Tempo and Mode in the Molecular Evolution of Viruses

A Taxonomy of Quantitative Differences in Parameters of Evolutionary Variation across Viruses

PhD Thesis

Hanouf Mohammed Alhamdan, BSc, MSc

Student number: 35392307

Faculty of Health and Medicine

Department of Biomedical and Life science

August 2025

I, Hanouf Mohammed Alhamdan, confirm that the work presented in this thesis is my own and has not been submitted in substantially the same form for the award of a higher degree elsewhere. Where information has been derived from other sources, I confirm this has been indicated in the thesis.

Submitted in part fulfilment of the requirements for the degree of Doctorate of Philosophy

Abstract

Understanding evolution in viruses is vital to identify factors that drive their emergence, adaptability, and pathogenicity. This project focuses on studying tempo (substitution rate with clock behaviour) and mode (selection pressure) of viral evolution across a wide range of viral taxonomy, particularly at the family and genus levels. Viral sequences were downloaded from GenBank and subjected to an automated parsing process. Datasets were then passed through stringent filtering criteria to ensure the production of high-quality alignments suitable for evolutionary analysis.

Temporal signal was measured using molecular clock methods to assess evolutionary rate estimates. 59% of alignments analysed showed sufficient temporal signal, allowing them to progress to substitution rate analysis. Substitution rate was measured using Bayesian frameworks with the best clock model (relaxed vs. strict clock) according to branch variation. The majority of alignments fell into moderate and fast evolving categories, while very slow and very fast were the lowest. According to temporal signal and substitution rate analysis findings, tempo did not align with viral taxonomy levels.

To study evolution mode, selection pressure was analysed over viral alignments using likelihood-based codon models, with 60% of alignments showing positive selection. Functional domains and GO annotations linked the majority of positive selected sites associated proteins to structural and replication processes. Contrary to common belief that selection in viruses is mainly driven by immune evasion, high number of positive selected sites were identified in non-surface proteins. Structural modelling further demonstrated how positive selection can impact protein surfaces and molecular interactions.

Results can contribute to an understanding how viruses evolve in general. By comparing tempo and mode across taxonomic levels, the project aims to evaluate evolutionary patter consistency within different taxonomies. Across all analyses, findings indicated that evolutionary dynamics do not uniformly follow taxonomic classification.

TABLE OF CONTENTS

List of Figures:

List of Tables:

List of Supplementary Tables:

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr Derek Gatherer for his continuous guidance, encouragement and invaluable support throughout my PhD journey. His expertise and advice have been crucial at every stage of my research.

I am also deeply grateful to Saudi Public Health Authority, for awarding me the scholarship that made this study possible. Their financial support provided me with the opportunity to pursue my research and academic goals.

My special thanks go to my parents and siblings for their love and prayers, and to my wonderful children for their patience and understanding during the long hours of work. I also want to thank my friends who stood by me with kindness and encouragement throughout this journey.

Chapter 1: Introduction

1.1 Viral genome data tsunami

Bacteriophage ΦX174 genome was the first sequence introduced in 1977 by Frederick Sanger (Sanger et al., 1977), nearly twenty years before the genome of a cellular organism, the *Haemophilus influenzae* bacteria (Fleischmann et al., 1995).

Additional to Sanger sequencing methods, high-throughput sequencing (HTS) techniques start to be standard methods for viral sequences production (Goodwin et al., 2016), although HTS methods have shown frequent sequencing errors in Illumina and MinION recorded from 0.1% to 12.7% (Bowden et al., 2019; Perez-Losada et al., 2020). The initiation of Next Generation Sequencing NGS transformed the aspect of genomics, turning a flowing stream to a tsunami of data. By enabling massive sequencing output, NGS reformed the field providing unparalleled sequencing depth and throughput (Fusté, 2012).

1.2 Genome sequences and evolution

"... [A] knowledge of sequences could contribute much to our understanding of living matter."

[Frederick Sanger] (Sanger, 2005)

Nucleic acids in polynucleotide chains holds the key to the genetic and biochemical traits of all terrestrial life forms. This role continues regardless of whether the genome is of DNA or RNA. While DNA serves as the primary genetic material in most organisms, many viruses, including influenzae viruses, coronaviruses, and retroviruses have RNA as their genetic material. RNA viruses are known to have genetic diversity due to their high mutation rate and error prone replication mechanisms (Domingo, 2001). Additionally, RNA genomes can function directly as mRNA or need reverse transcription into DNA, expanding their functional

diversity (Kolakofsky, 2015). Consequently, the capability to determine these sequences is important for advancing biological research (Heather and Chain, 2016).

The number of viral sequences has significantly increased, especially following the COVID-19 pandemic. According to the latest data from the NCBI Taxonomy database (accessed in May 2025), there are over 14,222,333 viral nucleotide sequence records in GenBank (Federhen, 2012).

Here a significant question in evolutionary biology can be asked: how genomic data utilized to find important information regarding evolution? It is important to measure evolution at the level of whole genomes since it is the main level for natural selection. Essential information bonds genotype and phenotype is carried by genome (Huynen and Bork, 1998). While natural selection works on phenotypes, in viruses these phenotypes arise from all genetic elements contribution (DeLong et al., 2022). Therefore, studying evolution at whole genome level is essential since the full genomic context shapes the viral phenotype (Loverdo and Lloyd-Smith, 2013).

Genome sequences are not only necessary to study viral evolution, but they are also essential to manage pandemics. They help in understanding how viruses mutate, spread, and interact to adapt to hosts. With all viruses surrounding from centuries, our understanding of pandemics is shaped by both historical records and molecular studies. Historical accounts of infectious diseases outbreaks date back to ancient Greece, where epidemics were described such as the Plague of Athens recorded 430 BCE. The ability to study ancient pathogens is expanded with archaeo-molecular data, such as smallpox, with ancient viral sequences recovered from human remains. Pandemics direct molecular analysis turning point started in 2003 in SARS outbreak (Flemming, 2023). Furthermore, the genetic sequencing allowed for a better understanding of viral evolution during H1N1 pandemic in 2009 (Duchene et al., 2020b).

Understanding viral evolution is our useful tool to understand viral nature and interactions. This is essential to find how viruses adapt, spread, and interact with their hosts. In the early stages of an epidemic, a virus newly introduced into a human population, undergoes rapid adaptation to improve its transmission and replication. This adaptation often leads to the accumulation of mutation, particularly in genes affecting infectivity and replication efficiency. For example, SARS-CoV-2 was estimated to acquire around 22 mutations per genome per year during its initial spread in humans. Many of these mutations resulted in amino acid changes in viral proteins, indicating the action of strong positive selection. While such changes enhance

transmissibility, they do not necessarily correlate with increased disease severity (Luo et al., 2021).

1.3 Studying viral evolution

When George Gaylord Simpson published his book “Tempo and Mode in Evolution” in 1945, during the pre-molecular genetics era when gene sequences were not yet available, he contribute to the understanding of evolutionary processes by distinguishing between the “Tempo” of evolution referring to the rate at which evolution occurs, and the “Mode” of evolution, which describes the type or pattern of evolutionary change (Simpson, 1945).

Later on, in 1995, Walter M. Fitch and Francisco J. Ayala has published a book with the title of “Tempo and Mode in Evolution Genetics and Palaeontology 50 years after Simpson” and they have some updates on the understanding of Tempo and Mode in evolution (Fitch and Ayala, 1995).

1.4 Tempo

1.4.1 Classical definition

Simpson highlighted the rates of evolution using data from zoology and palaeontology. With the ability to measure and interpret these rates with their acceleration and deceleration, fast or slow evolution can be recognized. Simpson described this using the term “tempo”.

1.4.2 Tempo and substitution rate

In the modern, post-Fitch and Ayala view, tempo is now regarded as the rate of change at the molecular level, specifically defined by the substitution rate.

1.4.2.1 Substitution rate

Among viruses, the errors rate during viral genome replication is known as the mutation rate. On the other hand, when the rate appears within all environment individuals and become fixed, is referred to as the substitution rate. While mutation rates are used to measure genetic diversity started in an offspring, substitution rates are used to measure the rate of a certain lineage or taxon evolving (Peck and Lauring, 2018).

One of the main concepts where neutral theory was built on, is the constant rates of protein change (Kimura, 1968). If many mutations do not affect fitness, they will not be affected by selection and will be left to chance. Because each neutral mutation has an equal probability of becoming fixed in a population, their rate of substitution depends on how often they occur. Therefore, Kimura (1968) suggested that the substitution rate have to be only based on the neutral mutation rate. The neutral theory's fundamental conclusion, which is that the mutation rate determines the neutral substitution rate, clearly predicts that the rate of genome evolution will differ in accordance with variations in the mutation rate (Bromham, 2020).

1.4.2.2 Molecular clock

Using genetic data, molecular clock means can estimate evolutionary rates and time frames. Molecular clock assumes that genetic change occurs at a constant pace throughout lineages, so estimations of these rates can be used to determine when evolutionary divergence events occurred across the Tree of Life. For this, the molecular clock is a crucial tool in phylogenetic study. In its most basic and original form, the molecular clock implies uniform rates across taxa and within lineages. A linear correlation for genetic distance and time is expressing molecular clock. The limitations of this basic model start to appear as growing evidence of differing rates within species, particularly among different taxa, prompted significant advancements in methodology. This has been facilitated by the Bayesian phylogenetic framework, the dramatic enhancements in computational capacity, and the increasing availability of genetic data (Ho and Duchene, 2014).

A number of molecular clock models are available, choosing the right one depends on many statistical factors, such as dataset size and computational needs. All models need calibration with time often from fossils records to estimate evolutionary timescales (Sauquet, 2013).

The strict molecular clock has the simplest model among other molecular clocks assuming a constant evolutionary rate across all phylogeny branches. With a single parameter, the rate

of evolution is measured in substitutions per site per year. Its commonly applied to intraspecific data where variation rate is expected to be low due to minimal differences in factors like generation time or DNA repair. The strict clock also serves as a null model to detect rate of variation and has been integrated to Bayesian phylogenetics for tree topology and calibrations (Brown and Yang, 2011).

Relaxed molecular clocks allow substitution rates to vary across branches, unlike strict clocks. Early models assumed rate changes were gradual and autocorrelated, influenced by life-history and environmental factors. In Bayesian frameworks, relaxed clocks are commonly used and fall into main types: autocorrelated clocks, where neighbouring branches have similar rates, and uncorrelated clocks, where each branch can have independent rate (Huelsenbeck et al., 2001).

1.5 Mode

1.5.1 Classical definition

According to Simpson, considering tempo as the rate of evolutionary change, then “mode” can be defined as the pattern of evolution either morphologically or genetically.

1.5.2 Mode and selection pressure

Mode is defined in this study as the kind of selection pressure following Fitch and Ayala. Three kinds of selection pressure mean evolutionary change undergo three modes.

- Positive selection

  The task of estimating the proportionate role of natural selection in determining the genetic variation observed across living creatures has captivated geneticists of population for times. According to one school of view, the majority of diversity within and between species is neutral selection. New mutations may become more common in the population as a result of events randomly, even if they do not offer a capability benefit to the life cell having them (Kimura, 1989; Nielsen, 2005).

Species divergence is mostly caused by positive Darwinian selection, which is also a significant source of evolutionary innovation. A broad understanding that neutral drift and positive selection both have main roles in evolutionary change has gradually replaced the Neutralist-Selectionist argument of the previous thirty years (Kosiol et al., 2008).

A long-standing debate in evolutionary genetics concerns the role that positive selection plays in molecular evolution. Modern evolutionary standards are on the neutral theory, which maintains that positive selection has only a modest influence on molecular changes and that genetic drift is the primary cause of most changes. (Smith, 1983). Nevertheless, it is becoming more and more evident that natural selection—positive and negative—is ubiquitous in many genomes, to such a degree, negative selection has been a null model to explain heterogeneity in genetic diversity levels throughout the genome. In fact, now researchers are more interested in determining "how frequent and strong is positive selection?" rather than "does positive selection present?" Consequently, a variety of methods using genomic data employed to measure the frequency and intensity of positive selection (Booker et al., 2017).

In order for evolution process to perform, natural selection is the main driving force to any life organism. Natural selection preserve the role and allows invention and adaptation (Spielman et al., 2019).

Natural selection testing are a valuable tool since they may be used objectively across the viral genome. Negative selection confirmation can disclose functionally constrained parts of the genome, whereas specific positive selection confirmation can classify genome parts where molecular activities might have deviated (Berrio et al., 2020); for instance this type of selection pressure is found in circumstances when a single mutation can boost viral gene’s ability to avoid the host immune system, similar to where a mutation can result in a selective benefit and quick fixation. Genomic scans in viruses under positive selection suggest that genes which evolve in adaptive manner are generally divided into three classes: immunological defence, chemosensory perceptivity and replication, with the bulk of these genes implicated by immune system (Kerns et al., 2008; Nielsen et al., 2005).

- Neutrality

  A neutral mutation is one which has no selective advantage or disadvantage. At the population levels, neutral mutations exhibit genetic drift, i.e., their frequencies vary at random. Point mutation will be mostly considered, there are also other kinds of mutation e.g., deletion, duplication, inversion, recombination etc.

- Constraint

  Constraint is the opposite of positive selection, as reserves a state of adaptation. Evolutionary constraint can be defined as the phenomena of limiting and restricting the adaptive evolution productions (Hansen, 2015).

  Although evolution can cause alterations in morphology, physiology and behaviour, it is often limited by various constraints. These include a lack of genetic variation, the loss of well-adapted genotypes, and trade-off arising from interactions between traits. Additionally, evolving multiple traits can be challenging. At the molecular level, genetic constraints can be evident through the loss of functional genes caused by mutational decay. Such limitations shows that although evolution is powerful, it does not act without bounds (Hoffmann, 2013).

- Measure of selection

  The synonymous/ non-synonymous substitution rate ratio dN/dS, also called Omega can be used to quantify selection pressure in coding sequences of protein. Omega measures the ratio of non-synonymous to synonymous change, distinguishing constraint, neutrality, and positive selection. Despite its drawbacks, the dN/dS ratio is simple and extensively utilised (Wilson and Consortium, 2020). This estimation's biological significance has usually been interpreted as follows:

  $dN/dS = 1$ represents a neutral evolution process.

  $dN/dS < 1$ represents a purifying (negative) selection process i.e., constraint.

  $dN/dS > 1$ indicates a process of diversifying (positive) selection.

dN/dS is used to measure selection in protein coding sequences from many species due to its well-recognised understanding and the availability of multiple applications to estimate this metric fast (Del Amparo et al., 2020).

1.6 Viral genome organization and evolutionary changes

1.6.1 Structure and function of viral genome

Viruses are intracellular parasites with compact genomes that rely entirely on the host's cellular machinery for replication and gene expression. Viral genomes may consist of either DNA or RNA, in single or double stranded forms, and can be arranged linearly, circularly or segmented into multiple nucleic acid molecules. Double stranded DNA (dsDNA) viruses often carry large genomes that replicates using host or viral DNA polymerases. Single stranded DNA (ssDNA) viruses must first convert their genomes into dsDNA before replication. Double stranded RNA (dsRNA) viruses often have segmented genomes and replicate using RNA-dependent-RNA polymerase enzymes that are either encoded within their genomes or delivered with the virion in infection. Single stranded RNA (ssRNA) viruses can be positive-sense functioning directly as mRNA, or negative-sense requiring transcription into complementary strands before translation (Chaitanya, 2019; te Velthuis, 2014).

