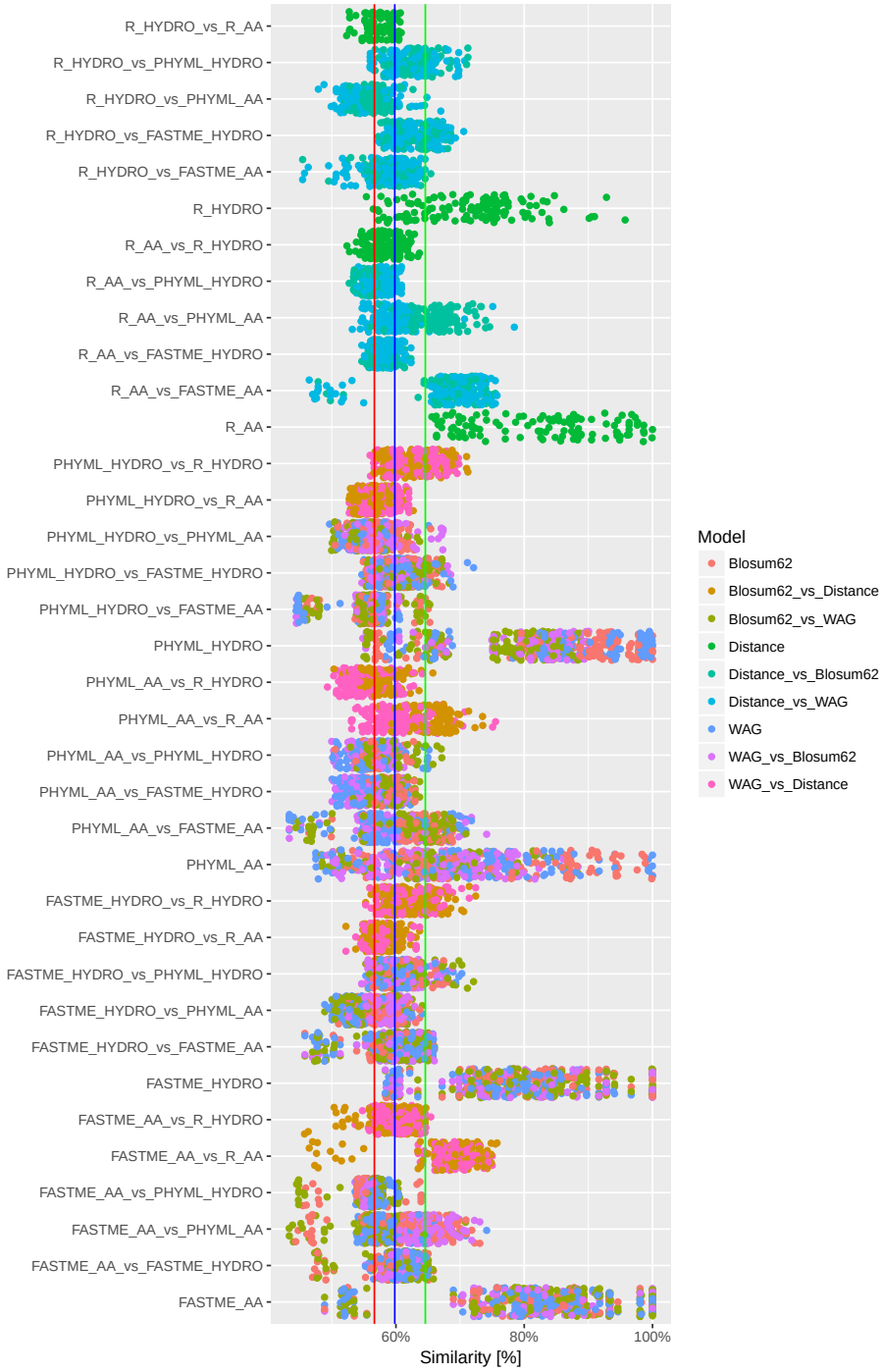

Figure 49: Calculated Distances Between all Topologies That Have Been Calculated Based on the Original Trees.

The combinations of substitution models and/or distance variation are color-coded. Vertical lines indicate the distribution o the calculated similarities with the lower quartile (red, 57%), median (blue, 60%) and upper quartile (green, 62%).

Figure 50: Calculated Distances Between all Topologies That Have Been Calculated Based on the Collapsed Trees.

For collapsing, the branch lengths of branches have been reduced to zero for all nodes that had a bootstrap support of less than 90%.The combinations of substitution models and/or distance variation are color-coded. Vertical lines indicate the distribution o the calculated similarities with the lower quartile (red, 57%), median (blue, 60%) and upper quartile (green, 62%).

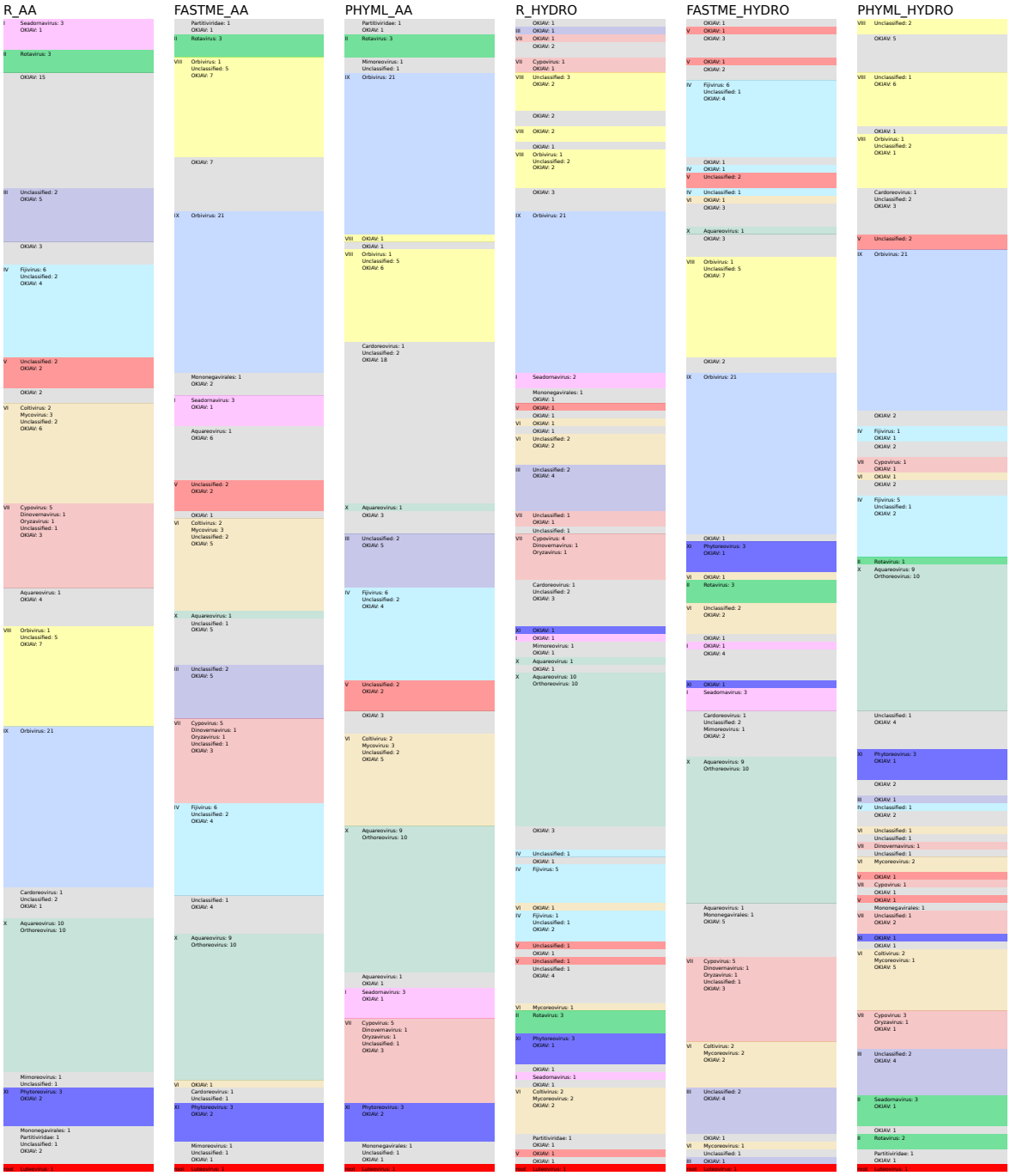

Figure 51: Schematic Topology of the Best Supported Phylogenies.

The colored boxes represent the groups that are considered to be stable with the gray boxes representing the unstable proportions of the phylogeny. The sizes of the colored boxes corresponds to the number of leaves in the respective group and the members of the respective groups are summarized by taxomy. The acronym OKIAV (**O**ne **KI**TE **A**ssociated **V**irus) indicates a potential virus obtained from the 1KITE transcriptomes.

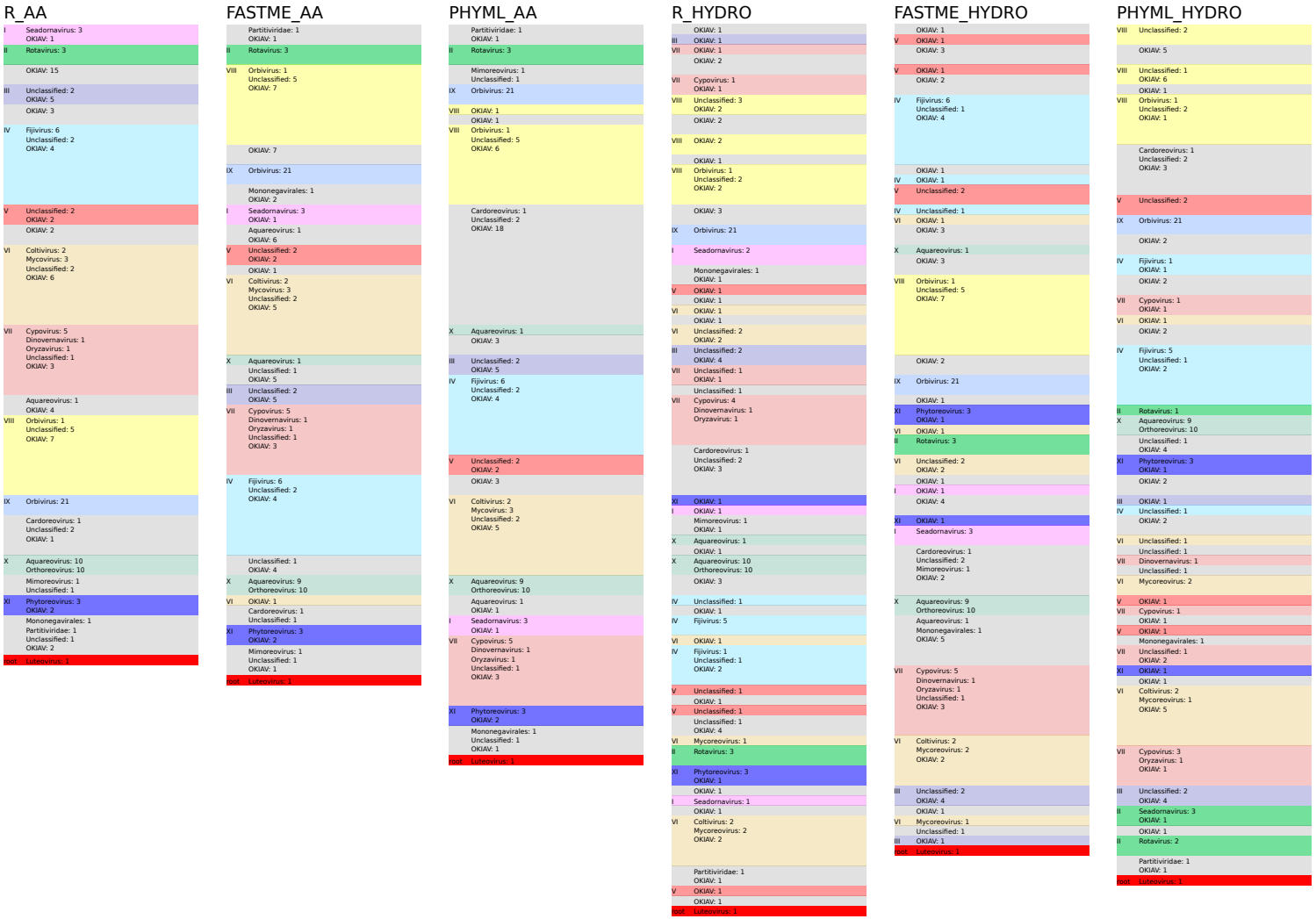

Figure 52: Adjusted Schematic Topology of the Best Supported Phylogenies Based on Fig. 51.

The sizes of the colored boxes has been adjusted for better readability. Large groups have been reduced by half their size and singletons have been doubled in size. The acronym OKIAV (**O**ne **KI**TE **A**ssociated **V**irus) indicates a potential virus obtained from the 1KITE transcriptomes.

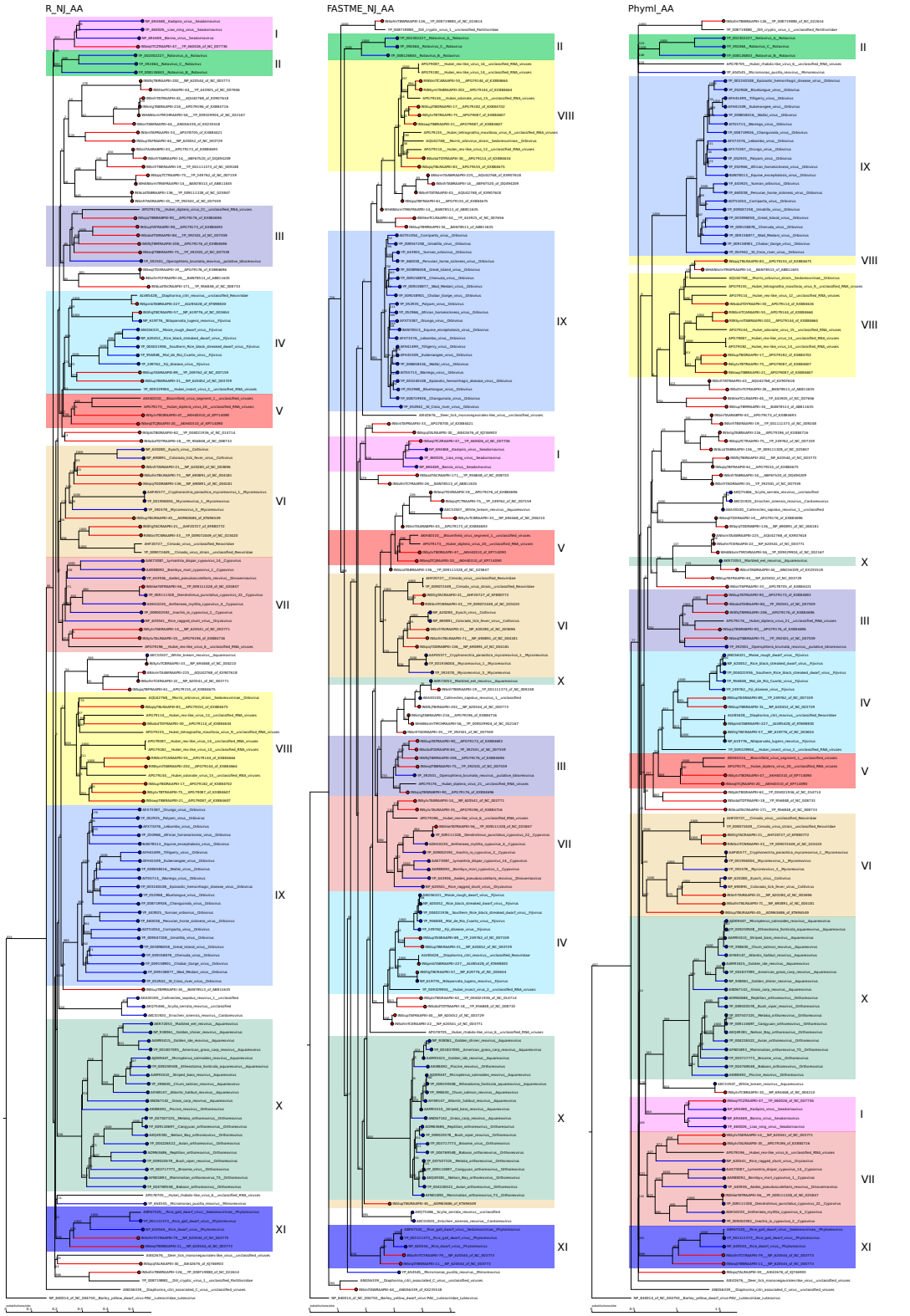

Figure 53: Detailed Best Supported Phylogenies for the Pure Amino Acid Alignment. Highlighted are the groups of more than three leaves that form more or less stable monophyla based on the NJ topology reconstructed by R. Blue branches indicate that the viruses were in the initial reference library, red branches indicate viruses from the 1KITE transcriptomes and black branches indicate references that have additionally been retrieved from the reciprocal BLAST against the NR. Tip labels contain the respective genera after the species names for reference viruses. Sequences from transcriptomes are labeled with the assembly ID followed by the genebank accessions for the protein and nucleotide sequence of the best BLAST match.

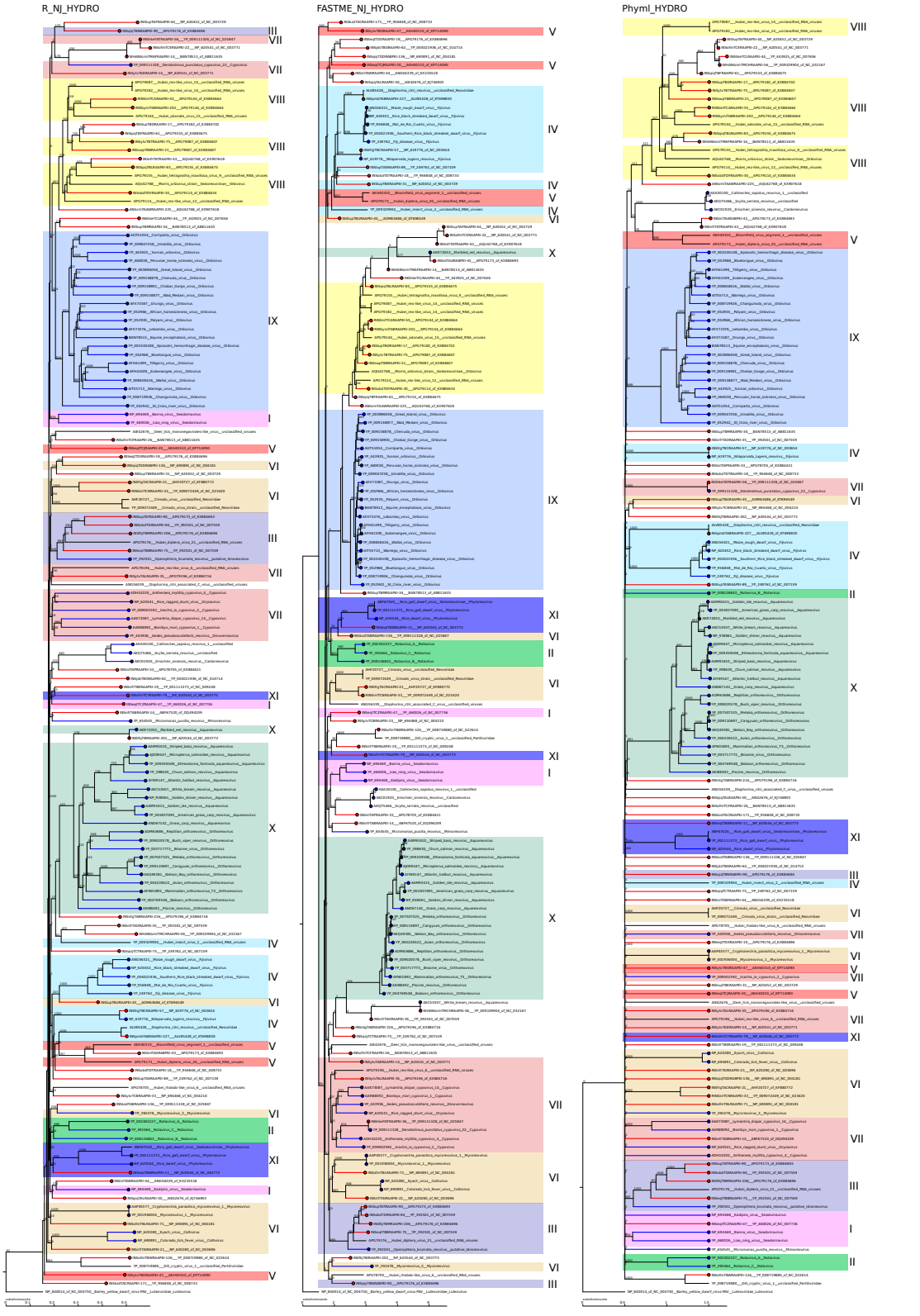

Figure 54: Detailed Best Supported Phylogenies for the Hydrophobicity Alignment.
Highlighted are the groups of more than three leaves that form more or less stable monophyla based on the NJ topology reconstructed by R in Fig. 53. Blue branches indicate that the viruses were in the initial reference library, red branches indicate viruses from the 1KITE transcriptomes and black branches indicate references that have additionally been retrieved from the reciprocal BLAST against the NR. Tip labels contain the respective genera after the species names for reference viruses. Sequences from transcriptomes are labeled with the assembly ID followed by the genebank accessions for the protein and nucleotide sequence of the best BLAST match.

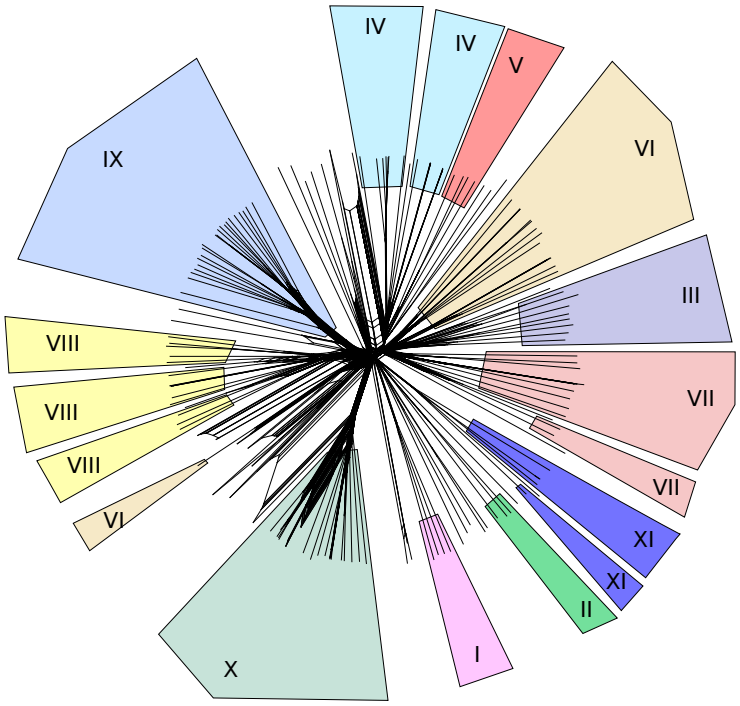

Figure 55: Neighbour-Network for the Pure Amino Acid Alignment by SplitsTree.
Highlighted are the groups of more than three leaves that form more or less stable monophyla based on the NJ topology reconstructed by R in Fig. 53.

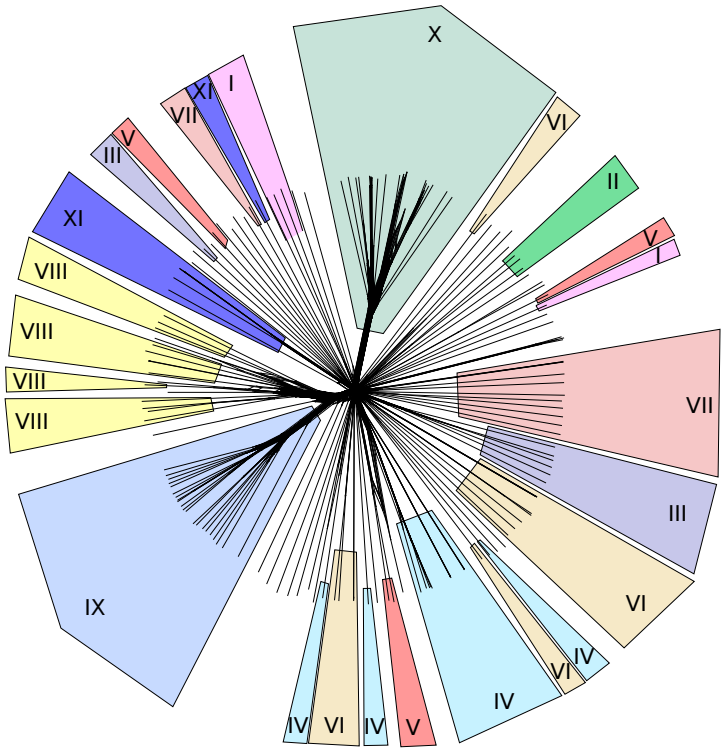

Figure 56: Neighbour-Network for the Hydrophobicity Alignment by SplitsTree.
Highlighted are the groups of more than three leaves that form more or less stable monophyla based on the NJ topology reconstructed by R in Fig. 53.