Virus particles consist of viral genome enclosed within a protein shell known as the capsid. In some viruses, this capsid is further enveloped by a lipid membrane derived from the host cell, which is embedded with viral proteins, typically those involved in host cell recognition and binding. This outer layer, known as viral envelope, along with the capsid, responsible of number of functions in the viral infection process. These include attachment to host cells, entry to the cell, uncoating of the viral genome, and packaging of new virions. Together, these structural components facilitate the transfer of genetic material and help in virus stability. Viral spike proteins are critical for mediating entry by binding to cellular receptors and initiating membrane fusion (Lucas, 2010; Murae et al., 2022).

In addition to structural components, viruses also encode various non-structural proteins (e.g., NSP1 to NSP10, NSP12 to NSP16) that are translated from the 5' region of the viral RNA genome and are involved in replication, transcription and evasion of host immune responses (Yadav et al., 2021). Accessory proteins, although not essential for viral

replication, contribute to viral pathogenicity and host adaptation, often modulating immune responses or enhancing replication efficiency under specific conditions (Laguette and Benkirane, 2015).

1.6.2 Genetic variation mechanism in viruses

Viral genomes evolve through several mechanisms that generate diversity at the cellular level. Single-nucleotide polymorphisms (SNPs) are frequent variations found within the human genome. On average they appear once every 500 to 1000 base pairs. This occurrence is less common in coding regions compared to non-coding regions, indicating that purifying selection has a main role in SNPs distribution. SNPs involve substitutions of individual bases and can result in silent, missense or nonsense mutation (Collins et al., 2004; Gorlov et al., 2006).

Insertion and deletions (indels) alter the genome by adding or removing nucleotides, potentially causing frameshifts or disrupting functional regions (e.g., as in HIV-1 and SARS-CoV-2 viruses). Indels play a key role in protein evolution, in RNA viruses, they help drive the development of new viral traits, such as change in host interaction and tropism (Bakhache et al., 2025; Fischer et al., 2021). See figure 1 for indels, SNP diagram.

Genetic recombination occurs when two viral genomes infect the same cell and exchange sequence segments, this process is one of the main viral evolution forces and can be seen more in positive strand RNA viruses. In segmented viruses, such as Influenzae A, reassortment involves exchange of entire genome segments between co-infecting strains, leading to new antigenic variants (Perez-Losada et al., 2015; Wang et al., 2022a). Additionally, through recombination, viruses can acquire new genetic recombination, resulting of the emergence of a new virus. For instance, the western equine encephalitis virus originated from a recombination event between Sindbis-like and Eastern equine encephalitic-like alphaviruses (Simon-Loriere and Holmes, 2011). See figure 2 for viral recombination and reassortment difference.

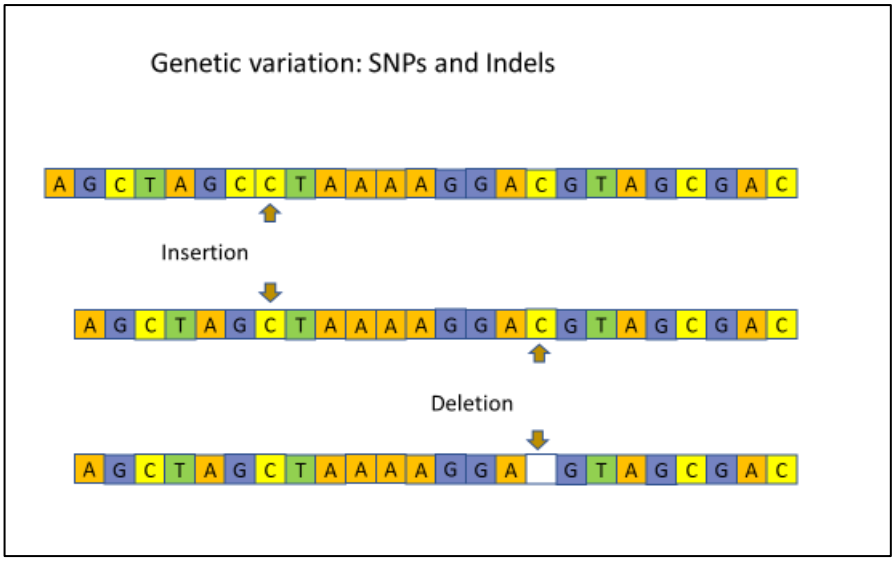

Figure 1: Illustration of genetic variation through single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels). An insertion occurs when nucleotide is added to the sequence (top), while a deletion occurs when nucleotides is removed from the sequence (bottom).

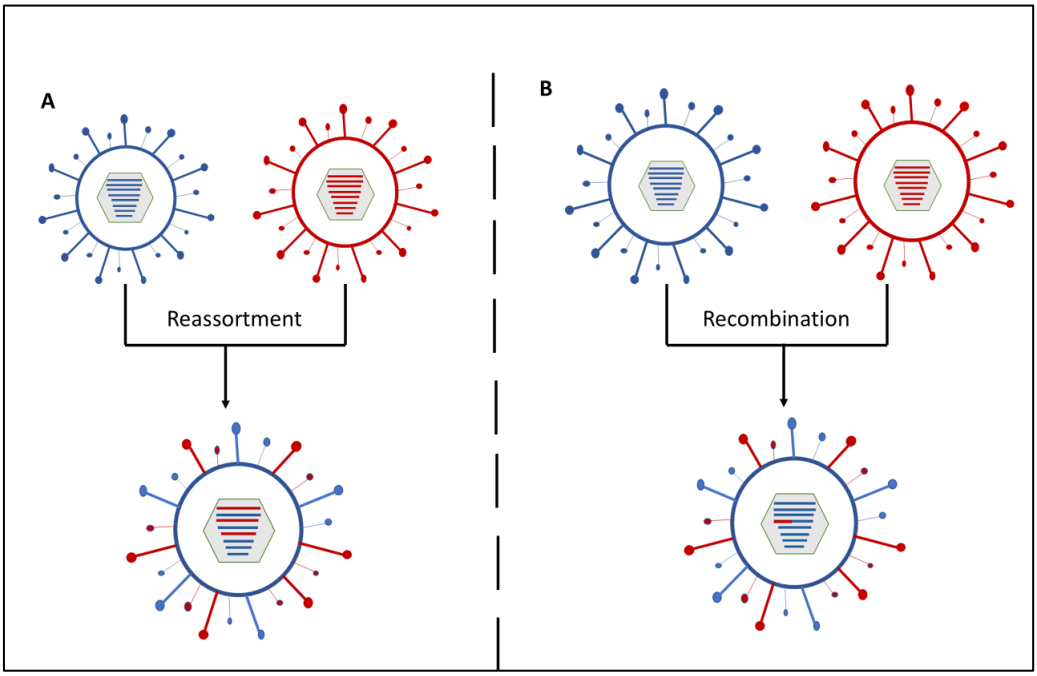

Figure 2: A diagram represents reassortment and recombination in viruses. (A) Reassortment occurs in viruses with segmented genomes (e.g., Influenza A virus, Rotavirus). When two related viruses co-infect the same host cell, entire genome segments are exchanged, producing a novel combination of parental segments. (B) Recombination occurs mainly in viruses with non-segmented genomes (e.g., SARS-CoV-2, Poliovirus). During replication, segments of genetic material from two parental genomes are joined within the same molecule, resulting in a sequence that combines genetic features from both parents.

1.7 Justification of project

The project aims to investigate evolutionary mechanisms causing viral diversity by studying tempo and mode of evolution across viral taxa. By looking at taxonomic differences, the project aims to establish a comprehensive understanding of how evolutionary forces as mutation, selection, molecular clock, and recombination interact to drive viral adaptation. The goal is to determine if tempo and mode show uniformity across taxonomic levels such as genus or family, or if these parameters vary between taxonomical levels. An additional objective is to explore whether specific evolutionary patterns, such as the presence or absence of recombination can serve as taxonomic markers, as seen in orthomyxoviruses where recombination is mostly absent.

Viruses leading to outbreaks may be subjected to new selection pressures and an expansion in virus population size, both of which may be reflected in changes in tempo, and mode and therefore are candidates for current or future pandemics.

Viruses known to be highly adaptable for their rapid mutation, high replication rates and ability to recombine, allowing them to overcome host immune system. The project begins by collecting viral genome sequences from public databases, followed by filtering, and processing them through number of softwares and models to have a general understanding of viral evolution computationally.

The project objectives aim to build a taxonomy of quantitative differences in viruses' evolutionary variations by studying tempo and mode among viruses. This project explores areas which have limited work in previous research. While many studies in virology are often presented through review-based approaches, a full taxonomic study across viral levels has not been produced. Once taxonomic approach is clear, this project aims to assess whether tempo and mode of evolution show consistency across different taxonomic levels such as genus or family, or these parameters vary independently on taxonomic classifications.

1.8 Technical approaches to study Tempo and Mode

1.8.1 Programming language

The fast progress of sequences production from different species increases the need of bioinformatic tools able to study and deal with such data. Automating large data processing is essential, with bioinformatics tools fetching and parsing genomic sequences from databases in different file formats become accessible.

Python and Biopython are open-source computational tools where could be found free online, accessible by all the main operating environments.

- Python

  Python syntax has easy structures, with an extensive range of libraries programming abilities can be used by large number of users. Python is an option for a systematic programming language since it has the ability to interface to adjusted code written in other languages e.g. C, C++ or even FORTRAN (Oliphant, 2007). Even the quantitatively challenging science of molecular dynamics has adopted Python (Hinsen, 2000).

  An additional feature in several programming languages that promotes better software design through easier maintenance and reusability is being an object-oriented programming (OOP). This paradigm structures software into reusable and adjustable components using classes and objects. Objects are self-contained units that hold data and interact with each other through defined methods to perform desired functions of an application. Classes serve as blueprints that define the properties and methods of objects. Once defined, classes can be used similarly to data types, which enhances flexibility in program design (Hirshfield and Ege, 1996).

  Python is natively object-oriented programming language that allows polymorphism, inheritance, and encapsulation. Considering everything even primitive types as an object. It can be used in both beginning practices and real-world applications for its syntax and reliable object model (Goldwasser and Letscher, 2007).

- Biopython

  Biopython has evolved into a collection of modules designed for programmers working in computational biology or bioinformatics to utilise in scripts or integrate into their individual software. The Biopython scheme offers a collection of Python software modules for sequence recovery, study, and management (Cock et al., 2009).

  Additionally, there are more programming languages with different packages that can be used to the aim of parsing records from genome sequences, below are examples of most used languages:

- Perl and BioPerl

  Perl was not natively object oriented at its initial release until a new version introduced in 1994 where it began supporting object-oriented programming features. However, Perl is still not natively object-oriented, objects are optional extra features (Prechelt, 2003).

  The BioPerl is a collection of programming tools written in Perl for handling sequence data, annotations, and alignments from various file formats and databases. It contains modules for processing outputs from sequence analysis programs and querying databases both locally and online. Perl programming language link software applications performing pipelines, also Perl convert file formats, and extract information from analysis outputs.

  The BioPerl toolkit aims to provide reusable modules with general routines for life-science data, reducing redundancy and promoting shared solutions. Once a routine is developed for parsing sequences from formats like EMBL and GenBank, it should not need to be rewritten (Stajich et al., 2002).

- BioC++

  Additional to BioPerl, more programming languages are C++ and Java. With the performance of C++, BioC++ is a toolkit that provides functions for sequence analysis, data processing, and algorithm construction. Java, known for its platform independence and large libraries, is used in bioinformatics through frameworks like BioJava. BioJava is a common choice for developing graphical user interface (GUI) based bioinformatics tools for its cross-platform compatibility. Software such as BEAST and TempEst are

built in Java to support interfaces that facilitate evolutionary analysis. Both BioC++ and BioJava provide an alternative to BioPerl and Biopython that enables manipulating large datasets and complex computational tasks (Comeau et al., 2014). Both Java and C++ are object-oriented languages.

- Bioconductor

  Bioconductor is an open-source repository of bioinformatics tools used for genomics, transcriptomics, and next generation sequencing systems. Bioconductor tools are written in R statistical programming language, installed, and adapted through an open-source model maintained by GitHub. R itself is an open-source extension of the older S language, with more statistical and graphing capabilities that can be expanded with packages from repositories like CRAN and Bioconductor (Sepulveda, 2020). Although R is not natively object-oriented language as Python and Java, it supports object-oriented programming OOP through many systems as S3, S4 and R6.

1.8.2 Database

Many biological investigations rely on saving and organising taxonomic data. Taxonomic data is frequently important metadata that aids in the organising of biology datasets and provides a modest method to direct and find the growing amount of data produced by genomic researches (Buchmann and Holmes, 2020).

The NCBI Taxonomy is an organised list of organismal names that spans all life fields, this arrangement is done on a hierarchy base. According to the finest authority in each taxonomic discipline and terminology system, these names are accurate, recent, and legitimate. To the extent possible, the taxonomy employed is cladistic, showing present understanding of relations within organisms, and it is updated on a regular basis to keep all additional data (Schoch et al., 2020).

- GenBank

  Our main source of data was GenBank, which is a complete open access database of nucleotide sequence that has features of biological and bibliographic explanation supporting. (Benson et al., 2012).

Virus classification involves organizing viruses names within a taxonomic system. The two main systems followed are the Baltimore and ICTV classifications, which categorize viruses based on phenotypic traits as morphology, genome type, replication, host type and related infections. Since phenotype-based classification requires intensive and manual efforts, some viruses remain unclassified in the ICTV system (Wang, 2015).

The Baltimore classification separate viruses into seven classes based on genome type and mRNA synthesis. While still broadly used, it does not explain evolutionary relationships. In contrast, the ICTV classification cover taxonomy beyond families and orders using structural and functional data (International Committee on Taxonomy of Viruses Executive, 2020). As a result, viruses new categorizing has replaced Baltimore at NCBI taxonomy according to evolutionary relations given by International Committee on Taxonomy of Viruses (Sayers et al., 2021).

- Viral section

  Viral GenBank records can be acquired by NCBI either a flat file or more structured formats through FTP server under the Viral section at ftp.ncbi.nlm.nih.gov/genbank. The FTP viruses only repository includes several types of compressed files, such as FASTA format, protein sequences, GenBank format flat file and coding sequences.

  Due to the demand increase on viral sequences analysis, viral genome sequences accumulate in global public databases. This highlighted the importance of creating public research infrastructure for sequences storage and analysis. This infrastructure contains main databases as those under the International Nucleotide Sequence Database Collaboration (INSDC) (Karsch-Mizrachi et al., 2012) including GenBank, European Molecular Biology Laboratory EMBL-EBL (Brooksbank et al., 2014) and DNA Database of Japan DDBJ (Kosuge et al., 2014), adding to the reference databases and the NCBI Viral Genome Resource. A reference genome is a curated sequence represented to be used as a standard for a given viral species. It is used as a reference point for genome annotation, comparison in viral evolutionary studies and pathogenicity. In the NCBI Viral Genome Project each "RefSeq" is derived from a complete genome record submitted to the INSDC, with additional annotations.

Reference genomes are uniquely identified by accession numbers starting with prefix 'NC_', distinguishing them from INSDC submissions (Brister et al., 2015).