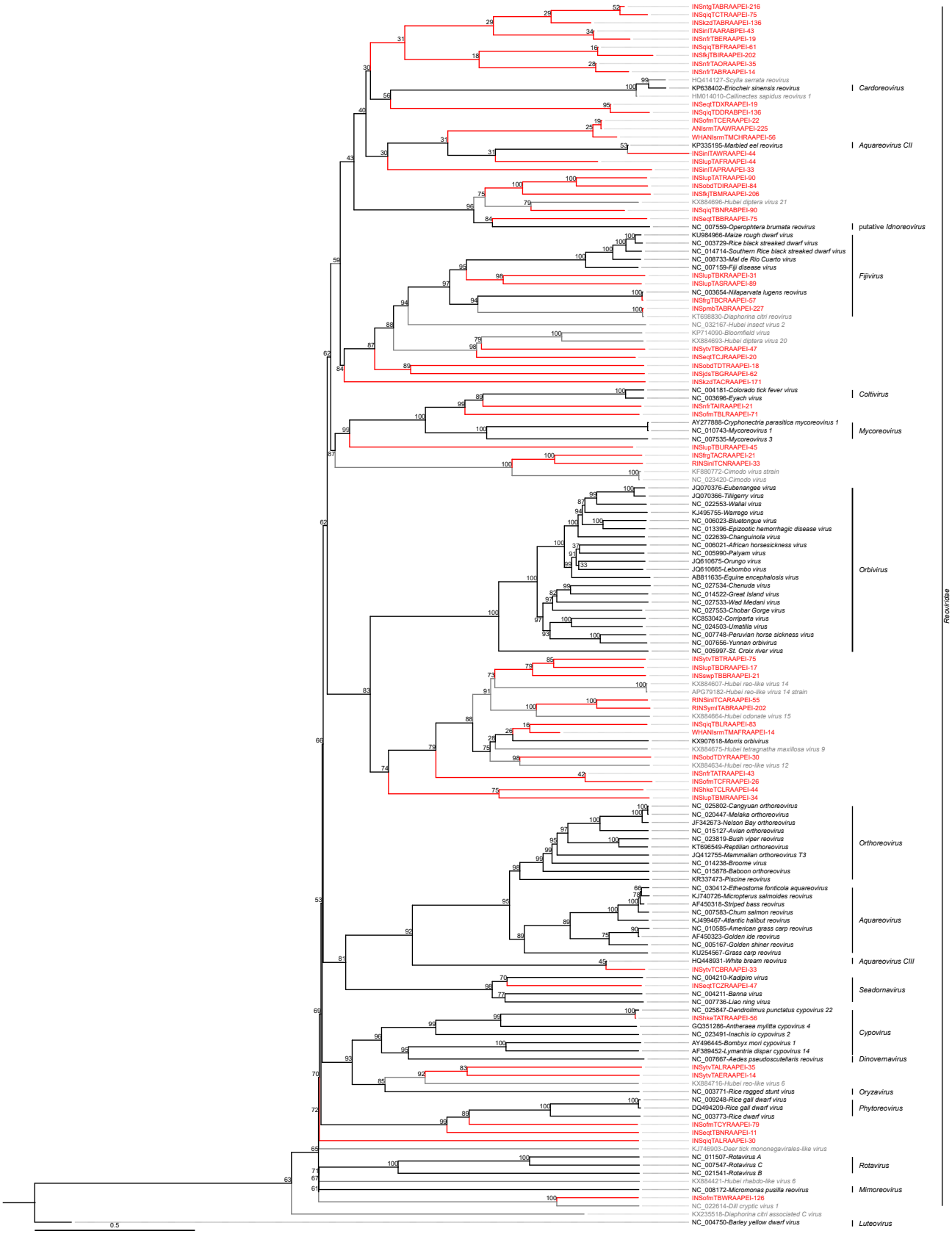

Figure 57: PhyML Tree With Transfer Bootstraps.

Black branches indicate that the viruses were in the initial reference library, red branches indicate viruses from the 1KITE transcriptomes and gray branches indicate references that have additionally been retrieved from the reciprocal BLAST against the NR.

4 Discussion

4.1 Preliminary Work

It was possible to retrieve verifiable viral sequences that are related to known viruses. However, the amount of potential viral sequences was more than expected. This leads to more questions about the reliability of genomic and transcriptomic data. It is of utmost importance to make sure that the sequencing data contains only sequences of the intended organism.

Especially in cases where a single individual cannot provide enough nucleic acid that can be cleanly extracted from a specific tissue, it is nearly impossible to only sequence the targeted nucleic acids. If the gut of the organism or even multiple organisms is part of the prepared sample, the microbiome and diet are part of the dataset as well. Without a proper reference genome that can be used for mapping, it is difficult to tell real host sequences apart from other organisms. The same is true for the association of a virus with its host. Thus, whenever there is a virus identified, it could have been ingested.

About 40% of the potential viral sequences were too short or too divergent to be included into the alignments that have been used for phylogenentic reconstruction. The reliability of those findings is questionable. There are two main reasons for the sequences being so short. First, RNA is degraded very fast compared to DNA (Ross, 1995). Thus it is reasonable to assume that depending on the age of the sample before actual sequencing took place, much of the RNA has already been degraded. If the sequence was of real viral origin from a remittent infection, its grade of decay is also expected to be higher. Secondly, it is a general problem in deep sequencing processes, that some sequences or regions are not sequenced based on primer design and other methodological errors (Laehnemann *et al.*, 2015). The following assembly steps rely on the amount of overlap of the reads. If several reads based on a single strand have no overlapping reads that connect them to each other. So despite a near full genome of a virus was within a sample, some areas could have been sequenced with a high coverage while other areas that are not represented at all.

Although it was possible to identify several true positive viruses, it is not clear whether the very short fragments are actually of viral origin. However, effort was taken to minimize false positives by matching the sequences via BLAST and InterProScan with larger databases. Since the template alignments used for the initial search contained only the RdRp-regions, it might have been possible to detect other ubiquitous domains that are similar. In the simplest case this could be other RNA-binding sites.

Searching transcriptomic data for unknown viruses using pHMMs is a promising method. HMMER3 is reasonably fast for the short viral sequences and allows a quick screening of mass data. The difficult task was to interpret the findings and put them into the right

context while making sure that the obtained sequences were actually of viral origin. Using several sequence matching methods for identifying other relationships and functions based on larger, non-viral databases is a critical step for verification, if no laboratory methods are applicable. These steps and gathering additional information about the potentially related viruses were only partially automated and it was a lot of manual effort involved to evaluate the potential viral sequences. However, the general concept and work-flow led to interesting results and was taken a step further in the respective chapters of TRAVIS.

4.2 TRAVIS

TRAVIS facilitates the automation of identification of potential viral sequences and delivers all necessary data that allows fast and direct interpretation of the results by researchers. As in the preliminary work, the use of Hidden Markov Models turned out to be very fast yet reliable for virus research. This has also been confirmed by Skewes-Cox *et al.* (2014). It was possible to retrieve nearly all true positives by using HMMSEARCH and JACKHMMER with only a fraction of the calculation time the other search tools needed. Although it is important to have several methods agree on what is supposed to be a potential viral sequence, a quick search by just using HMMER3 for preliminary studies can save a large proportion of the calculation time. The reciprocal BLAST against the NR still needs to be done for an easier detection of false positives and finding better matching viruses. One thing is always necessary to consider when using large public databases. They are often contaminated with *e.g.* human sequences (Longo *et al.*, 2011) and are not free of annotation errors. However, they are very useful when such things are heeded. Yet, to increase the speed of the reciprocal check of the suspicious sequence, the same database could also be used with JACKHMMER instead of BLAST.

The biggest drawback of TRAVIS is the generation of the reference library. It is not only advised to keep the reference library up to date but also the curation with metadata such as correct taxonomy is very time consuming and error-prone. Until now, TRAVIS needs a user specified reference library to run properly. In the case of Reoviridae, it was partially difficult to find the correct sequences for the viruses based on the ICTV taxonomy report. For example, *Aquareovirus* is comprised of *Aquareovirus A* to *Aquareovirus G* and *Mycoreovirus* of *Mycoreovirus 1* to *Mycoreovirus 2*. Some of the known viruses have been renamed and/or are listed under a different name on NCBI and it is difficult to determine whether the virus belongs to an ICTV-accepted genus or not. This was especially misleading when reconstructing the phylogenies in the case of *Marbled eel reovirus* and *White bream reovirus* that are supposed *Aquareoviruses* but are not monophyletically clading with the remaining *Aquareoviruses*. A 'blind' search that only takes a sequence database with no need for metadata is currently in development, however an integration of public, virus-specific databases such as vFam (Skewes-Cox *et al.*, 2014) is worth considering. Additionally, access to a local reference database for comparing results with yet unpublished viruses has to be implemented. TRAVIS has not been tested on large DNA viruses yet. But as some matches based on the NR in the search for Reoviridae show, the graphical display of very large sequence organizations needs to be optimized by *e.g.* adding an option for scaling of the TRAVIS Scavenger plots.

InterProScan was used extensively in the preliminary work and also was useful for *e.g.* determining the fallacious sequences for the run on Reoviridae. Despite its annotation

capabilities, it has not been implemented in TRAVIS for the following reasons. First, it is depending on an internet-connection. For example, if the necessary ports on the machine or the network, TRAVIS is running on, are blocked or the connection breaks in a larger process, many steps have to be rerun. Error tracing might be very complex and disarrayed. Second, a local installation of InterProScan is unfortunately rather complicated and requires several hundred gigabytes of databases that have to be updated regularly. Since the installation and usage of TRAVIS was supposed to be as easy as possible, this would have contradicted one of the main aims of the pipeline. Third, for many of the known Reoviridae, no useful protein domains could be detected (chapter 2.3.1). This makes it difficult for proteins without predictable domains to be properly compared and annotated. Therefore, a custom visualization for the direct sequence comparison was developed for TRAVIS. The calculation for these visualizations rely on BLAST and thus are fast and independent on known functional annotations. Hence there is no need to know the domain structure and functions to be able to identify similar sequences. However, additional domain search with InterProScan can provide more insight, if domains are detectable.

Another drawback in this run was the amount of reported false positives. This was mainly due to some proteins of *Reoviruses* that contain ubiquitously expressed domains which can be found in many genomes. Additionally, the sensitivity has been set very high with an overall e-value cutoff of 10^{-6}. Despite this was set on purpose to maximize the detection of 'real' viral sequences, it imposes an additional burden on the researcher that has to interpret the results. However, the approximate 42 million transcripts could have been reduced to about 2600 potential reoviral sequences where it was mostly easy to distinguish between true and false positives. For the searches with the pHMMs, it might have been possible that the alignments they were based on were suboptimal because of the low similarity of the individual sequences and eventually created misleading results. Generally, alignments of viruses might be suboptimal because there are many small areas that can match multiple times on the same sequence and thus create errors. This is well visible in the sequence organization plots created by TRAVIS Scavenger. However, apart for the e-value threshold, default settings have been used for the search tools to see how well they can handle diverse sequences and being set up by beginners. If other parameters are adjusted appropriately, the amount of false positives could likely be reduced while maintaining the high sensitivity. It cannot be completely ruled out that some the sequences that were labeled as false positives are indeed true positives that are just too divergent from the known viruses in the databases and thus make verification impossible.

The best matches were set as best matches subjectively based on the visualizations provided by TRAVIS Scavenger by the person that evaluated the results and thus were not purely based on objective criteria. This is the part where human interpretation is not

completely avoidable until now. The decision on best matches and especially true and false positives is on the edge of statistical measures combined with experience in virus annotation that algorithms cannot yet provide. In future, machine learning algorithms implemented in neural networks will be likely helpful in reducing subjective human bias in the evaluation (Jagadish *et al.*, 2014; Dunjko and Briegel, 2018).

There is a big caveat for all the obtained potential viral sequences. It is important to distinguish between the discovery of a virus and the detection of a nucleic acid sequence of potential viral origin (Calisher and Tesh, 2014). Despite it is possible to extract whole genomes worth of nucleic acids from samples it does not necessarily mean that the organism from which the sample originates actually suffers from a viral infection. Additionally, if fragments of a sequence were found to be potentially on one segment, they were combined (see chapter 2.3.5). A co-infection of two similar viruses cannot be completely ruled out. The artificial generation of chimeric sequences also impede the proper reconstruction of a phylogeny. Chimera are considered to be sequences that are derived from two different parents and can be a very problematic artifact in PCR-based sequencing methods (Wang and Wang, 1996; Ashelford *et al.*, 2005; Edgar *et al.*, 2011).

However there are good chances for the true positives to be fully functional viruses. Full virus genomes in the bivalves *Crassostrea gigas* and *Mytilus galloprovincialis* were extracted using bioinformatics and then confirmed as functional viruses in the laboratory (Rosani and Gerdol, 2017). Since virus databases have been augmented with reference sequences, this backwards approach to classical virus detection is feasible. The classical virus detection already allowed to identify viruses that are very distantly related to known viruses based on sequence similarity. For example, *Micromonas pusilla reovirus* has been extracted via classical laboratory procedure and shows amino acid identity of 8-10% to *Aquareovirus* and 21% to *Rotavirus A* for the RdRp (Attoui *et al.*, 2006a). Despite the low sequence similarity, the structure of the genome, and the function of the genes therein, it has been classified as a proposed the new genus *Mimoreovirus* within Reoviridae. This an example for the high diversity within the family. Additionally, VP1 of this virus was found similar to bacterial hemagglutinins at about 38-40%. Similarities to non-viral genes have also been reported for *Liao ning virus* (Attoui *et al.*, 2006b).

This is important in context with similarity estimations to potential viral sequences obtained from the 1KITE data. Especially considering the large amount of potential viral fragments that could have been found in the preliminary work, the findings can be regarded as support for the progressive hypothesis on the origin of viruses (see chapter 1.1; Wessner, 2010). If a combination of Insect and bacterial genes could make up a fully functional virus, even multiple origins of viruses could be worth considering.

The segmentation of Reoviridae might have several other implications as well. Re-

assortment is supposed to be an important mechanism in virus evolution (Domingo and Holland, 1997). In addition, as the assimilation of other foreign genes cannot be completely excluded and might lead to a higher fitness, for example by enabling the virus to infect another host. This could also explain the difference in number of segments for several Reoviridae (Attoui *et al.*, 2006a). Depending on the host and thus the available host genes that are used in virus proliferation, additional segments might be needed or not necessary and therefore can get assimilated or lost. Eventually this leads to the diversification of genome structure in the terms of number of segments. It can be speculated that the assimilation of host genes into a reoviral genome can be initiated by the addition of host mRNA into the virion.

The occurrence of viruses in insects does not necessarily have to be parasitic to the primary host. There are insects like parasitic wasps that live in symbiosis with viruses and those viruses are essential for the reproduction of their hosts (Burke *et al.*, 2014; Burke, 2016). The wasps lay their eggs into other animals they parasitize. The virus is transmitted during that process and interferes *e.g.* with the immune system of the infected host so that the eggs can hatch and feed on the host. This imposes the question whether those insects domesticated or even generated their symbiotic viruses from their own genome. The known symbiotic viruses are Polydnaviridae, which are not in the focus in this study, but TRAVIS probably can be used for studies on this subject as well. Despite the Polydnaviridae consist of two very divergent genera, they are also thought to have a common ancestor (Béliveau *et al.*, 2015).

The most difficult issue in this thesis was the inference of phylogeny. While telling true positives apart from false positives was possible, the diversity of obtained potential viral sequences was more difficult to interpret. Although all analyses were based on sequence similarity that could be very low in some cases, it has to be noted that similar does not necessarily mean that the sequences are homologous (Reeck *et al.*, 1987), but a phylogeny has to be based on homology (Stevens, 1984). Assuming a common origin of viruses in general and RNA viruses in particular with the RdRp as a central gene, the phylogeny of the viruses in this study was based on the implied homology of detected RdRps with similar sequences..

However, the high divergence based on the high mutation rates (Holland *et al.*, 1982) generally makes it difficult to infer a 'correct' alignment and eventually phylogenies that are based on this alignment. Viruses have a unique selective pressure and assuming new models on evolutionary traits like substitution rates can take this into consideration (Dimmic *et al.*, 2002; Dang *et al.*, 2010). The molecular clock of different strains of the same virus can vary and thus make tree inference more complex. Considering different evolutionary rates for different viruses and different strains could improve phylogenies (Dunham and Holmes,

2007). Yet, this is likely not possible to achieve for so many taxa. Additionally, the three-dimensional structure of the encoded proteins can give more insight on the actual similarity of functionality of the proteins. (Richards, 1977; Floudas *et al.*, 2006; Wright and Dyson, 1999). For example, T-coffee (chapter 2.1.1.7) is capable of using structural information to infer alignments. These features may contribute to phylogenies and compensate for the short genomes. Such structural data could be derived from sequence information as it is for viruses in VIPERdb (`http://viperdb.scripps.edu/`; Carrillo-Tripp *et al.*, 2009).

However, even if the alignments are optimal, it is not always possible to reconstruct stable phylogenies. Especially on studies where several genes have been concatenated for phylogenetic reconstruction, the change in gene composition has a significant impact on the inferred phylogenies (Shen *et al.*, 2017). Other problems occurred on very divergent deep branching datasets comparing Bacteria, Archaea and NCLDVs. These phylogenies were probably reconstructed using inappropriate methods (Forterre and Gaïa, 2016). This shows that the used methods for alignment and phylogenetic reconstruction have to be tailored to fit the dataset for proper inference of phylogenies. Additionally, for segmented viruses like Reoviridae, where horizontal gene transfer can happen, assuming a bifurcating phylogeny is not cogent since it does not reflect the actual biological history. This is not only the case for viruses but also *e.g.* for many prokaryotes (Gogarten and Townsend, 2005; Zhaxybayeva *et al.*, 2006). As previously stated, networks are suitable for showing the conflicting signals in multiple sequence alignments that are used for inference of phylogeny (Iranzo *et al.*, 2017; Bastkowski *et al.*, 2017) and thus deliver more informative phylogenies. However, it is worth considering to use different new algorithms for inferring phylogenies as well. For example, PhyQuart (Kück and Wägele, 2016; Kück *et al.*, 2017) is a split based phylogenetic reconstruction algorithm that is able to outperform ML based algorithms in terms of reconstructing the right topologies for very long sequences. It is not yet applicable for the short virus sequences but it is actively developed and enhanced functionality might help to resolve virus phylogenies eventually.

4.3 General Discussion

Many potential viral sequences could only be retrieved fragmentarily from the transcriptomes. This is mostly due to the fact that the assembly success for a transcriptome is determined by the sequencing efficiency, the assembly algorithm and the condition of the sample. Yet since the obtained potential viral sequences were actually expressed, the chances to have detected a real viral mRNA are high. However, according to the progressive hypothesis of virus origin (Wessner, 2010), inactive regions on the host DNA (introns) could have undergone a mutation that causes them to be transcribed. If such regions contained protein domains that can perform viral functions or improve the fitness of a virus that is currently infecting the cell, this gene could be integrated into the genome of the respective virus. Verification of the new found viruses *in vitro* via PCR or in cell cultures could not be done due to the fact that it was not possible to get aliquots of the original samples yet as it has been done in other studies (Rosani and Gerdol, 2017). Additionally, it is not completely possible to predict genome sizes of very distantly related viruses because the genome structure can change drastically within a group of viruses. These changes are *e.g.* repositioning, deletion or insertions of ORFs or even gain and loss of whole segments. Therefore the new found viruses in this study remain tentative until similar viruses are found that actually are fully characterized in laboratories based on cell culture or fresh samples from infected organisms (Calisher and Tesh, 2014).

TRAVIS is currently in a state that allows fully automated screening of data, yet several further improvements on functionality can be suggested. The amount of false positives still is very high and imposes a burden on the researcher. A check of the suspicious sequences against a small database containing ubiquitously expressed proteins like zinc fingers as shown in chapter 3.2.2 can at least flag fallacious sequences. TRAVIS is capable of the implementation of new own functions as well as additional third-party algorithms. This allows to add more search tools like Diamond (Buchfink *et al.*, 2014) or meta-classification tools like Kraken (Wood and Salzberg, 2014) or GOTTCHA (Freitas *et al.*, 2015). The latter ones could be especially useful for the identification of false positives. Also, filters and scaling options for the plots generated by Scavenger will allow to speedup the evaluation. The generation of a preliminary phylogenetic tree for all suspicious sequences and the respective references for a general overview during evaluation is planned as well.

The outlook on providing a sample, a reference database and getting a fully annotated virome for the sample including tentative phylogenies is very enticing. However, the exploration of viral diversity on trancriptomic data in general is expected to contribute to the efficiency of viral research by flagging sequences as potentially viral that have not been annotated otherwise. It will help to identify novel viruses in future metagenomic studies and medical treatment of patients that suffer from symptoms with unknown causes.

5 Summary

Most of the ongoing virus research is focused on mammalian and bird viruses, which are well known to be directly or indirectly associated with human diseases. While many viruses are transmitted by blood-feeding arthropods (Arboviruses), virus research on non-bloodfeeding arthropods has long been neglected. Within arthropods, insects are the most diverse animal group on earth and can be found in virtually every habitat. They play a key role in ecosystem health and thus set the basis for many environmental impact assessment studies. The under-estimation of viral diversity was recently made evident by broad sampling of arthropods and other invertebrates. Knowledge about viruses in insects can therefore give insight on the emergence and evolution of viruses. Discovery of yet unknown viruses and consequently, preparedness for emerging diseases are vital to prevent epidemics, especially in the context of globalization. Advancements in metagenomics with rapid growth of available gene databases in recent years have facilitated the exploration of virus diversity.

Transcriptomes from the '1000 Insect Transcriptome Evolution Project' (1KITE; `http://1kite.org`) have been screened for several groups of RNA viruses. In contrast to a genome, where DNA is sequenced, RNA of a sample is sequenced for a transcriptome. Therefore, only expressed genes of an organism is present in a transcriptome. However, it may contain RNA of viral origin as well. This dataset contains transcriptomes of over 1000 different arthropod species covering all extant orders of hexapods. The primary goal of 1KITE is to solve questions about the evolution of insects but in this study the focus is on the broad range of novel viruses that is expected to be within this large dataset.

Since viruses have very high mutation rates and databases have a bias towards viruses that have an impact on humans, livestock, and agriculture, it is required to combine expert knowledge with sensitive search algorithms and appropriate support tools. A new kind of bioinformatic consistency-based virus detection pipeline called TRAVIS (**TRA**nscriptome **VI**rus **S**canner) is proposed in this study. It is designed for the sensitive mass screening of transcriptomic data directed towards a specific virus group in order to find new, distantly related viruses in addition to closely related. It uses different search algorithms including BLAST, profile Hidden Markov Models (HMMER3) and a new *k*-mer approach implemented in MMSeqs2. The computational work-flow is mostly automated and delivers statistical and visual output for improved result evaluation.

Specific databases containing different groups of RNA-viruses were used to systematically scan the 1KITE transcriptomes. Hundreds of potential new viruses were identified and partially characterized. While some of those viruses could have been assigned to existing taxonomical groups, the phylogenetic distance of many findings indicate novel virus genera and families.

6 Appendix

The full appendix can be found in the digital supplementary material.

6.1 Related Publication

Käfer, S., Paraskevopoulou, S., Zirkel, F., Wieseke, N., Donath, A., Petersen, M., Jones, T. C., Middendorf, M.,Junglen, S., Misof, B., M., Drosten, C. (2019). Re-assessing the diversity of negative strand RNA viruses in insects. *Submitted manuscript*.

6.2 TRAVIS Documentation

6.2.1 Introduction

The configuration of TRAVIS is done by creating manifest files that contain all necessary information. These manifest files are actually plain-text comma separated value files (CSV) that you can edit either in a text editor of your choice or spreadsheet software like LibreOffice, OpenOffice or MS Excel. But when you export the CSVs make sure that the export has been done properly. That means opening it in a text editor and check whether the entries are actually separated by comma and not by semi-colon (the german version of Excel does that!). Another problem are quotes around the entries. TRAVIS does not like quotes. Also please only use alphanumeric characters, dashes and underscores for whatever you enter in the manifest files. Note that TRAVIS internally uses double and triple underscores as separators.

The beta version of TRAVIS is available at `https://github.com/kaefers/travis`. This guide is not comprehensive for all functionality as more features will be implemented in the future. It assumes that you have basic knowlege about the use of the Unix command line.

6.2.2 Concept and Workflow

Each of the TRAVIS subprograms (Henchman, Core and Scavenger) is called with a single manifest file (see chapter 6.2.4) as a parameter. If you want to have a completely automated run of TRAVIS without manual interaction, you can call them subsequently in *e.g.* a bash script.

```
$perl TRAVIS_Henchman_vX.pl TCC.csv
$perl TRAVIS_Core_vX.pl TCC.csv
$perl TRAVIS_Scavenger_vX.pl TCC.csv
```

However, TRAVIS Henchman creates a manifest file called 'Troubling TRAVIS Table' (see chapter 6.2.5), where all intended searches for TRAVIS Core are listed. You can adjust this table according to your specific needs. This can drastically reduce calculation time. It is also possible to manually create a TTT or use an old one and skip TRAVIS Henchman.

If you have a large dataset, you can run TRAVIS Scavenger on the same TCC while TRAVIS Core is still running in order to get the results that have already been generated. Because TRAVIS runs all intended searches completely for each sample and logs the results, it is also possible to resume calculations from the last processed sample.

Each sample gets an own set of output files based on the given ID in the sample library. These output files will be fastas, tables (CSV) and visualizations of (SVG) the matches. I recommend to open the SVGs in a web-browser because detailed information about the

matches will be displayed when you hover your cursor over certain elements. This has been tested with Mozilla Firefox and Google Chrome under Windows 10 and Ubuntu 16.04. These details can also be found in the corresponding CSV.