In addition to NCBI, there are other major databases offering publicly available viral sequences. These databases are maintained by organizations in different regions, including Europe and Japan. European Nucleotide Archive (ENA) hosted by EMBL-EBI (European Bioinformatics Institute). ENA content provides the scientific community with a wide range of services and covers a wide range of data types, from raw reads to asserted annotation. Submission services are designed to meet the needs of a wide variety of data sources, ranging from large sequencing centres to small-scale research laboratories. Several thousand active data submitters data consumers receive prompt assistance from ENA helpdesk (Toribio et al., 2017). Moreover, DNA Data Bank of Japan (DDBJ) that is hosted by National Institute of Genetics, Japan, is a nucleotide sequences public database. As part of its standard database function, the DDBJ has been gathering annotated nucleotide sequences since 1987 through researcher submissions of sequence data (Kodama et al., 2018).

1.8.3 Software tools analysing Tempo and Mode

- Rates and Dates

  TempEst software is a tool used to study and examine datasets containing sequences that are temporally sampled. For any sequence, the essential piece of information needed are molecular phylogeny and sample dates. The phylogenetic tree must be built using a method that allows variable branch lengths, such as Neighbour Joining (NJ) or Maximum Likelihood (ML). Other methods which assume equal branch lengths will not be suitable for this tool. TempEst applications determine whether the data has enough temporal signal to justify calculation of substitution rate, for the progress of “molecular clock” study. And recognise sequences with dissimilar genetic discrepancy and sample dates. Data issues like as annotation mistakes, contamination in sampling, sequence recombination, and alignment defects can all be identified by observing at the latter (Rambaut et al., 2016).

  TempEst determines whether a dataset shows clock-like behaviour by calculating the correlation between genetic distance that is measured as root-to-tip substitutions on a

phylogenetic tree, and temporal distance. A strong positive correlation indicates that substitutions accumulate over time in a clock-like way.

Also, a method developed that estimates time and rate in the absence of a clock is r8s (Kishino et al., 2001). r8s is a software program used to estimate molecular evolution rates and divergence time in phylogenetic trees. This software tools evaluates variability rates across phylogenetic tree branches, using both standard maximum likelihood methods and other approaches that relax the molecular clock assumption. r8s supports the incorporation of calibration points to convert relative times into absolute scales (Sanderson, 2003).

Although r8s tool used for molecular dating, it is a likelihood-based tool with fewer features and flexibility than TempEst.

Bayesian inference of phylogeny using Markov chain Monte Carlo (MCMC) methods has significantly simplified the analysis of complex and parameter rich models. These make Bayesian analysis optimum choice for studying substitution rates, as it can handle the complexity and variability in evolutionary data (Nylander et al., 2004).

BEAST is a large software set that aims to deliver a comprehensive framework for parameter estimate and theory using evolutionary models with their sequences as data sets. BEAST, which allows for the use of previous information in conjunction with data materials (Drummond and Rambaut, 2007).

The history of evolution on planet, the dynamics of past and present populations, and infections spreading can all be inferred from molecular sequences, fossil remnants and their geographical distributions. Integrating various data sources to understand evolution over the entire spectrum of spatial and temporal scales is one of the encounters facing modern evolutionary biology. The area is facing a changing to a more quantitative field. Molecular sequence data expansion with the growth of computational and mathematical methods for their study, marked the beginning of this transition. But this change is becoming more evident in other areas of evolutionary biology, as massive worldwide databases of information sources, including geographic distributions, population histories, and fossils, are being collected and made accessible to the public (Drummond et al., 2012).

The fast growing of pathogen genome sequencing in response to infectious diseases has been a driving force behind the development of BEAST. Specifically, viruses which

are known to evolve faster may now be monitored in almost real-time to comprehend their spread and evolutionary patterns (Suchard et al., 2018).

In phylogenetic analysis, several models are used to describe different aspects of the evolutionary process. These models are essential in softwares like BEAST, which use them to generate the appropriate phylogenetic trees. Below are the key models of evolution:

- Substitution models

  Substitution models define the probabilities of replacement of one amino acid or nucleotide by other over time. It is a type of Markov process that describes the rates and probabilities at which distinct substitutions happen along a tree branch. BEAST offers number of substitution models within BEAUti interface such as, (JC) models assumes equal base frequencies and substitution rates among all nucleotide pairs. (HKY) models distinguish between transition and transversion rates, (TN93) model incorporate two separate transition rates. (GTR) models allow different rates for all substitution types and different base frequencies. All these models can be combined with additional Gamma distribution setting to allow rate variation across sites (Drummond et al., 2012).

- Rate model among branches

  The rate model among branches determines how rates are distributed along different branches, they reflect heterogeneity in the evolutionary process among lineages. These models play a crucial role in processes used to estimate divergence time. Common types include strict clock which estimates a constant rate among branches and relaxed clocks, such as uncorrelated lognormal and exponential models, which allow rate variation among branches and help measure evolutionary differences (Brown and Yang, 2011).

- Tree model

  The tree model describes phylogenetic tree structure by representing relations among sequences. Several types of tree models are used in phylogenetic analysis, such as Neighbour Joining (NJ), Maximum Likelihood (ML), Bayesian Interference and coalescent models. Common tree models available in BEAST are mainly the coalescent-based models. The coalescent constant population model assumes a

stable population size over time, while coalescent exponential growth model is designed for population that can expand or shrink. Additionally, more flexible coalescent models like the Bayesian Skyline, Skygrid and Skyride where population size changes can be estimated over time using genetic data (Drummond et al., 2012).

Additional to BEAST, MrBayes is also a Bayesian inference tool for the purpose of creating phylogenetic trees and calculating evolutionary rates. Robust statistical analysis is performed by sampling posterior distributions using Markov Chain Monte Carlo (MCMC) sampling. Being compatible with some molecular evolution models as relaxed clock models, MrBayes also handle data on both nucleotides and amino acids. Although MrBayes works on phylogenetic interference, BEAST is employed mainly for the ease of use, excels in detailed temporal and evolutionary rate analysis (Huelsenbeck and Ronquist, 2001).

Also, BEAST2 was developed to overcome some limitations in BEAST software, as limited support to third party extension and inconsistent documentation. BEAST2 has the same Bayesian evolutionary analysis capabilities as BEAST (Bouckaert et al., 2014).

- Selection pressure

The use of phylogenetic methods to analyse DNA and protein sequences has been significant due to the rapid growth of molecular sequence data, driven by multiple genome sequencing initiatives. Currently, it is typical to perform phylogeny reconstruction utilising extensive datasets that include hundreds or even thousands of genes. Phylogenetic approaches are commonly employed to estimate genomes rates of evolution, as well as to identify evidence of natural selection. The evolutionary information obtained is then utilised to interpret genomic data. Both evolutionary conservation, which indicates the presence of negative purifying selection, and fast evolution, driven by positive Darwinian selection, have been used to identify areas of the genome that are functionally significant (Yang, 2005).

Sitewise likelihood ratio (SLR) programme identifies sites in coding DNA that exhibit either exceptional conservation or variability, indicating purifying or positive selection, respectively. It does this by examining the pattern of changes in an aligned set of

sequences on an evolutionary tree. The program determines the strength of selection at each site by comparing the rate of nonsynonymous (amino acid changes) substitutions to synonymous (silent) substitutions, which are presumed to evolve neutrally, unaffected by selection. SLR conducts a precise likelihood-ratio test for selection at every alignment site, making minimal assumptions about the selection occurrence and letting for different levels of evolutionary constraint at each site. This direct test assesses whether a specific site is evolving non-neutrally. The results from multiple sitewise tests are then adjusted for several comparisons, highlighting which sites show strong sign of purifying or positive selection, thereby indicating strong evidence of selection in the alignment. Otherwise, SLR can be configured to specifically detect unusual variable sites, identifying them and proofing positive selection appearance within the alignment (Massingham and Goldman, 2005). In their article, Massingham and Goldman compared SLR and PAML (referred to the Nielsen and Yang method NY). They mentioned that traditional methods as the ones used in PAML, assumes that selection pressure vary across sites according to statistical distribution, while SLR estimates selection at each site directly without requiring such assumption. This reduces the risk of false positives associated with model misspecification and makes SLR more applicable.

PAML is a software with a number of programs designed for conducting phylogenetic studies of DNA and protein sequences using the maximum likelihood (ML) approach. The programme can be utilised for the following purposes: (i) Estimating evolutionary parameters, as lengths of phylogenetic tree branches, ratio of transition/transversion, using the maximum likelihood method. (ii) Conducting likelihood ratio tests to evaluate hypotheses related to sequence evolution, (molecular clock). (iii) Calculating substitution rates at sites. (iv) Reconstructing phylogenetic trees using both maximum likelihood and Bayesian methods. PAML incorporates a range of evolutionary models, that account for different rates of evolution among sites, allow for the analysis of numerous gene sequence data together, and specifically designed for amino acid sequences (Yang, 1997).

Another software uses maximum likelihood-based methods to study positive and negative selection in sequences is Datamonkey. Rates of synonymous and non-synonymous substitutions are estimated using codon-based models of molecular evolution to identify codons or lineages under selection, even in the presence of

recombination. Statistical tools used in this platform ranges from rapid data exploration to much complicated models, and all are accessible through web interface (Pond and Frost, 2005; Poon et al., 2009).

1.9 Substitution rate in viruses

Molecular adaptation mechanism can be best studied using viral genes. High mutation rates and rapid replication of viral genomes are essential for evading host immune system. Viral populations fast evolving rates even faster than their hosts maintain their high mutation rates. For example, mutation rate of RNA viruses are more than 100 times higher than those of fungi and *Escherichia coli* (Li, 1997). However, genome functionality can be affected with very fast mutation, resulting in mutation rates themselves being a target for natural selection. Only specific mutations enable virus to evade host immune system without affecting the virus. Also identifying amino acids under different selection pressures can help in recognising immune system targets and finding highly virulent strains that contribute to vaccine development (Anisimova and Yang, 2004).

When viruses studied, especially RNA viruses, it is commonly considered that their high mutation rates, caused by error-prone replication, result in rapid evolution. However, substitution rates are influenced by both mutation and replication rates and are also affected by adaptive environments and population size fluctuations. If most mutations are effectively neutral and rates of replication remain consistent, viral evolution may adhere to a molecular clock. Then, analysing substitution rates provides valuable understanding to roles of genetic drift and natural selection in viral evolution and helps estimate divergence dates from data of genomic sequences (Jenkins et al., 2002). Zaire ebolavirus (EBOV) is an example of a virus where the replication rates are highly variable, it shows variable replication rates during its transition between hosts. In natural reservoir species, the virus may remain dormant for long periods, with low mutation and replication rates, such behaviour suggests that the viral evolution can remain switched off in its reservoir host. However, during cross-species transmission, human host environment can be subjected to relaxed purifying selection, leading to increased viral mutation rate during outbreaks (Holmes et al., 2016; Luo et al., 2020).

Viruses have remarkable high substitution rates for their dependence on RNA-dependent RNA polymerases that lack ability to correct errors during replication. RNA viruses typically display substitution rates that span from 10^{-2} to 10^{-5} substitutions per site per year. Majority of these viruses have substitution rates of approximately 10^{-3} substitutions per site per year. The high mutation rate of these viruses enables them to adapt to new environments (Holmes, 2003). For example, the influenza virus shows how high substitution rates contributes to antigenic drift, requiring frequent updates to seasonal vaccines to meet newly emerging strains. In the same way, HIV-1 undergoes fast evolution within individual hosts because of positive selection pressures that promote mutations which aid in evading the immune system, resulting in a high intra-host substitution rate (Nielsen and Yang, 1998).

On the other hand, certain RNA viruses have lower rates of substitution as a result of different biological factors. Simian foamy virus (SFV) and human T-cell lymphotropic virus (HTLV) record lower substitution rates, which can be due to their replication and periods of latency. The primary mode of replication for SFV is the integration of its DNA into the genomes of host cells, resulting in a gradual mutation. The main reason for SFV's low substitution rates is its latent state within hosts. Although SFV reverse transcriptase does not have additional mechanisms for correcting errors during reverse transcription, the virus's latency contributes to its low substitution rates (Switzer et al., 2005). HTLV has a low transmission rate between hosts and often spreads through clonal expansion of infected cells, resulting in lower overall substitution rates. Despite these exceptions, the high mutation rates observed in most RNA viruses highlight their ability for rapid evolution and adaptation (Lemey et al., 2005).

On the other hand, DNA viruses exhibit lower substitution rates, due to high-fidelity DNA polymerases with error-correcting mechanisms. However, some exceptions, such as single-stranded DNA (ssDNA) viruses like canine parvovirus and human parvovirus B19, show high substitution rates similar to RNA viruses, highlighting the diversity in viral evolutionary dynamics. Overall, high substitution rates in viruses play a crucial role in their ability to evade host immune responses and survive in different environments, which ultimately contributes to their evolutionary success (Duffy et al., 2008).

1.10 Selection pressure in viruses

Selection pressure plays a critical role in the evolutionary process of viruses, affecting their ability to avoid host immune responses and adjust to different environments. Measuring substitution rates, enables understanding selective pressures on virus populations. Following examples of selection pressure in viruses demonstrate how positive selection allows the evasion of the immune system, while negative selection preserves vital viral functions.

As previously mentioned, when the rate of nonsynonymous substitutions (dN) is higher than synonymous substitutions (dS), indicates presence of positive selection. While antigenic variation is driven by positive selection, negative selection works on conserving viral functions by eliminating mutations with reducing non-synonymous substitution rates (Frank, 2002).

For example, the surface glycoprotein gp120 of HIV-1, which is responsible for host cell entry, shows strong positive selection at several amino acid sites. These sites, located in variable regions of the protein, allow the virus to escape neutralising antibodies and adapt to host environments. Studying 186 HIV-1 subtype B sequences showed presence of positive selection at 33 different sites. These locations are mainly found in the exposed regions of the gp120 protein, highlighting the virus's ability to adapt and avoid detection by the immune system (Yamaguchi-Kabata and Gojobori, 2000).

While negative selection tends to eliminate mutations that result in amino acid change, positive selection allows such variant to persist and even replace the original population, favouring their survival. One example of the impact of positive selection on viral evolution is the Influenza A virus, this virus represents selection pressure impact on antigenic drift, a process where regular accumulation of mutations in viral antigens, such as hemagglutinin (HA), enables the virus to escape host immunity. Influenza A isolates were collected between 1983 and 1997 in a study revealed that alterations in amino acids at specific sites that were under positive selection could accurately estimate the future success of viral lineages. Typically, these mutations, which modified epitopes recognised by antibodies, enabled the virus to avoid pre-existing immunity in the host population, leading to development of new dominant strains. Studies shown that positive selection pressures on the (HA) genes of H1, H3, and H5 subtypes are key to the evolution of antigenic and receptor binding sites. Another example comes from picornaviruses, where VP1, the most visible and immunodominant protein in the viral capsid, is a well-known target of positive selection (Bush et al., 1999; Shi et al., 2009).

The foot-and-mouth disease virus (FMDV) is a prime example of positive selection in viral evolution. Haydon et al. (2001) discovered 17 sites under strong positive selection in FMDV isolates. Out of these 17 locations, 12 were already known to be regions where escape mutants developed under monoclonal antibody pressure in experimental studies. This emphasises the significance of changes in amino acids that help the virus in avoiding the immune response of the host. It also highlights the importance of combining natural variations with experimental data to understand viral evolution.

While surface proteins, especially envelope or capsid components are targets for immune driven positive selection, other researchers highlighted adaptive evolution in non-surface proteins such as polymerases and accessory (functional) proteins (Redondo et al., 2021). For example, PB2 E627K mutation in influenzae polymerase enhances replication in mammalian hosts by optimizing activity at human airway temperatures (Long et al., 2013). Similarly, studies on SARS-CoV-2 revealed that mutations in accessory proteins like ORF8 contributed to immune evasion and host adaptation (Thorne et al., 2022). These examples illustrate how amino acid changes in internal viral proteins can be positively selected when they enhance replication efficiency or host range, even in the absence of direct immune pressure.

1.10.1 Protein structure in selection

Protein tertiary structure plays an important role in understanding protein evolution by imposing structural constraints that develop dependencies on sequence positions. Incorporating tertiary structure mostly generates additional data, revealing how phenotype impacts genotype evolution. Progression in computational biology allowed structural data integration into evolutionary models, which highlight the relationship between structure and evolutionary rates across diverse protein families (Choi et al., 2007). Selection on protein sequences is shaped by number of biochemical and biophysical factors that influence function and evolutionary stability. Selection has an influence on protein function as well as on the preservation of stability and orientations of functional residues (Chi and Liberles, 2016).

Viruses maintain their life cycle by adapting immune evasion mechanism, to ensure survival within a host. Some viruses as picornaviruses and retroviruses have the ability to evade host

immune system by altering their envelope glycoprotein to escape the recognition of major histocompatibility (MHC) class I molecules (Vossen et al., 2002). Viral infections start with viral proteins binding to host cell surface components, such as proteins, lipids, or glycans initiating entry process. These interactions facilitate the virus attachment to host cell and trigger subsequent steps in the infection cycle. Imaging techniques and structural biology advances as x-ray or electron microscopy studies receptor-virus interactions in details, which enables to determine many cellular processes (Koehler et al., 2020; Wang et al., 2024). In molecular recognition, surface residues which are more variable than conserved core residues are vital, and positive selection frequently targets them. Without disrupting overall structural stability, mutations in these regions can have a remarkable impact on protein interactions and evolution (Kini and Chan, 1999).

1.11 Aim and hypothesis of the study

The aim of this project is to investigate whether tempo (substitution rate) and mode (selection pressure) of viral evolution exhibit consistent patterns across taxonomic levels, and whether theses evolutionary parameters can serve as reliable taxonomic markers. By studying viruses from a wide range of families and genera, the project seeks to establish a taxonomy of quantitative differences in viral evolutionary dynamics.

This work also explores the role of other evolutionary forces such as recombination and molecular clock behaviour and how they interact with tempo and mode to shape viral evolution diversity. A key aspect is to determine whether evolutionary parameters reflect formal taxonomic classifications or vary independently of them.

If tempo and mode of viral evolution are closely linked to taxonomic levels, then consistent evolutionary patterns will be observed across related viruses within the same genus or family. Alternatively, if these parameters vary independently, it suggests that evolutionary behaviour is shaped more by ecological or functional constraint than by taxonomic relationships.

Chapter 2: Materials and Methods

2.1 Software and scripting:

2.1.1 Virtual machine

Methods were performed within a virtual machine which is a remote and distributed computer environment that may run a different operating system.

Virtualization produces a virtual version of operating system, data storage device or even a computer that is not an independent device but appears as a single physical entity to the operator.

Work done on VMware version number 11.3.0.29534, build number 18090558. An access to a remote desktop connection was established to the VM using X2Go Client version number 4.1.2.2 (Baur, 2023). X2Go created an Ubuntu open source version, Ubuntu 20.04.6 LTS, licensed XFCE version number 4.16 desktop environment which enables working remotely (Fourdan, 2011).

2.1.2 GenBank parsing tools

All input data records were initially processed by applying scripts written in Python 3, using modules described in Table1. Number of Python packages and modules were imported to perform several programming tasks in written scripts, see Table 1 for python packages details and Table 2 for scripts list. For example, pandas and date time modules were imported to calculate and standardise differences in collection dates formats annotated in GenBank records, specifically at three digits availability, these imported packages from python worked on date format uniformity by adding dummy days or months, thus if collection date was submitted on NCBI as 2012, then it will be modified to 30-Jun-2012 and

Aug/2010 will be modified to 15-Aug-2010. Also, an additional flag “m” will be added to the modified sequence header, and “u” to the unmodified date.

Biopython 1.79 offers Python libraries for computational biology and bioinformatics, widely used for solving bioinformatics problems and integrating into custom software (Chang, 2020; Cock et al., 2009).

The Bio.SeqIO module is the main Biopython sequence parser, it offers a straightforward interface for writing biological sequence files in multiple formats.

Table 1: Illustrates imported modules used to achieve results and their description (Guido, 1990).

Module	**Description**
>>> from Bio import SeqIO	SeqIO sequence input/output, used for the purpose of reading GenBank sequences.
>>> import pandas as pd	Pandas is a data analysis and modelling library, which was used for date format conversion e.g. DDMMYY
>>> from datetime import datetime	Datetime stores classes for changing dates and times, it is used to convert dates to a desired format.
>>> import re	Re allows checking a given string matches a regular expression. In our script regular expressions used for dates checking if any of them are missed date, month, year.
>>> import os	Os allows the use operating system-dependent functions on the go, it was used to create directory using python in a specific location.
>>> import shutil	Shutil is a high-level operative file module that supports a variety of operations and file collections. There are functions that assist file copying and removal in particular.
>>> import sys	Sys runs as an access to some variables that the interpreter uses, as well as functions that have close relationships with the interpreter.
>>> from collections import counter	Counter is a class offered by collections module, it is used to identify the most frequent items and carrying out mathematical operations on counts.

2.1.3 Directory creation

- Taxonomy hierarchy directory

  Using computational tools mentioned, coding scripts were written to parse viruses records from GenBank format files. The output was achieved by creating a directory called “Viruses” that contains folders and subfolders named according to viruses’ taxonomy hierarchy classification as present in each GenBank record. The “Viruses” directory works as the main directory with sub directories for every descending level copying taxonomy hierarchy. At the last folder order reached, a single or multiple FASTA formatted files are created and saved in text files. The file name has a title of the “organism” name similar to GenBank record feature, in each of these files all the available sequences for one organism are listed separated with headers, under each header the concatenated coding sequence “codome” for one species is saved, each header will contain the following information: accession number + country + collection date in a uniform format and an indication if the date is modified or unmodified e.g. >MH017546.1/Ghana/m/30-Jun-2014, where the first part after the arrow is the accession number in GenBank record followed by the country of origin, next **m** stands for a modified date and the last part is the modified date. Figure 3 is an example of a descending order of folders and subfolders starting with “Viruses” directory reaching to FASTA format file with more than one sequence for Fulton virus, starting from super kingdom level "Viruses", then clade level "Riboviria"**,** after that kingdom level "Orthonavirae", follow there is phylum level "Negarnaviricota", later on genus "Unclassified_Negarnaviricota" and the last level the species name is "Fulton_virus" where it will be saved in FASTA format file called Fulton_virus.fasta.

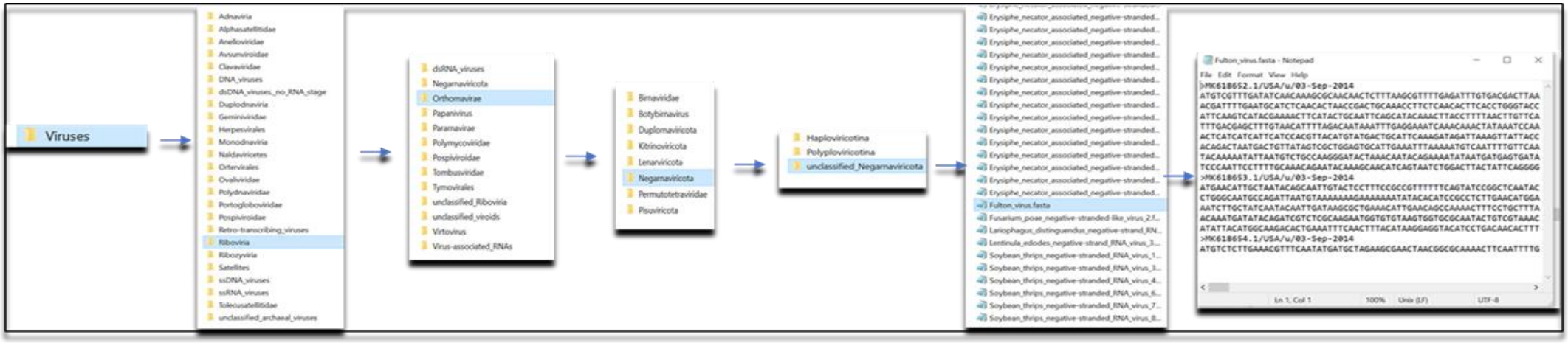

Figure 3: Example of hierarchical folder structure for parsed viral sequences. This shows a stepwise view of descending directories within the main **"Viruses"** folder, generated after sequence data were parsed and organized from GenBank records. The hierarchy begins at the highest taxonomic level, where folders are arranged by major viral groups (e.g., *Riboviria*). Each subsequent level represents a more specific taxonomic rank, progressing through phylum, order, family, genus, and subgenus (where applicable). The final folder contains sequence files in FASTA format for individual viral species, each named according to the species in each record. The rightmost panel illustrates an example of a FASTA file opened in a text editor, showing the sequence header (including accession number, collection date and country of origin) followed by the nucleotide sequence

- Genomes and reference genomes directory

  Reference genomes records were parsed with a different criterion than other viral genomes, see section 2.3. Concatenated reference viral sequences were stored in text files with only accession number in the header and then saved according to taxonomical hierarchy inside the main “Viruses” directory at the last order directory similar to the method followed with the viral genome records.

2.1.4 Scripting

The majority of data processing and analysis steps in this project were automated using Python scripts. These scripts were written to perform tasks and automate software running larger datasets. Utilizing a range of previously mentioned Python modules and packages. Below is a description of each script and its function within the analysis pipeline.

- Script 1: GenBank parsing

  Script #1 performs the initial parsing and saving of coding sequences for all viral and reference genome GenBank records. The script parses viral genome sequences and filter them according to the presence of collection date and CDS. It then builds a structured directory system following viruses taxonomy. The script processes the coding sequences following an inclusion criterion to assure no stop codons produced in the produced codomes. Once validated and FASTA files created and saved, the script generates summary spreadsheets listing parsed sequences counts, species names, taxonomy hierarchy levels and collection dates.

- Script 2: Ref_codomes directory

  Script #2 was designed to organize and filter FASTA files based on the presence of corresponding reference genomes. The script takes FASTA files created from script#1 and save only files that have matching reference codome files into a second directory, regardless of taxonomy. This ensures that downstream analysis is conducted only on viral species with available reference genomes.

- Script 3: Segments alignments

Script #3 automates the use of CD-Hit program (further discussed in section 2.3.3). The script takes all reference codome sequences as an input and runs CD-Hit to identify clusters of highly similar sequences within each reference genome.

- Script 4: Directory structure

  Script #4 creates species specific directory which includes both Ref-codomes FASTA and all corresponding codome FASTA sequences for the species. This script mainly serves segmented viruses, where each segment was treated as a separate species during analysis.

- Script 5: Filtering codomes and Ref-codomes

  Script #5 applies length-based filtering criteria to refine codomes and reference codome sequences saved in the directory created at script 4. This script checks each Ref-codome file and keeps it only if its length is <= 50 kb. Then, it saves only codome sequences if they are 75% to 100% of the length of their corresponding Ref-codome.

- Script 6: Nucleotide translation

  Script #6 is responsible for translating nucleotide coding sequences (codomes and Ref-codomes) into their corresponding amino acid sequences. The output consists of translated peptide sequences transcodomes and Ref-transcodomes. This step prior to pairwise sequence alignments.

- Script 7: Pairwise alignments

  Script #7 performs pairwise sequence alignment between codome sequences and their corresponding Ref-codomes, also between transcodomes and their respective Ref-transcodomes. The script uses alignment tool to compute the similarity percentage between each sequence pair. The output contains files with identity scores, which serves as a quality checking step for identity.

- Script 8: Filtering pairwise aligned sequences

  Script #8 processes the pairwise alignment identity scores outputs from script 7 and applies sequence similarity thresholds to determine codomes saved for the tempo and mode studying. The script filters codomes and transcodomes with low similarity scores and save the high-quality codome sequences for further multiple sequence alignment step.

- Script 9: Calculation of codomes collection dates

  Script #9 calculates the earliest and latest collection dates for each species aligned codome sequences after filters applied. This script uses spreadsheet generated from script #1 as an input file, which contains metadata of parsed codomes collection dates. Using Python modules, the script calculates earliest and latest collection dates for each dataset.

- Script 10: Multiple sequence alignment

  Script #10 performs multiple sequence alignment of codome sequences for each species using the alignment tool MAFFT. The script automates MAFFT to produce align output files for every dataset.

- Script 11: XML creation

  Script #11 generates XML files from the aligned codome sequences produced by MAFFT. The generated XMLs work as input files for BEAST run which hold all metadata from aligned codome sequences and also define Bayesian models and parameters chosen for BEAST further run.

- Script 12: BEAST

  Script #12 automates the performance of BEAST analyses using XMLs generated from script 11. It automates BEAST run by applying BEAST -overwrite command on each XML file. The produced output is a .log file for each dataset holds parameters estimates.

- Script 13: PHYLIP files creation

  Script #13 converts FASTA codome sequences to PHYLIP format, this script starts the stage of SLR files creation, (description of files types from script 13 to 17 are mentioned in section 2.8.1)

- Script 14: Distance matrix calculation

  Script #14 calculates a distance matrix for each PHYLIP alignment file producing dismat files.

- Script15: Tree files creation

  Script #15 generates neighbor-joining phylogenetic trees from dismat input files.

- Script 16: Sequence files creation

  Script #16 reconvert the PHYLIP files into SLR compatible sequence files.

- Script 17: Control files generation

  Script #17 generates a control (CTL) file for each dataset which contains path of three components: the aligned sequence file, the corresponding tree file and the output file location. These CTL files are the input for SLR analysis.

- Script 18: SLR run

  Script #18 automates the execution of SLR analysis for each alignment. It applies SLR_shared command to run SLR and produce output file saved in the same directory for the corresponding species alignment.

- Script 19: SLR outfile data exporting

  Script #19 extracts key information from SLR output files generated by script 18. It reads each output files to identify sites with positive selection and save results in a separate spreadsheet.

- Script 20: Peptide location calculation

  Script #20 maps positively selected amino acid sites back to their corresponding coding sequences (CDSs) using GenBank records. The script calculates the peptide level location of each CDS, and mature peptide regions if appeared. Then positively selected residues' locations are listed in a spreadsheet with their corresponding gene, product, protein ID.

- Script 21: InterProscan

  Script #21 performs functional annotation of all proteins under positive selection using InterProscan domain. This script once applied on an input file contains list of protein IDs, it then generates a spreadsheet contains: Gene Ontology terms, Pfam domain names, and other associated functions for each protein ID.

- Script 22: FASTA splitting

  Script #22 splits a FASTA file containing all selected proteins sequences into individual FASTA files, each representing a single protein.

- Script 23: Blastall

  Script #23 applies the blastall command to each individual protein FASTA file (generated from script 22), searching against the Protein Data Bank (PDB) to identify structurally similar proteins. The scripts output a spreadsheet listing all PDB matches and percentage identity scores for each protein ID.