For details on the general concept see chapter 2.3.2.1.

6.2.3 Installation

TRAVIS is written in PERL and should work out of the box on most UNIX systems. If you have compiled versions of HMMER3, BLAST+, MMSeqs2 and MAFFT, you are good to go. You can specify the paths in the configuration file. However, if you have the programs installed and working with shortcuts/aliases, you can also use these. A combination of HMMER3 (v. 3.1b2), BLAST+ (v. 2.6.0) , MMseqs2 (v. 5437c6334d659119089cd8758a63838c29753048) and MAFFT (v. 7.302) worked well on Ubuntu 16.04 LTS but i guess, other versions won't make problems as long the respective developers do not change their parameter calls or output format.

6.2.4 TRAVIS Control Center (TCC)

This is the main configuration file where all necessary parameters are entered. Parameter names and examples can be found here.

6.2.4.1 database_name

Contains the name of the database to be generated by TRAVIS Henchman. If you already have a prepared database that you want to use as it is, you can skip TRAVIS Henchman, modify TCC, and start TRAVIS Core.

```
database_name,reo_full
```

6.2.4.2 resume_calculation

In case of crashes, you can resume the calculation based on the last save point. That save point is the last completely searched sample.

```
resume_calculation,1 or 0 encoding on/off
```

6.2.4.3 sample_dir

Specifies the path to the nucleotide data.

```
sample_dir,/TRAVIS/assemblies/fastas/
```

6.2.4.4 ORF_dir

Specifies the path to where the ORF data should be stored.

```
ORF_dir,/TRAVIS/assemblies/ORF_data/
```

6.2.4.5 ORF_length

Sets limits to the ORFs to be extracted in number of amino acids.

```
min_ORF_length,50
max_ORF_length,3000
```

6.2.4.6 sample_library

Specifies the path to the sample library with 'filename','ID','factor1','factor2','factor3'... You can add as many factors as you want depending on the information you need to be associated with the results later on.

```
sample_library,/TRAVIS/reo_full/reo_full_sample_library.csv
```

Required:

- 1st column has to be the filename of the sample
- 2nd column has to be a unique name or ID
- any number of columns containing any information

Table 30: Example of a sample library

lalala

filename	assembly_ID	order	family	genus_species	sample_location	sample_date
INSnfrTBERAAPEI-19.fasta	INSnfrTBERAAPEI-19	Coleoptera	Gyrinidae	Gyrinus_marinus	Hoehbeck_Pevestorf	11-Aug_2011

6.2.4.7 reference_library

Specifies the path to the reference library. As TRAVIS is relying on NCBI up to now, it is necessary to specify either accession numbers (separated by '&') in the reference library or the path to the respective assembly report on the NCBI-FTP server, if available. However, you as well need to specify the NT accession number and the PID of the 'main' gene. you can add as many factors as you want. these can be used for naming the references and sorting them into subgroups

```
reference_library,/TRAVIS/reo_full/reo_full_reference_library.csv
```

Required:

- 1st column has to be an acronym or ID
- 2nd column has to be a unique name

- any number of columns containing any information
- the last three columns have to be 'all_NT_ACC,<main>_NT_ACC,<main>_PID' where <main> can be replaced by a meaningful name

Table 31: Example of a reference library

Instead of providing single accession numbers, you can add the path to an NCBI assembly report (.txt), that you can get from `https://www.ncbi.nlm.nih.gov/assembly`.

Acronym	Name	Family	Genus	all_NT_ACC	RdRp_NT_ACC	RdRp_PID
APRV	Aedes_pseudoscutellaris_reovirus	Reoviridae	Dinovernavirus	/url/to/assembly_report*	NC_007667	YP_443936.1
AHRV	Atlantic_halibut_reovirus	Reoviridae	Aquareovirus	KJ499467&KJ499468&KJ913664	KJ499467	AIY69147.1

6.2.4.8 Local Reference Databases

Specifies the path to local reference databases. This database saves everything related to the reference library so you do not have to download everything from NCBI over and over again.

```
local_reference_database,/TRAVIS/reo_full/reo_local_reference_database.csv
reference_fastas,/TRAVIS/reo_full/references/
reference_gbx,/TRAVIS/reo_full/references/genebank/
```

6.2.4.9 header_names

Names of the columns that you want to drag through the whole analysis included in the header of the reference sequences. They have to be identical to the column names in your reference library.

```
header_names,Name&Genus
```

6.2.4.10 split_references

Names of the columns that you want to split the references by. So you can *e.g.* create subgroups by genus or family.

```
split_references,Genus&Family
```

6.2.4.11 sample_subset

If this is set to 'main_positive', only company sequences will be searched if main sequences were found in the respective sample. This can still be changed in TTT before running TRAVIS Core.

```
sample_subset,main_positive or all
```

6.2.4.12 result_dir

All relevant results will be stored here if not declared otherwise.

```
result_dir ,/TRAVIS/reo_full/
```

6.2.4.13 TTT

Specifies the path to the Troubling TRAVIS Table.

```
TTT,/TRAVIS/reo_full/reo_full_TTT.csv
```

6.2.4.14 nCPU

Specifies how many processors can be used.

```
nCPU,6
```

6.2.4.15 max_references

Limits how many references will be plotted in the Scavenger output.

```
max_references ,3
```

6.2.4.16 HMMER3

Specifies paths and settings of HMMER3.

```
hmmbuild ,/TRAVIS/travis_programs/hmmer-3.1b2/binaries/hmmbuild
hmmsearch ,/TRAVIS/travis_programs/hmmer-3.1b2/binaries/hmmsearch
hmmsearch_settings ,-E 1.00E-6
jackhmmer ,/TRAVIS/travis_programs/hmmer-3.1b2/binaries/jackhmmer
jackhmmer_settings ,-E 1.00E-6
```

6.2.4.17 MAFFT

Specifies paths and settings of MAFFT. In my experience, if you want to use a portable version of MAFFT, the proper $PATHs have to be configured. By specifying the location of the MAFFT_BINARIES, i could easily solve issues regarding that.

```
mafft ,/TRAVIS/travis_programs/mafft-7.302/mafftdir/bin/mafft
mafft_settings ,--maxiterate 1000 --genafpair --adjustdirection --reorder
mafft_binaries ,/TRAVIS/travis_programs/mafft-7.302/mafftdir/libexec/
```

6.2.4.18 MMSeqs2

Specifies paths and settings of MMSeqs2. 'minimal_cluster_size' is for the clustering of the company sequences by TRAVIS Henchman.

```
mmseqs,/TRAVIS/travis_programs/mmseqs2_SSE4/bin/mmseqs
mmseqs_cluster_settings,-c 0.01 -v 0 --cluster-mode 0 -s 7.5 --mask 0
mmseqs_search_settings,--max-seqs 10 -e 1.00E-6
minimal_cluster_size,2
```

6.2.4.19 BLASTP

Specifies paths and settings of BLASTP.

```
blastp,/TRAVIS/travis_programs/ncbi-blast-2.6.0+/bin/blastp
blastp_settings,-evalue 1.00E-6 -max_target_seqs 10
makeblastdb,/TRAVIS/travis_programs/ncbi-blast-2.6.0+/bin/makeblastdb
blastp_db,/TRAVIS/blast_DBs/nr
```

6.2.5 Troubling TRAVIS Table (TTT)

You can *e.g.*:

- switch off searches
- check and modify alignments that are the basis for hmmsearch
- change combination of search tools on certain groups/clusters
- add other proteins/groups/clusters to the 'main' pool

Table 32: Example of a TTT

type: main or company, sample_subset: all or main_positive

group_name	type	fasta_name	alignment_name	number_of_sequences	sample_subset	search_tools	run
RdRp_all	main_RdRp	RdRp_all.fasta	RdRp_all_aln.fasta	73	all	hmmer&jackhmmer&mmseqs&blastp	on

7 Acknowledgments

I would like to thank Prof. Dr. Bernhard Misof and Prof. Dr. Christian Drosten alike for giving me the opportunity to work with them on this project. They were always a reliable source of advice and motivation. I also thank Prof. Dr. Lukas Schreiber and Prof. Dr. Ullrich Wüllner for being the third and fourth assessor of this thesis.

Of all the great colleagues across the associated Institutes I want to highlight Dr. Florian Zirkel, Dr. Sandra Junglen and Dipl-Biol. Malte Petersen for being valuable teachers. MSc. Sofia Paraskevopoulou has my gratitude for all the work on post-evaluation we split and the mutual teaching.

I also would like to thank the DGF for funding this project.

Nevertheless important I thank my family, and especially Stefanie Bruhn, for the support over all these years.

8 References

Akiva, E., Brown, S., Almonacid, D. E., Barber 2nd, A. E., Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C. (2013). The structure–function linkage database. *Nucleic Acids Research*, 42(D1):D521–D530.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Anthony, S., Maan, N., Maan, S., Sutton, G., Attoui, H., & Mertens, P. (2009). Genetic and phylogenetic analysis of the core proteins VP1, VP3, VP4, VP6 and VP7 of *Epizootic haemorrhagic disease virus* (EHDV). *Virus Research*, 145(2):187–199.

Anthony, S. J., Islam, A., Johnson, C., Navarrete-Macias, I., Liang, E., Jain, K., Hitchens, P. L., Che, X., Soloyvov, A., Hicks, A. L., Ojeda-Flores, R., Zambrana-Torrelio, C., Ulrich, W., Rostal, M. K., Petrosov, A., Garcia, J., Haider, N., Wolfe, N., Goldstein, T., Morse, S. S., Rahman, M., Epstein, J. H., Mazet, J. K., Daszak, P., & Lipkin, W. I. (2015). Non-random patterns in viral diversity. *Nature Communications*, 6(1).

Anzola, J. V., Dall, D. J., Xu, Z., & Nuss, D. L. (1989). Complete nucleotide sequence of *Wound tumor virus* genomic segments encoding nonstructural polypeptides. *Virology*, 171(1):222–228.

Anzola, J. V., Xu, Z., Asamizu, T., & Nuss, D. L. (1987). Segment-specific inverted repeats found adjacent to conserved terminal sequences in *Wound tumor virus* genome and defective interfering RNAs. *Proceedings of the National Academy of Sciences*, 84(23):8301–8305.

Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., & Weightman, A. J. (2005). At least 1 in 20 16s rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, 71(12):7724–7736.

Attoui, H., Billoir, F., Biagini, P., de Micco, P., & de Lamballerie, X. (2000). Complete sequence determination and genetic analysis of *Banna virus* and *Kadipiro virus*: proposal for assignment to a new genus (*Seadornavirus*) within the family Reoviridae. *Journal of General Virology*, 81(6):1507–1515.

Attoui, H., Fang, Q., Jaafar, F. M., Cantaloube, J.-F., Biagini, P., de Micco, P., & de Lamballerie, X. (2002). Common evolutionary origin of aquareoviruses and orthoreoviruses revealed by genome characterization of *Golden shiner reovirus*, *Grass*

carp reovirus, *Striped bass reovirus* and *Golden ide reovirus* (genus *Aquareovirus*, family Reoviridae). *Journal of General Virology*, 83(8):1941–1951.

Attoui, H., Jaafar, F. M., Belhouchet, M., de Micco, P., de Lamballerie, X., & Brussaard, C. P. D. (2006a). *Micromonas pusilla reovirus*: a new member of the family Reoviridae assigned to a novel proposed genus (*Mimoreovirus*). *Journal of General Virology*, 87(5):1375–1383.

Attoui, H., Jaafar, F. M., Belhouchet, M., Tao, S., Chen, B., Liang, G., Tesh, R. B., de Micco, P., & de Lamballerie, X. (2006b). *Liao ning virus*, a new chinese *Seadornavirus* that replicates in transformed and embryonic mammalian cells. *Journal of General Virology*, 87(1):199–208.

Attoui, H., Mendez-lopez, M. R., Rao, S., Hurtado-Alendes, A., Lizaraso-Caparo, F., Jaafar, F. M., Samuel, A. R., Belhouchet, M., Pritchard, L. I., Melville, L., Weir, R. P., Hyatt, A. D., Davis, S. S., Lunt, R., Calisher, C. H., Tesh, R. B., Fujita, R., & Mertens, P. P. (2009). *Peruvian horse sickness virus* and *Yunnan orbivirus*, isolated from vertebrates and mosquitoes in Peru and Australia. *Virology*, 394(2):298–310.

Attoui, H., Mohd Jaafar, F., Belhouchet, M., Biagini, P., Cantaloube, J.-F., de Micco, P., & de Lamballerie, X. (2005). Expansion of family Reoviridae to include nine-segmented dsRNA viruses: Isolation and characterization of a new virus designated *Aedes pseudoscutellaris reovirus* assigned to a proposed genus (*Dinovernavirus*). *Virology*, 343(2):212–223.

Attoui, H., Stirling, J. M., Munderloh, U. G., & Burroughs, J. N. (2001). Complete sequence characterization of the genome of the *St.Croix River virus*, a new *Orbivirus* isolated from cells of *Ixodes scapularis*. *Journal of General Virology*, 82(4):795-804.

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., Roma-Mateo, C., Theodosiou, A., & Mitchell, A. L. (2012). The prints database: a fine-grained protein sequence annotation and analysis resourceits status in 2012. *Database*, 2012.

Baker, T., Olson, N., & Fuller, S. (1999). Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiology and Molecular Biology Reviews*, 63(4):862–922.

Ballinger, M. J., Bruenn, J. A., Hay, J., Czechowski, D., & Taylor, D. J. (2014). Discovery and evolution of bunyavirids in arctic phantom midges and ancient bunyavirid-like sequences in insect genomes. *Journal of Virology*, 88(16):8783–8794.

Bastkowski, S., Mapleson, D., Spillner, A., Wu, T., Balvociute, M., & Moulton, V. (2017). SPECTRE: a Suite of PhylogEnetiC Tools for Reticulate Evolution. *Bioinformatics*, 34(6):1056-1057.

Bateman, A. (2004). The Pfam protein families database. *Nucleic Acids Research*, 32(90001):138D–141.

Beatman, E. L., Massey, A., Shives, K. D., Burrack, K. S., Chamanian, M., Morrison, T. E., & Beckham, J. D. (2016). Alpha-synuclein expression restricts RNA viral infections in the brain. *Journal of Virology*, 90(6):2767–2782.

Bedarf, J. R., Hildebrand, F., Coelho, L. P., Sunagawa, S., Bahram, M., Goeser, F., Bork, P., & Wüllner, U. (2017). Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naïve Parkinsons disease patients. *Genome Medicine*, 9(1).

Belaganahalli, M. N., Maan, S., Maan, N. S., Nomikou, K., Guimera, M., Brownlie, J., Tesh, R., Attoui, H., & Mertens, P. P. C. (2013). Full genome sequencing of *Corriparta virus* identifies *California mosquito pool virus* as a member of the *Corriparta virus* species. *PLoS ONE*, 8(8):e70779.

Belaganahalli, M. N., Maan, S., Maan, N. S., Nomikou, K., Pritchard, I., Lunt, R., Kirkland, P. D., Attoui, H., Brownlie, J., & Mertens, P. P. C. (2012). Full genome sequencing and genetic characterization of *Eubenangee viruses* identify *Pata virus* as a distinct species within the genus *Orbivirus*. *PLoS ONE*, 7(3):e31911.

Belaganahalli, M. N., Maan, S., Maan, N. S., Pritchard, I., Kirkland, P. D., Brownlie, J., Attoui, H., & Mertens, P. P. C. (2014). Full genome characterization of the culicoides-borne marsupial orbiviruses: *Wallal virus*, *Mudjinbarry virus* and *Warrego viruses*. *PLoS ONE*, 9(10):e108379.

Benveniste, R. E. & Todaro, G. J. (1974). Evolution of c-type viral genes: inheritance of exogenously acquired viral genes. *Nature*, 252(5483):456.

Bhargava, P. & Mowry, E. M. (2014). Gut microbiome and multiple sclerosis. *Current Neurology and Neuroscience Reports*, 14(10):492.

Bibby, K. (2013). Metagenomic identification of viral pathogens. *Trends in Biotechnology*, 31(5):275–279.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., *et al.* (2009). *De novo* transcriptome assembly with abyss. *Bioinformatics*, 25(21):2872–2877.

Bordewich, M. & Semple, C. (2005). On the computational complexity of the rooted subtree prune and regraft distance. *Annals of Combinatorics*, 8(4):409–423.

Brahic, M. (2010). Multiple sclerosis and viruses. *Annals of Neurology*, 68(1):6–8.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The prodom database of protein domain families: more emphasis on 3D. *Nucleic Acids Research*, 33(suppl_1):D212–D215.

Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using diamond. *Nature Methods*, 12(1):59.

Bundesregierung (2017). Antwort der Bundesregierung auf die kleine Anfrage der Abgeordneten Steffi Lemke, Harald Ebner, Bärbel Höhn, weiterer Abgeordneter und der Fraktion Bündnis 90/die Grünen: Insekten in Deutschland und Auswirkungen ihres Rückgangs. `http://dip21.bundestag.de/dip21/btd/18/131/1813142.pdf/`. [Drucksache 18/12859, Online; accessed 08 June 2018].

Burke, G. R. (2016). Analysis of genetic variation across the encapsidated genome of *Microplitis demolitor bracovirus* in parasitoid wasps. *PLOS ONE*, 11(7):e0158846.

Burke, G. R., Walden, K. K. O., Whitfield, J. B., Robertson, H. M., & Strand, M. R. (2014). Widespread genome reorganization of an obligate virus mutualist. *PLoS Genetics*, 10(9):e1004660.

Bányai, K., Borzák, R., Ihász, K., Fehér, E., Dán, î, Jakab, F., Papp, T., Hetzel, U., Marschang, R. E., & Farkas, S. L. (2014). Whole-genome sequencing of a *Green bush viper reovirus* reveals a shared evolutionary history between reptilian and unusual mammalian orthoreoviruses. *Archives of Virology*, 159(1):153–158.

Bányai, K., Dandár, E., Dorsey, K. M., Mató, T., & Palya, V. (2011). The genomic constellation of a novel avian *Orthoreovirus* strain associated with runting-stunting syndrome in broilers. *Virus Genes*, 42(1):82–89.

Béliveau, C., Cohen, A., Stewart, D., Periquet, G., Djoumad, A., Kuhn, L., Stoltz, D., Boyle, B., Volkoff, A.-N., Herniou, E. A., Drezen, J.-M., & Cusson, M. (2015). Genomic and proteomic analyses indicate that *Banchine* and *Campoplegine polydnaviruses* have similar, if not identical, viral ancestors. *Journal of Virology*, 89(17):8909–8921.

Calisher, C. & Mertens, P. (1998). Taxonomy of *African horse sickness viruses*. In: *African Horse Sickness*, pages 3–11. Springer Verlag.

Calisher, C. H. & Tesh, R. B. (2014). Two misleading words in reports of virus discovery: little things mean a lot. *Archives of virology*, 159(8):2189–2191.

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

Carrillo-Tripp, M., Shepherd, C. M., Borelli, I. A., Venkataraman, S., Lander, G., Natarajan, P., Johnson, J. E., Brooks, C. L., & Reddy, V. S. (2009). VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Research*, 37(Database):D436–D442.

Chen, H., Smith, G., Li, K., Wang, J., Fan, X., Rayner, J., Vijaykrishna, D., Zhang, J., Zhang, L., Guo, C. (2006). Establishment of multiple sublineages of H5N1 *Influenza virus* in asia: implications for pandemic control. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8):2845–2850.

Coffey, L. L., Page, B. L., Greninger, A. L., Herring, B. L., Russell, R. C., Doggett, S. L., Haniotis, J., Wang, C., Deng, X., & Delwart, E. L. (2014). Enhanced arbovirus surveillance with deep sequencing: identification of novel *Rhabdoviruses* and *Bunyaviruses* in australian mosquitoes. *Virology*, 448:146–158.

Cook, S., Chung, B. Y.-W., Bass, D., Moureau, G., Tang, S., McAlister, E., Culverwell, C. L., Glücksman, E., Wang, H., Brown, T. D. K., Gould, E. A., Harbach, R. E., de Lamballerie, X., & Firth, A. E. (2013). Novel virus discovery and genome reconstruction from field RNA samples reveals highly divergent viruses in dipteran hosts. *PLoS ONE*, 8(11):e80720.

Crick, F. H. (1968). The origin of the genetic code. *Journal of Molecular Biology*, 38(3):367–379.

Dang, C., Le, Q., Gascuel, O., & Le, V. (2010). FLU, an amino acid substitution model for *Influenza* proteins. *BMC Evolutionary Biology*, 10(1):99.

Davison, A. J., Siddell, S., Mushegian, A., King, A. M. Q., Lefkowitz, E. J., Harrach, B., Kuhn, J. H., Knowles, N. J., Kropinski, A., Simmonds, P., Zerbini, F. M., Dutilh, B., Harrison, R., Junglen, S., Krupovic, M., Nibert, M. L., Rubino, L., Sabanadzovic, S., & Varsani, A. (2017). *Virus Taxonomy: The classification and nomenclature of viruses: the online (10th) report*. International Committee on Taxonomy of Viruses.

de Wit, E. & Munster, V. J. (2013). MERS-Cov: the intermediate host identified? *The Lancet. Infectious Diseases*, 13(10):827.

Deng, X.-X., Lü, L., Ou, Y.-J., Su, H.-J., Li, G., Guo, Z.-X., Zhang, R., Zheng, P.-R., Chen, Y.-G., He, J.-G., & Weng, S.-P. (2012). Sequence analysis of 12 genome segments of *Mud crab reovirus* (MCRV). *Virology*, 422(2):185–194.

Desper, R. & Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*, 21(3):587–598.

Deter, A. (2017). Bundesregierung bestaetigt Insektensterben durch intensive Landwirtschaft. `https://www.topagrar.com/news/Home-top-News-Bundesregierung-bestaetigt-Insektensterben-durch-intensive-Landwirtschaft-8427172.html/`. Online; accessed 08 June 2018.

Dimmic, M. W., Rest, J. S., Mindell, D. P., & Goldstein, R. A. (2002). rtREV: An amino acid substitution matrix for inference of *Retrovirus* and reverse transcriptase phylogeny. *Journal of Molecular Evolution*, 55(1):65–73.

Dinh, P. N., Long, H. T., Tien, N. T. K., Hien, N. T., Mai, L. T. Q., Phong, L. H., Van Tuan, L., Van Tan, H., Nguyen, N. B., Van Tu, P., *et al*. (2006). Risk factors for human infection with avian *Influenza A* H5N1, Vietnam, 2004. *Emerging Infectious Diseases*, 12(12):1841.

Distéfano, A. J., Conci, L. R., Muñoz Hidalgo, M., Guzmán, F. A., Hopp, H. E., & del Vas, M. (2003). Sequence and phylogenetic analysis of genome segments S1, S2, S3 and S6 of *Mal de rio cuarto virus*, a newly accepted *Fijivirus* species. *Virus research*, 92(1):113–121.

Domingo, E. & Holland, J. (1997). RNA virus mutations and fitness for survival. *Annual Reviews in Microbiology*, 51(1):151–178.

Duncan, R., Corcoran, J., Shou, J., & Stoltz, D. (2004). *Reptilian reovirus*: a new fusogenic *Orthoreovirus* species. *Virology*, 319(1):131–140.

Dunham, E. J. & Holmes, E. C. (2007). Inferring the timescale of *Dengue virus* evolution under realistic models of DNA substitution. *Journal of Molecular Evolution*, 64(6):656–661.

Dunjko, V. & Briegel, H. J. (2018). Machine learning & artificial intelligence in the quantum domain: a review of recent progress. *Reports on Progress in Physics*, 81.7 (2018): 074001.

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics*, 14(9):755–763.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10):e1002195.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200.

Eriksson, N., Pachter, L., Mitsuya, Y., Rhee, S.-Y., Wang, C., Gharizadeh, B., Ronaghi, M., Shafer, R. W., & Beerenwinkel, N. (2008). Viral population estimation using pyrosequencing. *PLoS Computational Biology*, 4(5):e1000074.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376.

Fields, B., Knipe, D., Howley, P., & Griffin, D. (2007). *Fields Virology. 5th Edition*. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Fife, D. (2017). fifer: A biostatisticians toolbox for various activities, including plotting, data cleanup, and data analysis. R package version 1.1.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., *et al*. (2016). InterPro in 2017: beyond protein family and domain annotations. *Nucleic Acids Research*, 45(D1):D190–D199.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., *et al*. (2015). The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285.

Floudas, C., Fung, H., McAllister, S., Mönnigmann, M., & Rajgaria, R. (2006). Advances in protein structure prediction and *de novo* protein design: A review. *Chemical Engineering Science*, 61(3):966–988.

Forterre, P. & Gaïa, M. (2016). Giant viruses and the origin of modern eukaryotes. *Current Opinion in Microbiology*, 31:44–49.

Freitas, T. A. K., Li, P.-E., Scholz, M. B., & Chain, P. S. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. *Nucleic Acids Research*, 43(10):e69–e69.

Fricke, W. F., Rasko, D. A., & Ravel, J. (2009). The role of genomics in the identification, prediction, and prevention of biological threats. *PLoS Biology*, 7(10):e1000217.

Fullwood, M. J., Wei, C.-L., Liu, E. T., & Ruan, Y. (2009). Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research*, 19(4):521–532.

Gaboriaud, C., Bissery, V., Benchetrit, T., & Mornon, J. (1987). Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Letters*, 224(1):149–155.

Gao, R., Cao, B., Hu, Y., Feng, Z., Wang, D., Hu, W., Chen, J., Jie, Z., Qiu, H., Xu, K., *et al*. (2013). Human infection with a novel avian-origin *Influenza A* (H7N9) virus. *New England Journal of Medicine*, 368(20):1888–1897.