Table 2 below lists scripts written and used specifying input and output data with modules imported:

Table 2: List of all scripts written to perform methods, with the type of input and output files and modules imported.

	Script	**Input**	**Output**	**Modules/ Applications**
1	GenBank parsing script	1. Viral GenBank records version # 2.44 2. RefSeq viral records	1. FASTA date collected codome and Ref-codomes sequences 2. Taxonomy hierarchy directory	SeqIO Pandas Re Os Datetime Counter
2	Ref_codomes directory script	FASTA text files	Arranged species directory.	Shutil
3	Segments alignments script	Ref_codomes	Text files identifying viral cluster segments	Os Cd-hit
4	Directory structure script	Codomes and Ref-codomes for all species	New arrangement: each species has a separate directory with all segments in sub-level	Os shutil
5	Filters applied on codomes and Ref-codomes directory	The output directory from script #4	1. Directory with codomes 75% and more of Ref-codome length. 2. All Ref-codomes are <= 50kb.	Os SeqIO Shutil
6	Nucleotide translating script	Codomes and Ref-codomes	Translated peptide sequences; Transcodomes and Ref-transcodomes	EMBOSS Transeq Os Shutil
7	Pairwise alignment script	1. Codomes and Ref-codomes. 2. Transcodomes and Ref-transcodomes	Pairwise aligned sequences with identity score percentage.	EMBOSS Needle Os Shutil
8	Filters applied on pairwise aligned sequences script	Needle pairwise alignment output in text files for both	3. Codomes to Ref-codomes identity score >= 90%.	Os Sys

		1. DNA level in codomes with trancodomes. 2. Peptide level in transcodomes with Ref-transcodomes.	4. Transcodomes to Ref-transcodomes identity score 50%.	
9	Codomes collection date calculation script	Spreadsheet of codomes collection dates out of script #1	Earliest and latest collection date for each aligned codomes dataset.	Datetime Os SeqIO Sys
10	Multiple sequence alignment script	Codomes sequences	MAFFT aligned data sets	MAFFT Os Sys
11	XML creation script	Aligned codomes data set	XMLs of aligned codomes sequences data set available	SeqIO Os Datetime
12	BEAST run	XMLs of codomes sequences data	log files with BEAST output	Beast -overwrite Os
13	PHYLIP files creation script	MAFFT aligned codome sequences data sets	PHYLIP format files	Os Seqret
14	Distance matrix calculation script	PHYLIP files	Dismat files	Os Fdnadist
15	Tree files creation script	Dismat files	Tree files	Os Fneighbor
16	Sequence files creation script	PHYLIP files	Seq files	Os Seqret
17	Control files generation script	1. Seq files paths 2. Tree files paths 3. Out files paths	SLR ctl files	Os
18	SLR run	Ctl files	Outfiles as text documents	Os SLR_shared
19	SLR outfile data exporting	Outfiles	Spreadsheet with all species data for positive selection	Os
20	Peptide location calculation script	1. Viral GenBank records for species with positive selection sites. 2. Viral GenBank records with matpeptide in CDS.	1. CDS and matpeptide location for peptides sequences. 2. Protein Id, gene, locus and notes for each coding part of the sequence.	Os SeqIO
21	InterProscan script	Text file contains all selected protein IDs	InterProscan data listed for all selected proteins	Os InterProscan
22	FASTA split script	Text file contains all selected protein in FASTA format	Directory contains each protein in FASTA format file	SeqIO
23	Blastall script	Directory contains each protein in FASTA format file	Protein matches in PDB database	Blastall Os

2.1.5 Scripts flowcharts

In order to enhance the clarity of processes done through scripting, flowcharts were employed. Initial flowcharts were generated by importing python "pyflowchart" library version 0.2.3, resulting in an abstract presentation of each flowchart, later each script output was exported to flowchart.js to draw the textual presentation of the flowchart showing; decision points, processes, arrows and boxes shapes (Adriano Raiano; Warehouse project, April 2018). The last step was drawing each flowchart using app.diagrams.net online tool. Figures 2 to 11 represents 10 scripts processes; each numbered according to original scripts number in Table2. Below is detailed description for all flowcharts displaying scripts.

- Script 1 flowchart: GenBank records parsing

  1. Start: the process begins with two data inputs from GenBank.

  Input1: viral genome records (GenBank release v2.44)

  Input2: reference genome records (downloaded Aug 2021)

  2. Processing viral GenBank records (input1):

  Check presence of collection date and CDS → if either is missing, then record is discarded.

  Check date format (DD-MM-YYYY); the collection date must follow the required format → if not, a dummy or modified date is inserted.

  Print header; a FASTA header is generated containing the accession number, date and country.

  3. Processing Reference genomes (input2)

  Check presence of CDS only → if not present discard the record:

  → if present, continue to print header with accession number only.

  4. Directory and taxonomy setup

  Create "Viruses" directory and taxonomy-based subdirectories:

  Records are sorted and saved according to their taxonomic classification with each level (e.g., family, genus, species) represented as a subfolder.

  5. Quality control on coding sequences: applies on both viral and reference genomes.

  Check for:

  - Sequence length a multiple of 3 value.
  - Correct start codon values.
  - Presence of N residue < 25% of sequence length.

→ sequence pass these checks proceed to the next step.

6. Create codomes and Ref-codomes:
   Valid coding sequences are saved as codomes or Ref-codomes in FASTA format, organized within the appropriate taxonomy-based directory.
7. Output files:
   Script generates a spreadsheet listing; all parsed sequences species names and number of records, a text file for each species documenting taxonomy hierarchy and a spreadsheet listing collection dates for all parsed codomes.
8. End

See figure 4 for Script 1 flowchart.

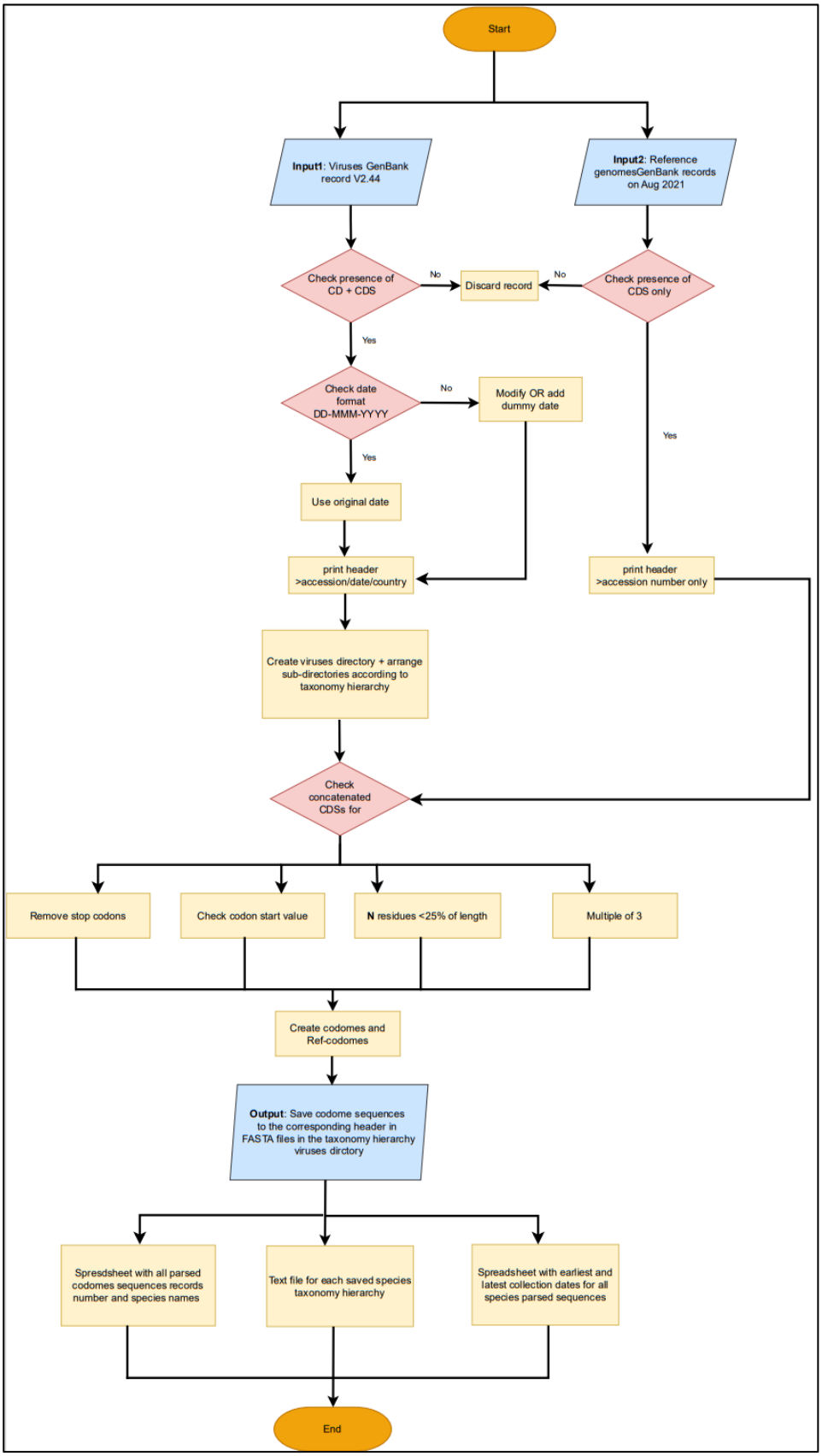

Figure 4: Script #1 flowchart, showing workflow for viral GenBank and Reference genome parsing, starting with two input files the GenBank record for viruses V2.44 and the downloaded Reference genomes on August 2021 and then ending with creating output viruses directory contain FASTA files with parsed codome and Ref-codome sequences.

- Script 4 flowchart: Directory structure

  This flowchart outlines the steps used to structure the parsed sequences by species, ensuring each directory contains both codomes and their corresponding reference sequence.

  1. Start: the process begins with an input directory that includes all FASTA files for codomes and Ref-codomes.
  2. Check Ref-codomes for NC_ identifier: each Ref-codome file is scanned for a NC_ accession prefix → if the identifier is not present, the entry is discarded.
  3. Create species specific directory:

     For entries with NC_ identifier, the script creates a new folder specific to that species. This folder contains:

     - Ref-codome FASTA
     - All codomes FASTA files associated with that species.
  4. Output:

     The result is a structured directory containing subdirectories per NC_ reference, each holding the relevant codome and reference codome sequences.
  5. End

  See figure 5 for script 4 flowchart.

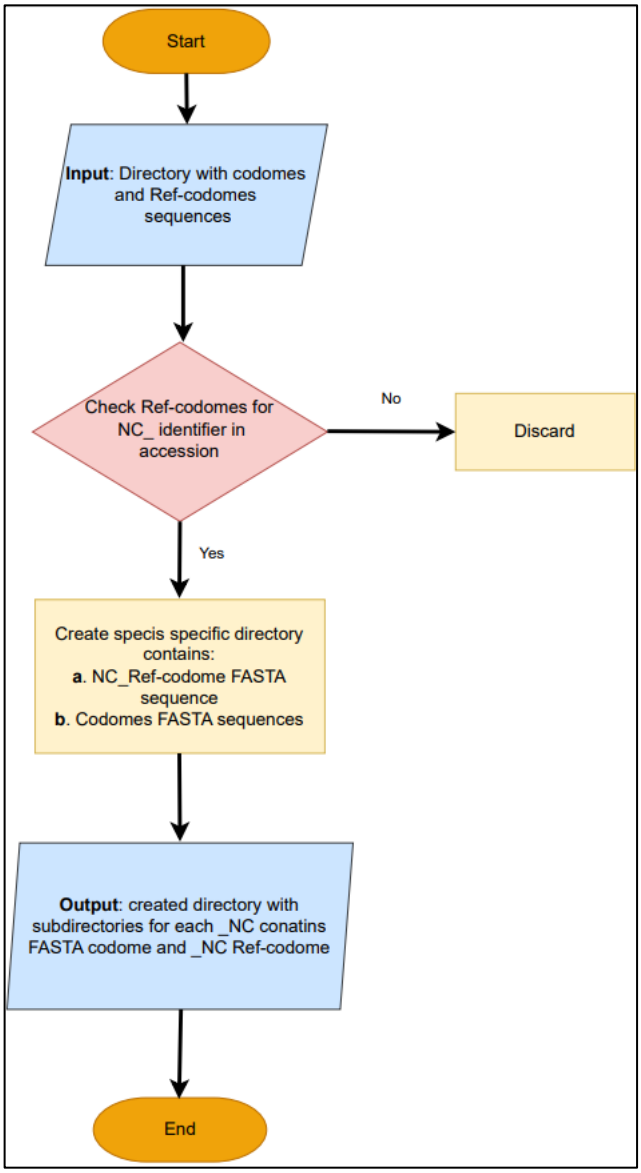

Figure 5: Script #4 flowchart. Showing workflow of directory structure creation for codomes and Ref-codomes. Starting by the directory contains all codomes and Ref-codomes and ending with an output of a directory contains subdirectories for each Ref-codome has NC_ identifier in accession with its corresponding codomes.

- Script 5 flowchart: Filtering codomes and Ref-codomes
  1. Start: the process begins with the species-specific directories generated in script 4 as an input, which contains both codomes and Ref-codomes FASTA files.
  2. Ref-codomes length check: each Ref-codome is checked for sequence length.

  → if the Ref-codome is larger than 50 kb, then it is discarded.

  → 50 kb or less, then proceed to the next step.

  3. Codome length filtering:

  Each codome is compared to the length of the retained Ref-codome.

  → Codomes are kept as saved only if their length is between 75% and 100% of the Ref-codome length.

  → Codomes outside this range are discarded.

  4. Output: The updated directory includes only:
     - Ref-codomes <= 50 kb.
     - Codomes that are 75 – 100% of the Ref-codome length.
  5. End

  See figure 6 for script 5 flowchart.

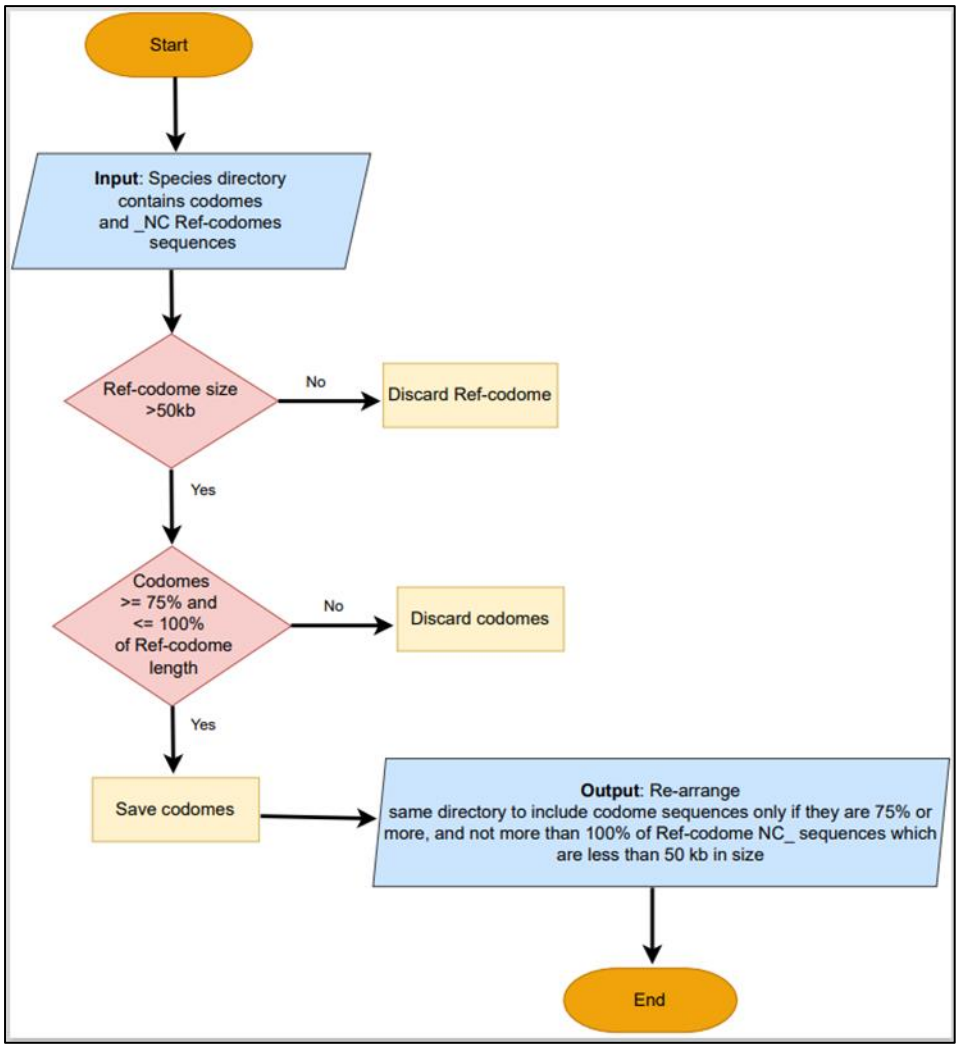

Figure 6: Script #5 flowchart. Showing workflow steps of codomes and Ref-codome length base filtering. Starting by directory created from previous script containing codomes and NC_ Ref-codomes, and ending up with arranged directory with NC_ filtered size and codomes filtered length.

- Script 7 flowchart: Pairwise alignment

1. Start: the script begins with a directory containing species specific files:
   - Codomes and Ref-codomes (.fasta and ref.fasta)
   - Transcodomes and Ref-transcodomes (.pep and ref.pep)
2. Loop through files:

The script checks each file name and groups:

- .pep files with their corresponding ref.pep file.
- .fasta files with their corresponding ref.fasta file.

3. Apply EMBOSS Needle:

For each matching pair, the EMBOSS Needle program is used to generate a global pairwise alignment.

4. Output:

The result is a set of alignments reports containing:

- Identity percentage between each codome and Ref-codome.
- Identity percentage between each transcodome and Ref-transcodome.

5. End.

See figure 7 for script 7 flowchart.

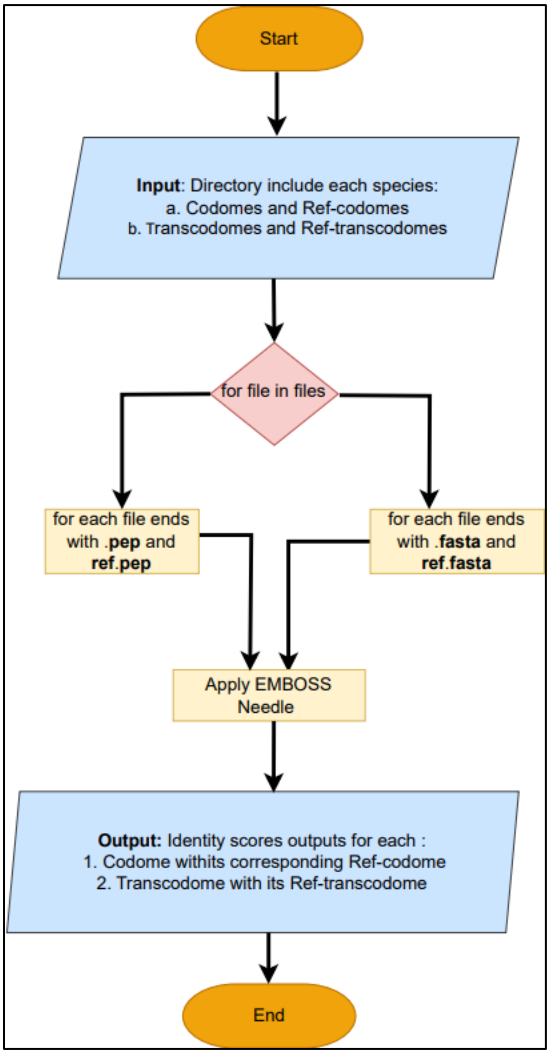

Figure 7: Script #7 flowchart. Showing workflow of pairwise alignment of codomes and with their Ref-codomes and also transcodomes with their corresponding Ref-transcodomes. Starting with a directory contains all codomes, transcodomes, Ref-codomes and Ref-transcodomes, and ending up with pairwise alignment identity scores.

- Script 8 flowchart: Filters on pairwise aligned sequences

  This flowchart outlines how codome and transcodome identity scores are evaluated to determine whether sequences are kept or discarded.

  1. Start: the input consists of text files containing identity scores from the EMBOSS Needle alignments as both DNA and peptide levels.
  2. Check DNA similarity scores (codomes vs. Ref-codome)

  → if the identity is 0 to 50 %, then codome is discarded.

  → if the identity is 90 to 100%, then codome is accepted and saved.

  → if the identity is 50 to 89%, a second filter is applied using peptide similarity score.

  3. Check peptide similarity scores (transcodomes vs. Ref-transcodomes)

  → if identity is 90 to 100%, the codome is accepted and saved.

  → if identity is < 90%, the codome is then discarded.

  4. Save results:

  Headers of accepted codomes are saved into text files.

  Codome sequences that pass all filters are collected and saved for multiple sequence alignment.

  5. Output: final output includes:
     - Codomes with >= 90% nucleotide identity.
     - Codomes with 50 to 89% nucleotide identity but >= 90% peptide identity.
  6. End.

  See figure 8 for script 8 flowchart.

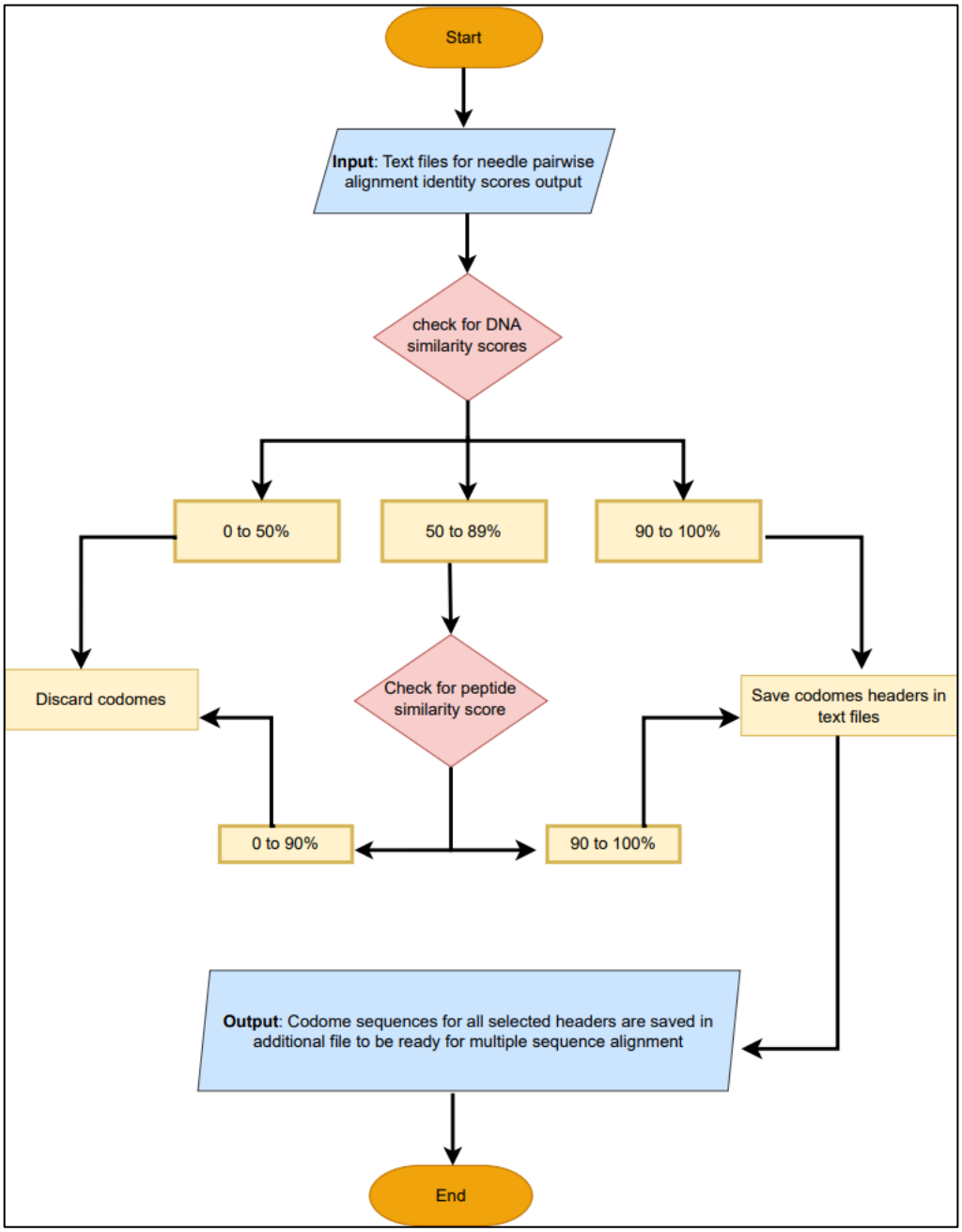

Figure 8: Script #8 flowchart. Showing workflow of filtering codome sequences based on pairwise alignment identity scores for codome (DNA or RNA) and transcodome (peptide) levels. Starting with text files containing pairwise alignment identity scores, and ending with filtered codomes according to threshold identity chosen for both codomes and transcodomes scores.

- Scripts 11 and 12 flowchart: XML creation and BEAST run

  This flowchart combines the process of scripts 11 and 12, covering the transition from sequence alignment to BEAST output generation.

  1. Start:

  The process begins with a directory of MAFFT-aligned codomes datasets.

  2. Define taxa IDs and taxon references:

  For each sequence, an ID is constructed using accession, country and collection date. These are used to calculate time values for tip dating.

  3. Generate tree and site models:

  Tree models used an uncorrelated log-normal relaxed clock.

  Site models use the GTR substitution model, including rate parameters and frequency settings.

  4. Define MCMC and operators:

  MCMC chain settings and BEAST operators are defined for each dataset to configure XML.

  5. Create XML files:

  An XML file is generated for each alignment dataset, containing all BEAST required setting.

  6. Run BEAST:

  The script loops through all .xml files and run beast -overwrite command on each alignment.

  7. Output:

  For each dataset, BEAST generates a .log file in the same directory, containing the output of the evolutionary rate estimation and model performance.

  8. End

  See figure 9 for scripts 11 and 12 flowchart.

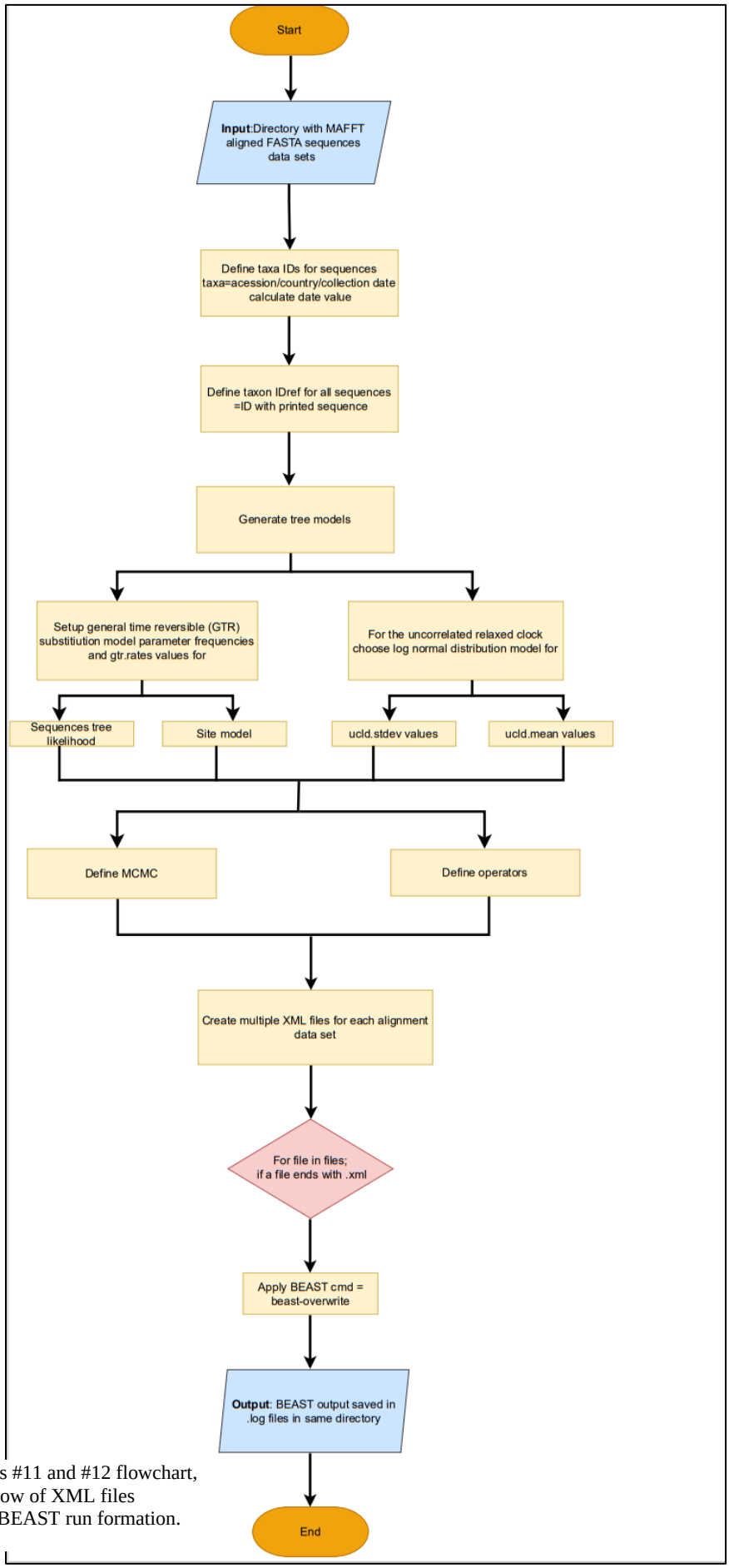

Figure 9: Scripts #11 and #12 flowchart, showing workflow of XML files generation and BEAST run formation.

- Scripts 13 to 17 flowchart: SLR preparation workflow

  1. Start:

  The process begins with MAFFT-aligned codomes in FASTA format as input.

  2. Convert FASTA to PHYLIP:

  Using seqret command, FASTA alignments are converted into PHYLIP format.

  3. Distance matrix calculation:

  The PHYLIP files are passed to fdnadist, which calculates pairwise distances and produce dismat files.

  4. Tree file generation:

  The dismat files are used by fneighbor to construct neighbor-joining trees, outputting tree files for each dataset.