Girard, M. P., Tam, J. S., Assossou, O. M., & Kieny, M. P. (2010). The 2009 A (H1N1 *Influenza virus* pandemic: A review. *Vaccine*, 28(31):4895–4902.

Gogarten, J. P. & Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679.

Goodier, J. L. & Kazazian Jr, H. H. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell*, 135(1):23–35.

Graham, R. I., Rao, S., Possee, R. D., Sait, S. M., Mertens, P. P., & Hails, R. S. (2006). Detection and characterisation of three novel species of reovirus (Reoviridae), isolated from geographically separate populations of the winter moth *Operophtera brumata* (Lepidoptera: Geometridae) on Orkney. *Journal of Invertebrate Pathology*, 91(2):79–87.

Green, E. D. & Guyer, M. S. (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature*, 470(7333):204–213.

Guindon, S., Delsuc, F., Dufayard, J.-F., & Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. In: *Bioinformatics for DNA sequence analysis*, pages 113–137. Springer.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3):307–321.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2012). Tigrfams and genome properties in 2013. *Nucleic Acids Research*, 41(D1):D387–D395.

Hales, S., De Wet, N., Maindonald, J., & Woodward, A. (2002). Potential effect of population and climate changes on global distribution of dengue fever: an empirical model. *The Lancet*, 360(9336):830–834.

Halloran, A., Vantomme, P., Hanboonsong, Y., & Ekesi, S. (2015). Regulating edible insects: the challenge of addressing food security, nature conservation, and the erosion of traditional food culture. *Food Security*, 7(3):739–746.

Hancock, K., Veguilla, V., Lu, X., Zhong, W., Butler, E. N., Sun, H., Liu, F., Dong, L., DeVos, J. R., Gargiullo, P. M., *et al.* (2009). Cross-reactive antibody responses to the 2009 pandemic H1N1 *Influenza virus*. *New England journal of medicine*, 361(20):1945–1952.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., & VandePol, S. (1982). Rapid evolution of RNA genomes. *Science*, 215(4540):1577–1585.

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host & Microbe*, 10(4):368–377.

Holmes, E. C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St George, K., Grenfell, B. T., Salzberg, S. L., Fraser, C. M., Lipman, D. J. (2005). Whole-genome analysis of human *Influenza A virus* reveals multiple persistent lineages and reassortment among recent H3N2 viruses. *PLoS Biology*, 3(9):e300.

Hulo, N. (2006). The PROSITE database. *Nucleic Acids Research*, 34(90001):D227–D230.

Huson, D. H. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267.

Iranzo, J., Krupovic, M., & Koonin, E. V. (2017). A network perspective on the virus world. *Communicative & Integrative Biology*, 10(2):e1296614.

Jaenike, J. (2012). Population genetics of beneficial heritable symbionts. *Trends in Ecology & Evolution*, 27(4):226–232.

Jagadish, H., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94.

Jiang, B. D. X. H. T., Li, M., Tromp, J., & Zhang, L. (2000). On computing the nearest neighbor interchange distance. In: *Discrete Mathematical Problems with Medical Applications: DIMACS Workshop Discrete Mathematical Problems with Medical*

Applications, December 8-10, 1999, DIMACS Center, volume 55, pages 125. American Mathematical Soc.

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, 11(1):431.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, 451(7181):990.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., *et al*. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

Junglen, S. (2016). Evolutionary origin of pathogenic arthropod-borne viruses: a case study in the family Bunyaviridae. *Current Opinion in Insect Science*, 16:81–86.

Junglen, S. & Drosten, C. (2013). Virus discovery and recent insights into virus diversity in arthropods. *Current Opinion in Microbiology*, 16(4):507–513.

Junier, T. & Zdobnov, E. M. (2010). The newick utilities: high-throughput phylogenetic tree processing in the unix shell. *Bioinformatics*, 26(13):1669–1670.

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066.

Katzourakis, A. & Gifford, R. J. (2010). Endogenous viral elements in animal genomes. *PLoS Genetics*, 6(11):e1001191.

Kilpatrick, A. M., Kramer, L. D., Jones, M. J., Marra, P. P., & Daszak, P. (2006). *West nile virus* epidemics in North America are driven by shifts in mosquito feeding behavior. *PLoS biology*, 4(4):e82.

Koonin, E. V., Dolja, V. V., & Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology*, 479:2–25.

Koonin, E. V. & Novozhilov, A. S. (2009). Origin and evolution of the genetic code: the universal enigma. *IUBMB life*, 61(2):99–111.

Kück, P. & Longo, G. C. (2014). Fasconcat-g: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology*, 11(1):81.

Kück, P., Wilkinson, M., Gross, C., Foster, P. G., & Wägele, J. W. (2017). Can quartet analyses combining maximum likelihood estimation and Hennigian logic overcome long branch attraction in phylogenomic sequence data? *PLoS ONE*, 12(8):e0183393.

Kück, P. & Wägele, J. W. (2016). Plesiomorphic character states cause systematic errors in molecular phylogenetic analyses: a simulation study. *Cladistics*, 32(4):461–478.

Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2015). Denoising DNA deep sequencing data: high-throughput sequencing errors and their correction. *Briefings in Bioinformatics*, 17(1):154–179.

Lam, S. D., Dawson, N. L., Das, S., Sillitoe, I., Ashford, P., Lee, D., Lehtinen, S., Orengo, C. A., & Lees, J. G. (2015). Gene3d: expanding the utility of domain assignments. *Nucleic Acids Research*, 44(D1):D404–D409.

Lander, E., Linton, L., Birren, B., Nusbaum, C., Zody, M., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al*. (2001). Initial sequencing and analysis of the human genome, international human genome sequencing consortium.[erratum to document cited in ca134: 217890]. *Nature (London, UK)*, 412:565–566.

Leather, S. (2018). Ecological armageddon–more evidence for the drastic decline in insect numbers. *Annals of Applied Biology*, 172(1):1–3.

Lee, N., Hui, D., Wu, A., Chan, P., Cameron, P., Joynt, G. M., Ahuja, A., Yung, M. Y., Leung, C., To, K., *et al*. (2003). A major outbreak of severe acute respiratory syndrome in Hong Kong. *New England Journal of Medicine*, 348(20):1986–1994.

Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution*, 32(10):2798–2800.

Lemoine, F., Entfellner, J.-B. D., Wilkinson, E., Correia, D., Felipe, M. D., Oliveira, T., & Gascuel, O. (2018). Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, 556(7702):452.

Letunic, I., Doerks, T., & Bork, P. (2014). Smart: recent updates, new developments and status in 2015. *Nucleic Acids Research*, 43(D1):D257–D260.

Li, C.-X., Shi, M., Tian, J.-H., Lin, X.-D., Kang, Y.-J., Chen, L.-J., Qin, X.-C., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2015). Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses. *eLife*, 4.

Li, Y., Wang, H., Nie, K., Zhang, C., Zhang, Y., Wang, J., Niu, P., & Ma, X. (2016). VIP: an integrated pipeline for metagenomics of virus identification and discovery. *Scientific reports*, 6(1).

Lin, H.-H. & Liao, Y.-C. (2017). drVM: a new tool for efficient genome assembly of known eukaryotic viruses from metagenomes. *GigaScience*, 6(2):1–10.

Longo, M. S., O'Neill, M. J., & O'Neill, R. J. (2011). Abundant human DNA contamination identified in non-primate genome databases. *PLoS ONE*, 6(2):e16410.

Madera, M. & Gough, J. (2002). A comparison of profile hidden markov model procedures for remote homology detection. *Nucleic Acids Research*, 30(19):4321–4328.

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I. (2014). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, 43(D1):D222–D226.

Massey, A. R. & Beckham, J. D. (2016). Alpha-synuclein, a novel viral restriction factor hiding in plain sight. *DNA and Cell Biology*, 35(11):643–645.

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2015). Panther version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Research*, 44(D1):D336–D342.

Misof, B., Liu, S., Meusemann, K., Peters, R. S., Donath, A., Mayer, C., Frandsen, P. B., Ware, J., Flouri, T., Beutel, R. G., Niehuis, O., Petersen, M., Izquierdo-Carrasco, F., Wappler, T., Rust, J., Aberer, A. J., Aspock, U., Aspock, H., Bartel, D., Blanke, A., Berger, S., Bohm, A., Buckley, T. R., Calcott, B., Chen, J., Friedrich, F., Fukui, M., Fujita, M., Greve, C., Grobe, P., Gu, S., Huang, Y., Jermiin, L. S., Kawahara, A. Y., Krogmann, L., Kubiak, M., Lanfear, R., Letsch, H., Li, Y., Li, Z., Li, J., Lu, H., Machida, R., Mashimo, Y., Kapli, P., McKenna, D. D., Meng, G., Nakagaki, Y., Navarrete-Heredia, J. L., Ott, M., Ou, Y., Pass, G., Podsiadlowski, L., Pohl, H., von Reumont, B. M., Schutte, K., Sekiya, K., Shimizu, S., Slipinski, A., Stamatakis, A., Song, W., Su, X., Szucsich, N. U., Tan, M., Tan, X., Tang, M., Tang, J., Timelthaler, G., Tomizuka, S., Trautwein, M., Tong, X., Uchifune, T., Walzl, M. G., Wiegmann, B. M., Wilbrandt, J., Wipfler, B., Wong, T. K. F., Wu, Q., Wu, G., Xie, Y., Yang, S., Yang, Q., Yeates, D. K., Yoshizawa, K., Zhang, Q., Zhang, R., Zhang, W., Zhang, Y., Zhao, J., Zhou, C., Zhou, L., Ziesmann, T., Zou, S., Li, Y., Xu, X., Zhang, Y., Yang, H., Wang, J., Wang, J., Kjer, K. M., & Zhou, X. (2014b). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346(6210):763–767.

Modrow, S., Falke, D., Truyen, U., & Schätzl, H. (2010). *Molekulare Virologie*. Spektrum Akademischer Verlag.

Mokili, J. L., Rohwer, F., & Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology*, 2(1):63–77.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., & Worm, B. (2011). How many species are there on earth and in the ocean? *PLoS Biology*, 9(8):e1001127.

Moriyasu, Y., Maruyama-Funatsuki, W., Kikuchi, A., Ichimi, K., Zhong, B., Yan, J., Zhu, Y., Suga, H., Watanabe, Y., Ichiki-Uehara, T., Shimizu, T., Hagiwara, K., Kamiunten, H., Akutsu, K., & Omura, T. (2007). Molecular analysis of the genome segments S1, S4, S6, S7 and S12 of a rice gall dwarf virus isolate from Thailand; completion of the genomic sequence. *Archives of Virology*, 152(7):1315–1322.

Morse, S. S., Mazet, J. A., Woolhouse, M., Parrish, C. R., Carroll, D., Karesh, W. B., Zambrana-Torrelio, C., Lipkin, W. I., & Daszak, P. (2012). Prediction and prevention of the next pandemic zoonosis. *The Lancet*, 380(9857):1956–1965.

Munang'andu, H. M., Mugimba, K. K., Byarugaba, D. K., Mutoloki, S., & Evensen, O. (2017). Current advances on virus discovery and diagnostic role of viral metagenomics in aquatic organisms. *Frontiers in Microbiology*, 8.

Nakada, S., Creager, R., Krystal, M., Aaronson, R., & Palese, P. (1984). *Influenza C virus* hemagglutinin: comparison with *influenza A* and *B virus* hemagglutinins. *Journal of Virology*, 50(1):118–124.

Nakashima, N., Koizumi, M., Watanabe, H., & Noda, H. (1996). Complete nucleotide sequence of the *Nilaparvata lugens reovirus*: a putative member of the genus *Fijivirus*. *Journal of General Virology*,77(1):139–146.

NCBICoordinators (2016). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 44(Database issue):D7.

Neumann, G., Noda, T., & Kawaoka, Y. (2009). Emergence and pandemic potential of swine-origin H1N1 *Influenza virus*. *Nature*, 459(7249):931.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217.

Nouri, S., Salem, N., Nigg, J. C., & Falk, B. W. (2015). A diverse array of new viral sequences identified in worldwide populations of the asian citrus psyllid (*Diaphorina citri*) using viral metagenomics. *Journal of Virology*, pages JVI–02793.

Nunes, M. R., Contreras-Gutierrez, M. A., Guzman, H., Martins, L. C., Barbirato, M. F., Savit, C., Balta, V., Uribe, S., Vivero, R., Suaza, J. D., *et al.* (2017). Genetic characterization, molecular epidemiology, and phylogenetic relationships of insect-specific viruses in the taxon negevirus. *Virology*, 504:152–167.

Oates, M. E., Stahlhacke, J., Vavoulis, D. V., Smithers, B., Rackham, O. J., Sardar, A. J., Zaucha, J., Thurlby, N., Fang, H., & Gough, J. (2014). The superfamily 1.75 database in 2014: a doubling of data. *Nucleic Acids Research*, 43(D1):D227–D233.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290.