  5. Sequence file formatting:

  In parallel, PHYLIP files are reconverted via seqret into SLR-compatible sequence files.

  6. Control file creation:

  Finally, for each dataset, a CTL file is generated pointing to the required paths: sequence file, tree file, and output destination.

  7. Output:

  For each alignment, a complete set of files (tree, sequence, and CTL) is saved in the respective species subdirectory, ready to be used by SLR analysis.

  8. End.

  See figure 10 for scripts 13 to 17 flowchart.

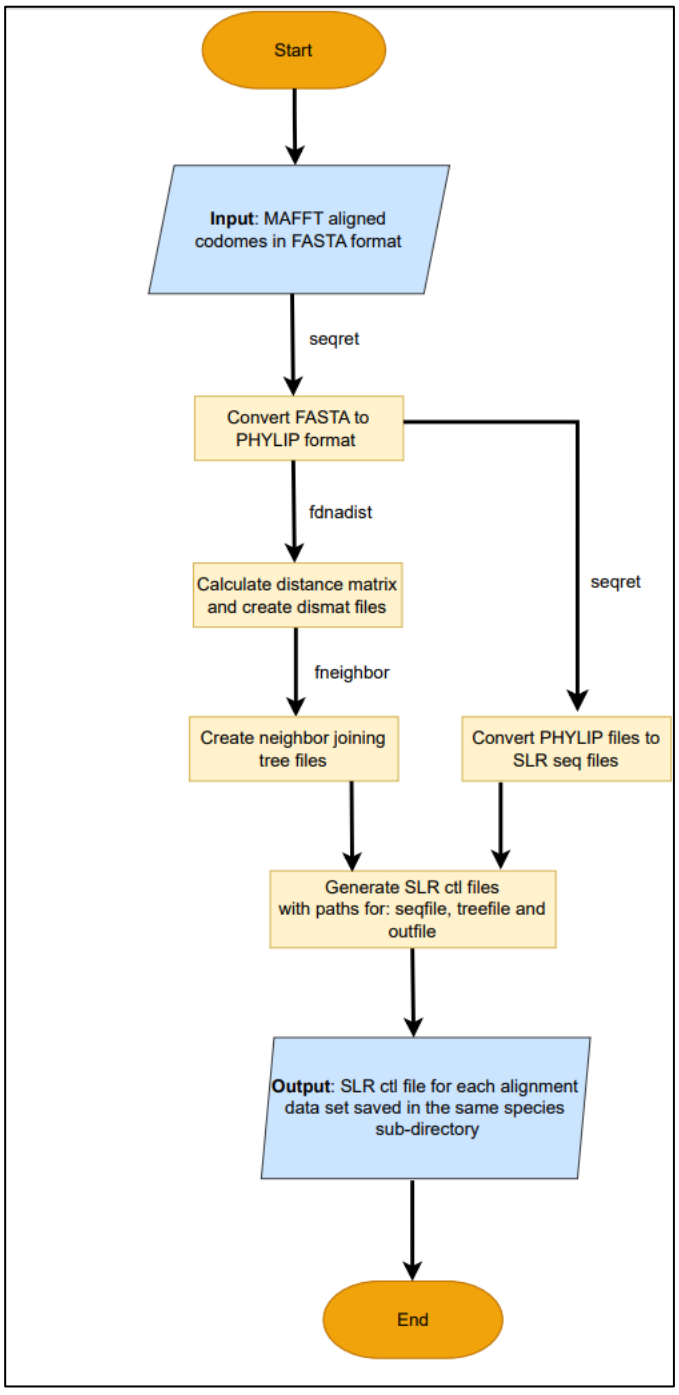

Figure 10: Scripts #13 to #17 process flowchart. Showing steps of file creation as a preparation for SLR analysis. Starting by FASTA format for multiple sequence aligned codomes, and ending up with directory contains subdirectories for each dataset with CTL and corresponding files for the same species for SLR run to follow.

- Scripts 18 and 19 flowchart: SLR run and output extraction

  1. Start:

  The process begins with a directory containing SLR .ctl control files.

  2. Loop through files:

  For each file ending in .ctl, the script applies the SLR command line tool to run the analysis.

  3. Generates SLR output files:

  For every alignment dataset, an .out file is produced and saved in the corresponding species subdirectory. These files contain all SLR results including sitewise likelihood ratios and statistical significance.

  4. Extract significant sites:

  The .out files are parsed to extract locations and counts of positively selected sites.

  5. Output:

  A summary spreadsheet is created, listing selection results across all species. This output includes the number of positive sites and their positions for each alignment.

  6. End.

  See figure 11 or scripts 18 and 19 flowchart.

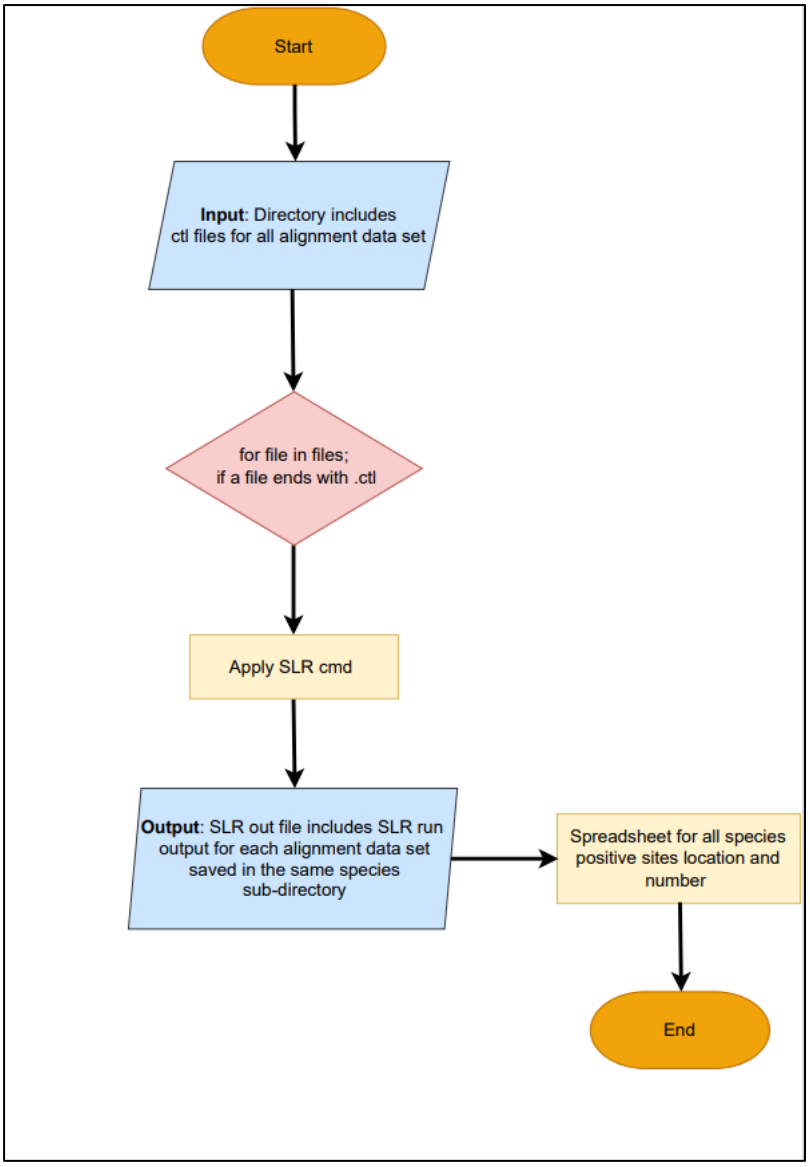

Figure 11: Scripts #18 to #19 flowchart. Showing SLR run performance through command and results extraction. Starting by input directory contains all CTL files for each alignment dataset in a subdirectory, and after SLR command run over all CTL files, the script ends with a spreadsheet for positive sites for all species data.

- Script 20 flowchart: Peptide location calculation

1. Start:

The process begins with a text file containing species names and GenBank accession numbers for all viral sequences found to have positively selected sites in the SLR analysis.

2. Import GenBank records:

Using Entrez Python package, all relevant GenBank records are retrieved based on the accession numbers provided.

3. Parse CDS features:

For each GenBank record the script examines the CDS feature subfields to determine the start and end positions of the coding sequence. These nucleotide positions are translated into peptide to match the amino acid scale used in SLR outputs.

4. Mature peptide

→if a mat_peptides is present, then parse the region.

5. Generate annotation files:

For each GenBank record, the script creates:

- File listing CDS and mat_peptide locations in peptide coordinates.
- Associated feature annotations from GenBank, including gene name, product, locus tag, protein ID and notes.

6. Integrate with SLR results:

The script uses the list of positively selected sites produced from script 19 and matches each site to its corresponding CDS or mat_peptide region.

7. Output:

Two spreadsheets are generated:

- One listing selected sites along with detailed CDS features annotations.
- Second for mat_peptide regions, where applicable.

8. End.

See figure 12 for script 20 flowchart.

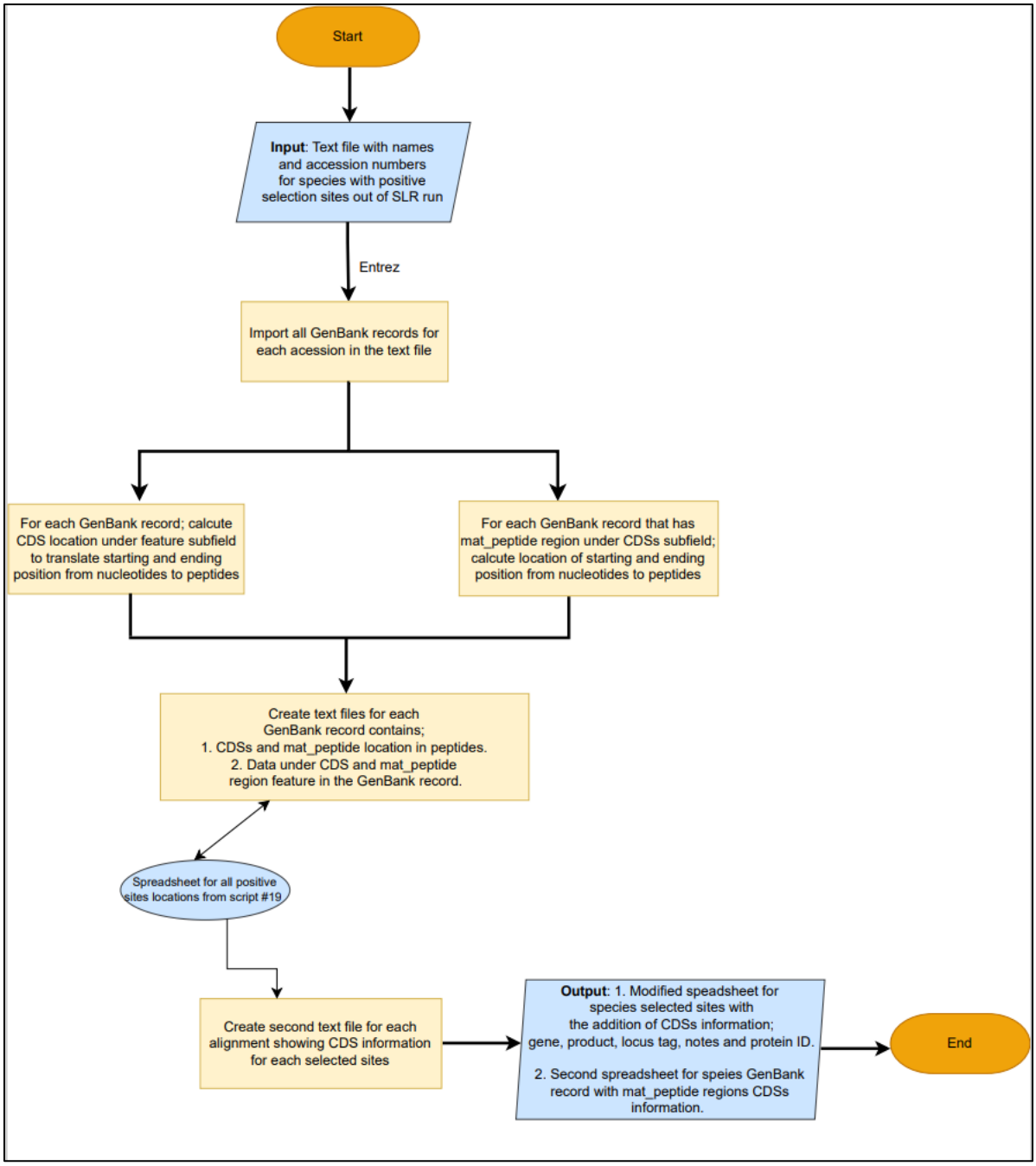

Figure 12: Script #20 flowchart. Showing workflow of peptide locations translation from GenBank records for all CDSs and mature peptide regions, with specific features recording. Starting by a text file contains all accessions for records found to have selection in the SLR run, and ending by an output of two spreadsheets with species with positive selection and corresponding features in CDSs, also for mat-peptides regions.

- Scripts 21 to 23 flowchart: Protein function and structure annotation

1. Start:

The input includes a text file of protein IDs and FASTA sequences for all selected proteins.

2. InterProscan annotation:

InterProscan is applied to the protein IDs, generating annotations that include:

- Gene Ontology terms.
- InterPro and Pfam domains.
- Protein lengths and names.

3. Functional data retrieval:

Additional information is fetched using:

- Quick GO (for GO terms definitions).
- Pfam-leacy (for clan and domain details).

4. FASTA splitting:

The full FASTA file is split into one .fasta file per protein, preparing for blast run.

5. BLAST search:

Using blastall command, each protein is searched against the PDB database to find structurally similar proteins.

6. Output:

- Spreadsheet detailing the GO and Pfam functional annotations for all selected proteins.
- Second spreadsheet listing PDB similarity scores for each protein.

7. End.

See figure 13 for scripts 21 to 23 flowchart.