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., De Castro, E., Baratin, D., Cuche, B. A., Bougueleret, L., Poux, S., *et al*. (2014). Hamap in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Research*, 43(D1):D1064–D1070.

Peiris, J. S., Yuen, K. Y., Osterhaus, A. D., & Stöhr, K. (2003). The severe acute respiratory syndrome. *New England Journal of Medicine*, 349(25):2431–2441.

Pellmyr, O. (1992). Evolution of insect pollination and angiosperm diversification. *Trends in Ecology & Evolution*, 7(2):46–49.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., & Claverie, J.-M. (2004). The 1.2-megabase genome sequence of mimivirus. *Science*, 306(5700):1344–1350.

Reeck, G. R., De Haen, C., Teller, D. C., Doolittle, R. F., Fitch, W. M., Dickerson, R. E., Chambon, P., McLachlan, A. D., Margoliash, E., Jukes, T. H.(1987). Homology in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell*, 50(5):667.

Reis-Filho, J. S. (2009). Next-generation sequencing. *Breast Cancer Research*, 11(3):S12.

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. *Trends in Genetics*, 16(6): 276-277.

Richards, F. M. (1977). Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering*, 6(1):151–176.

Robinson, G. E., Hackett, K. J., Purcell-Miramontes, M., Brown, S. J., Evans, J. D., Goldsmith, M. R., Lawson, D., Okamuro, J., Robertson, H. M., & Schneider, D. J. (2011). Creating a buzz about insect genomes. *Science*, 331(6023):1386–1386.

Rosani, U. & Gerdol, M. (2017). A bioinformatics approach reveals seven nearly-complete RNA-virus genomes in bivalve RNA-seq data. *Virus Research*, 239:33–42.

Rosario, K. & Breitbart, M. (2011). Exploring the viral world through metagenomics. *Current Opinion in Virology*, 1(4):289–297.

Rosenberg, D. M., Danks, H., & Lehmkuhl, D. M. (1986). Importance of insects in environmental impact assessment. *Environmental Management*, 10(6):773–783.

Ross, J. (1995). mRNA stability in mammalian cells. *Microbiological Reviews*, 59(3):423–450.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425.

Samways, M. J. (1993). Insects in biodiversity conservation: some perspectives and directives. *Biodiversity & Conservation*, 2(3):258–282.

Sand, A., Holt, M. K., Johansen, J., Brodal, G. S., Mailund, T., & Pedersen, C. N. (2014). tqdist: a library for computing the quartet and triplet distances between binary or general trees. *Bioinformatics*, 30(14):2079–2080.

Sayers, E. (2010). A general introduction to the e-utilities. *Entrez Programming Utilities Help [Internet]. Bethesda: National Center for Biotechnology Information*.

Schlee, M., Roth, A., Hornung, V., Hagmann, C. A., Wimmenauer, V., Barchet, W., Coch, C., Janke, M., Mihailovic, A., Wardle, G., *et al*. (2009). Recognition of 5 triphosphate by rig-i helicase requires short blunt double-stranded RNA as contained in panhandle of negative-strand virus. *Immunity*, 31(1):25–34.

Shen, H., Ma, Y., & Hu, Y. (2015). Near-full-length genome sequence of a novel *Reovirus* from the chinese mitten crab, *Eriocheir sinensis*. *Genome Announcements*, 3(3):e00447–15.

Shen, X.-X., Hittinger, C. T., & Rokas, A. (2017). Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution*, 1(5).

Shi, M., Lin, X.-D., Tian, J.-H., Chen, L.-J., Chen, X., Li, C.-X., Qin, X.-C., Li, J., Cao, J.-P., Eden, J.-S., Buchmann, J., Wang, W., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2016a). Redefining the invertebrate RNA virosphere. *Nature*, 540(7634):539–543.

Shi, M., Lin, X.-D., Vasilakis, N., Tian, J.-H., Li, C.-X., Chen, L.-J., Eastwood, G., Diao, X.-N., Chen, M.-H., Chen, X., Qin, X.-C., Widen, S. G., Wood, T. G., Tesh, R. B., Xu, J., Holmes, E. C., & Zhang, Y.-Z. (2016b). Divergent viruses discovered in arthropods and vertebrates revise the evolutionary history of the Flaviviridae and related viruses. *Journal of Virology*, 90(2):659–669.

Sigrist, C. J., De Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., & Xenarios, I. (2012). New and continuing developments at prosite. *Nucleic Acids Research*, 41(D1):D344–D347.

Silva, S. P., Dilcher, M., Weidmann, M., Carvalho, V. L., Casseb, A. R., Silva, E. V. P., Nunes, K. N. B., Chiang, J. O., Martins, L. C., Vasconcelos, P. F. C., & Nunes, M. R. T. (2013). Changuinola virus serogroup, new genomes within the genus *Orbivirus* (family Reoviridae) isolated in the brazilian amazon region. *Genome Announcements*, 1(6):e00940–13–e00940–13.

Skewes-Cox, P., Sharpton, T. J., Pollard, K. S., & DeRisi, J. L. (2014). Profile hidden markov models for the detection of viruses within metagenomic sequence data. *PLoS One*, 9(8):e105067.

Slater, G. S. C. & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6(1):31.

Small, C., Barro, M., Brown, T. L., & Patton, J. T. (2007). Genome heterogeneity of SA11 *Rotavirus* due to reassortment with O agent. *Virology*, 359(2):415–424.

Sonnhammer, E. L. L., von Heijne, G., & Krogh, A. (1998). A hidden markov model for predicting transmembrane helices in protein sequences. *Ismb*, 6(1):175–182.

Steinegger, M. & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big data: astronomical or genomical? *PLOS Biology*, 13(7):e1002195.

Stevens, P. F. (1984). Homology and phylogeny: morphology and systematics. *Systematic Botany*, pages 395–409.

Suyama, M., Torrents, D., & Bork, P. (2006). Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34(suppl_2):W609–W612.

Takahashi, K. & Nei, M. (2000). Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. *Molecular Biology and Evolution*, 17(8):1251–1258.

Taniguchi, S., Maeda, K., Horimoto, T., Masangkay, J. S., Puentespina, R., Alvarez, J., Eres, E., Cosico, E., Nagata, N., Egawa, K., Singh, H., Fukuma, A., Yoshikawa, T., Tani, H., Fukushi, S., Tsuchiaka, S., Omatsu, T., Mizutani, T., Une, Y., Yoshikawa, Y., Shimojima, M., Saijo, M., & Kyuwa, S. (2017). First isolation and characterization of pteropine *Orthoreoviruses* in fruit bats in the Philippines. *Archives of Virology*, 162(6):1529–1539.

Tokarz, R., Williams, S. H., Sameroff, S., Leon, M. S., Jain, K., & Lipkin, W. I. (2014). Virome analysis of *Amblyomma americanum*, *Dermacentor variabilis*, and *Ixodes scapularis* ticks reveals novel highly divergent vertebrate and invertebrate viruses. *Journal of Virology*, 88(19):11480–11492.

Upadhyaya, N. M., Ramm, K., Gellatly, J. A., Li, Z., Kositratana, W., & Waterhouse, P. M. (1998). *Rice ragged stunt oryzavirus* genome segment S4 could encode an RNA dependent RNA polymerase and a second protein of unknown function. *Archives of Virology*, 143(9):1815–1822.

Vasilakis, N., Forrester, N. L., Palacios, G., Nasar, F., Savji, N., Rossi, S. L., Guzman, H., Wood, T. G., Popov, V., Gorchakov, R., (2013). *Negevirus*: a proposed new taxon of insect-specific viruses with wide geographic distribution. *Journal of Virology*, 87(5):2475–2488.

von Bonsdorff, C. H. & Maunula, L. (1998). Short sequences define genetic lineages: phylogenetic analysis of group A *Rotaviruses* based on partial sequences of genome segments 4 and 9. *Journal of General Virology*, 79(2):321–332.

Wang, G. C. & Wang, Y. (1996). The frequency of chimeric molecules as a consequence of PCR co-amplification of 16s rRNA genes from different bacterial species. *Microbiology*, 142(5):1107–1114.

Wang, Q., Jia, P., & Zhao, Z. (2013). VirusFinder: Software for efficient and accurate detection of viruses and their integration sites in host genomes through next generation sequencing data. *PLoS ONE*, 8(5):e64465.

Watson, J. (1990). The human genome project: past, present, and future. *Science*, 248(4951):44–49.

Watson, J. D. & Crick, F. H. (1953). The structure of DNA. In: *Cold Spring Harbor symposia on quantitative biology*, volume 18, pages 123–131. Cold Spring Harbor Laboratory Press.

Weinbauer, M. G. & Rassoulzadegan, F. (2004). Are viruses driving microbial diversification and diversity? *Environmental Microbiology*, 6(1):1–11.

Wessner, D. R. (2010). The origins of viruses. *Virology*, 3(9):37.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkinson, M. (1996). Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology and Evolution*, 13(3):437–444.

Wood, D. E. & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46.

Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331.

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L.-S. L., Natale, D. A., Vinayaka, C., Hu, Z.-Z., Mazumder, R., Kumar, S., Kourtesis, P. (2004). PIRSF: family classification system at the protein information resource. *Nucleic Acids Research*, 32(suppl_1):D112–D114.

Xiong, M., Zhao, Z., Arnold, J., & Yu, F. (2011). Next-generation sequencing. *Journal of BioMed Research*, 2010.

Xu, Z., Choi, J., Lu, W., & Ou, J.-h. (2003). *Hepatitis C virus* F protein is a short-lived protein associated with the endoplasmic reticulum. *Journal of Virology*, 77(2):1578–1583.

Yang, X., Charlebois, P., Gnerre, S., Coole, M. G., Lennon, N. J., Levin, J. Z., Qu, J., Ryan, E. M., Zody, M. C., & Henn, M. R. (2012). *De novo* assembly of highly diverse viral populations. *BMC Genomics*, 13(1):475.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2016). ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8(1):28–36.

Zdobnov, E. M. & Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848.

Zhang, G. (2015). Genomics: Bird sequencing project takes off. *Nature*, 522(7554):34.

Zhao, G., Krishnamurthy, S., Cai, Z., Popov, V. L., Travassos da Rosa, A. P., Guzman, H., Cao, S., Virgin, H. W., Tesh, R. B., & Wang, D. (2013). Identification of novel viruses using VirusHunter – an automated data analysis pipeline. *PLoS ONE*, 8(10):e78470.

Zhao, G., Wu, G., Lim, E. S., Droit, L., Krishnamurthy, S., Barouch, D. H., Virgin, H. W., & Wang, D. (2017). VirusSeeker, a computational pipeline for virus discovery and virome composition analysis. *Virology*, 503:21–30.

Zhao, S., Liang, C., Hong, J., & Peng, H. (2003a). Genomic sequence analyses of segments 1 to 6 of *Dendrolimus punctatus cytoplasmic polyhedrosis virus*. *Archives of Virology*, 148(7):1357–1368.

Zhao, S., Liang, C., Hong, J., Xu, H., & Peng, H. (2003b). Molecular characterization of segments 7–10 of *Dendrolimus punctatus cytoplasmic polyhedrosis virus* provides the complete genome. *Virus Research*, 94(1):17–23.

Zhaxybayeva, O., Gogarten, J. P., Charlebois, R. L., Doolittle, W. F., & Papke, R. T. (2006). Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Research*, 16(9):1099–1108.

Zheng, Y., Gao, S., Padmanabhan, C., Li, R., Galvez, M., Gutierrez, D., Fuentes, S., Ling, K.-S., Kreuze, J., & Fei, Z. (2017). VirusDetect: An automated pipeline for efficient virus discovery using deep sequencing of small RNAs. *Virology*, 500:130–138.

Zhou, C., Liu, S., Song, W., Luo, S., Meng, G., Yang, C., Yang, H., Ma, J., Wang, L., Gao, S., Wang, J., Yang, H., Zhao, Y., Wang, H., & Zhou, X. (2018). Characterization of viral RNA splicing using whole-transcriptome datasets from host species. *Scientific Reports*, 8(1).

Zhou, N. N., Senne, D. A., Landgraf, J. S., Swenson, S. L., Erickson, G., Rossow, K., Liu, L., Yoon, K.-j., Krauss, S., & Webster, R. G. (1999). Genetic reassortment of avian, swine, and human *Influenza A viruses* in american pigs. *Journal of Virology*, 73(10):8851–8856.

Zimorski, V., Ku, C., Martin, W. F., & Gould, S. B. (2014). Endosymbiotic theory for organelle origins. *Current Opinion in Microbiology*, 22:38–48.

Zouache, K., Michelland, R. J., Failloux, A.-B., Grundmann, G. L., & Mavingui, P. (2012). *Chikungunya virus* impacts the diversity of symbiotic bacteria in mosquito vector. *Molecular Ecology*, 21(9):2297–2309.