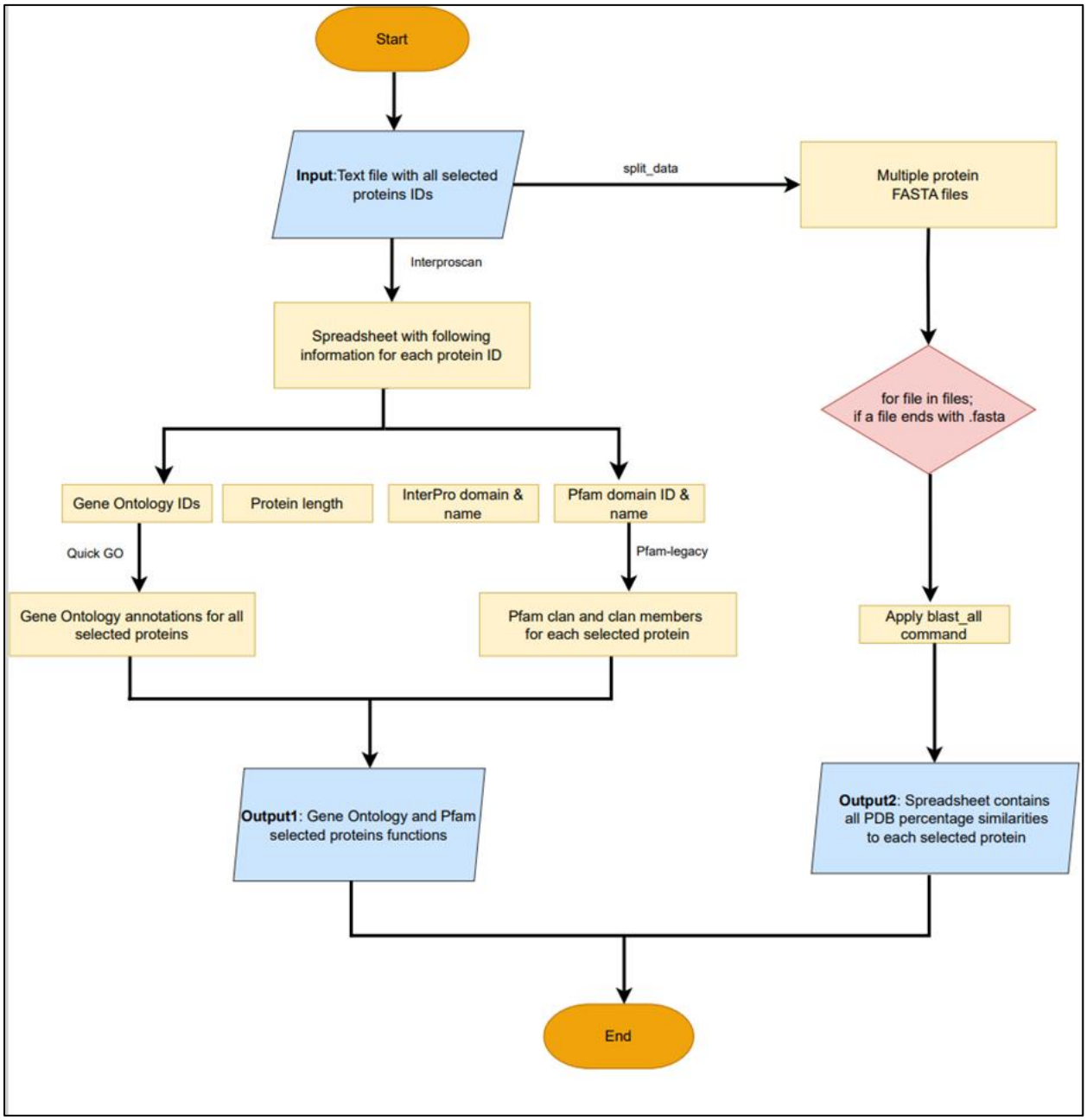

Figure 13: Scripts #21 to #23 flowchart. Showing workflow and steps of extracting functional annotations of proteins under positive selection. Starting by a text file with IDs for all proteins found to be under positive selection, and ending with two outputs the first is GO annotations with Pfam domains for protein list. The second output is a spreadsheet with PDB matches and similarities after blastall command applied.

2.2 Samples

2.2.1 GenBank records

Work was done on metadata from publicly available sequences, the main source of data is a GenBank record file. The viral section of GenBank, release 244.0 was downloaded from National Library of Medicine on 12th August 2021, sequences were initially in GenBank format.

GenBank records components: a GenBank record has many fields and sub fields for each record. Starting with Locus, where accession number can be found with the date of release and number of base pairs which reflect the whole length of the sequence, see Figure 14. Accession number is a unique primary key assigned to each GenBank record, which includes both a sequence and associated annotations. The accession number occurs on the ACCESSION line of a GenBank record and stay the same throughout its lifespan, even if the sequence or annotations change (Benson et al., 2012). Version number used to track sequence data changes, in case of any updates, a version suffix is added to accession number in the format "Accession.version", this identification appears on the VERSION line of the GenBank flat file. Source, where the name of the species can be found. Additionally, the same name of the species also can be found under Organism, with the full taxonomy hierarchy for the same species. Also, in features section there are more than a sub field as:

a) Source: contains the information on type of isolate, host, country of origin and collection date.

b) Gene: provides gene orientation and location in the DNA sequence in a starting and ending positions. Gene is recognized as a region of biological significance. The gene's extent is determined by the farthest 5' and 3' features.

c) "CDS": an identifier appears as a qualifier for coding sequence CDS feature which includes amino acid translation. Coding sequence with a range location for start and stop codons is the nucleotides part which matches with the amino acids sequence in a protein (National Library of Medicine, 1999). See Figure 14.

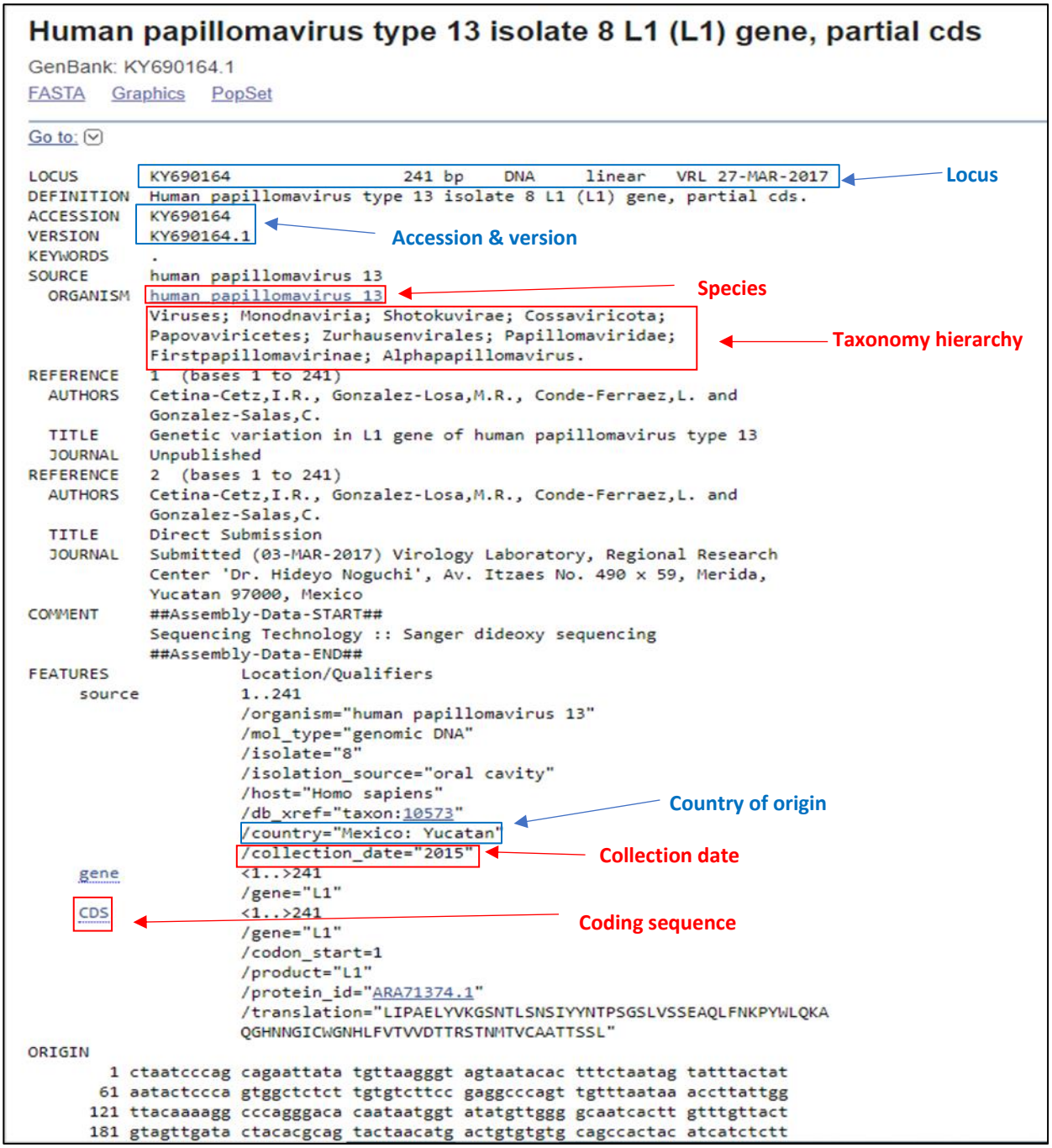

Figure 14: Example of a GenBank record for a viral sequence, illustrating annotation fields used in the sequence parsing workflow. The record shown corresponds to Human papillomavirus 13. Major sections of the GenBank flat file format are highlighted, including the LOCUS line (providing accession number, sequence length, molecule type, topology, division code, and date of submission), DEFINITION (a concise description of the sequence), ACCESSION and VERSION identifiers, and the SOURCE/ORGANISM information detailing taxonomic levels. Additionally, FEATURES outlining genomic elements with qualifiers such as gene name, protein ID, isolation source, host, geographic location, and collection date. In the annotated version, fields and subfields relevant to the filtering and classification pipeline are marked in blue boxes (e.g., host, collection date, gene name), while key functional annotations required for downstream analyses are marked in red boxes. These fields were programmatically extracted during the automated parsing process.

2.2.2 RefSeq and reference genome

A reference genome sequence is a high-quality, comprehensive representation of a specific organism's genome. It serves as a standardized framework for comparing and analysing the genomes of individuals or populations within the same species.

2.3 Data collection and structure

2.3.1 Download all GenBank viral sequences.

On August 2021 all viral GenBank records were downloaded through FTP with version number 2.44.

2.3.2 Download all reference viral sequences.

On September 2021, GenBank viral reference genomes were downloaded from NCBI.

2.3.3 Segmented Genomes

When a virus species has a segmented genome, every segment then was treated as a separate viral species corresponding to the matching reference genome to each segment. Segments species arrangement started by filtering all viral reference sequences to only include sequences with (NC_) in the accession number.

- CD-HIT cluster alignment

  A code was written in script #3 to automate CD-HIT, which is a program used to cluster and compare protein or nucleotide sequences to reduce sequence redundancy in large datasets (Fu et al., 2012). CD-HIT was used to remove duplicates in reference genomes, the input was all reference viral genome sequences, and the output was listing sequences accessions with NC_ header for each reference genome. This considered as an indicator

for quality or best curated reference genome. Figure 15 has examples of CD-HIT outputs for 3 references genomes identifying number of clusters in each one. The example on the left shows CD_HIT output for Zaire Ebola virus non-segmented genome with one cluster only, the following example in the middle shows Faba bean necrotic stunt virus segmented genomes with 8 clusters, and the last one at the right is the Rotavirus I segmented genome with 11 clusters.

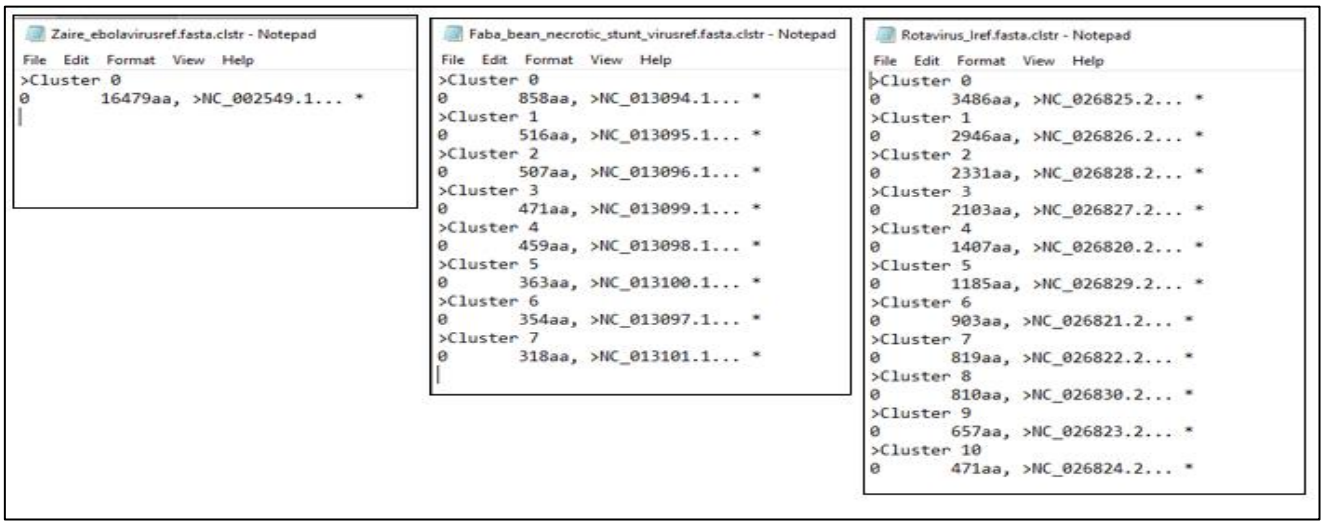

Figure 15: Screenshot for three viruses output from CD-hit showing ranges for number of segments starting with NC_, the first output from the left is the CD-hit output for Zaire ebolavirus a non-segmented genome with only one segment, the following is Faba bean necrotic stunt virus which its genome contain of 7 segments, and the last example from the right is Rotavirus I CD-hit output showing accessions of 10 segments.

- Directory files arrangements

  Later, using script #4 “Viruses” directory has a second arrangement different from the first taxonomic hierarchy classification, it has been modified to include only viral species with NC_ reference genome identifier. The first level in the directory has all viral species in directories, following level will contain FASTA files for viral sequences and FASTA files for NC_ reference sequences. This was done to distinguish segmented from non-segmented viruses, see Figure 16 for an example.

Figure 16: A screenshot of the second species directory arrangements according to presence of NC_ identifier in header, the example highlighted Abutilon mosaic segmented virus with two segments. Each segment has NC_ identifier which contain FASTA parsed Ref-codome. And the original FASTA file which has all parsed codomes for corresponding species.

2.3.4 Codomes creation

- Parsing GenBank records

  Parsing as a term is the process of studying character strings in order to link them to the underlying grammar's syntactic units. In computational biology, the term parsing refers to the process of extracting relevant data from files of various formats. As it executes an organised image of every feature and its relevant qualifier in GenBank then import them in an ordinary database management structure, extracting the needed data from a database entry for later analysis is a continual need in biological sequence analysis. This group of databases allows local control of GenBank entries, including indexing, recovery, and study of data and sequences on a computer (D'Addabbo et al., 2004).

- Date collected CDS from all viral genomes

  As described above, the viral GenBank records were downloaded, then script #1 was written to parse data with a specific information and save them in FASTA files. Among the listed fields and features in subfields section within one GenBank record as mentioned previously we are accepting only records with the presence of:

- Collection date: selected sequences should contain collection dates in features in order to be able to measure molecular clock speed (tempo).
- Coding sequence: collected FASTA sequences contains only the coding sequence part CDS and excluding the last stop codons. Coding sequence "codome" is needed to estimate the selection pressure accurately (mode).

Codome: CDS or CoDing Sequence, is a nucleotide sequence that matches to the amino acid sequence in a protein. Deciphering the entire coding potential or protein coding sequence CDS region of each gene is a critical step in the study of genomic information (Furuno et al., 2003). Through the project CDS is given a terminology "codome" which means the concatenated coding parts of the genome to allow analysis of selection in coding sequences, see Figure 17.

- CDS from reference genome sequences

  As viral reference genomes were used as a guide for further studies, Ref-genomes records were parsed using script #1 to produce reference codomes "Ref-codomes" that contains CDS part only.

- Taxonomic directory structure

  As an output of parsing script #1, taxonomy hierarchy, accession number, species name and country of origin, were collected from the GenBank record and saved in the structured format within a taxonomic directory structure. Later, using script #2 all Ref-genomes matches viral species were combined in one directory format for further methods applied.

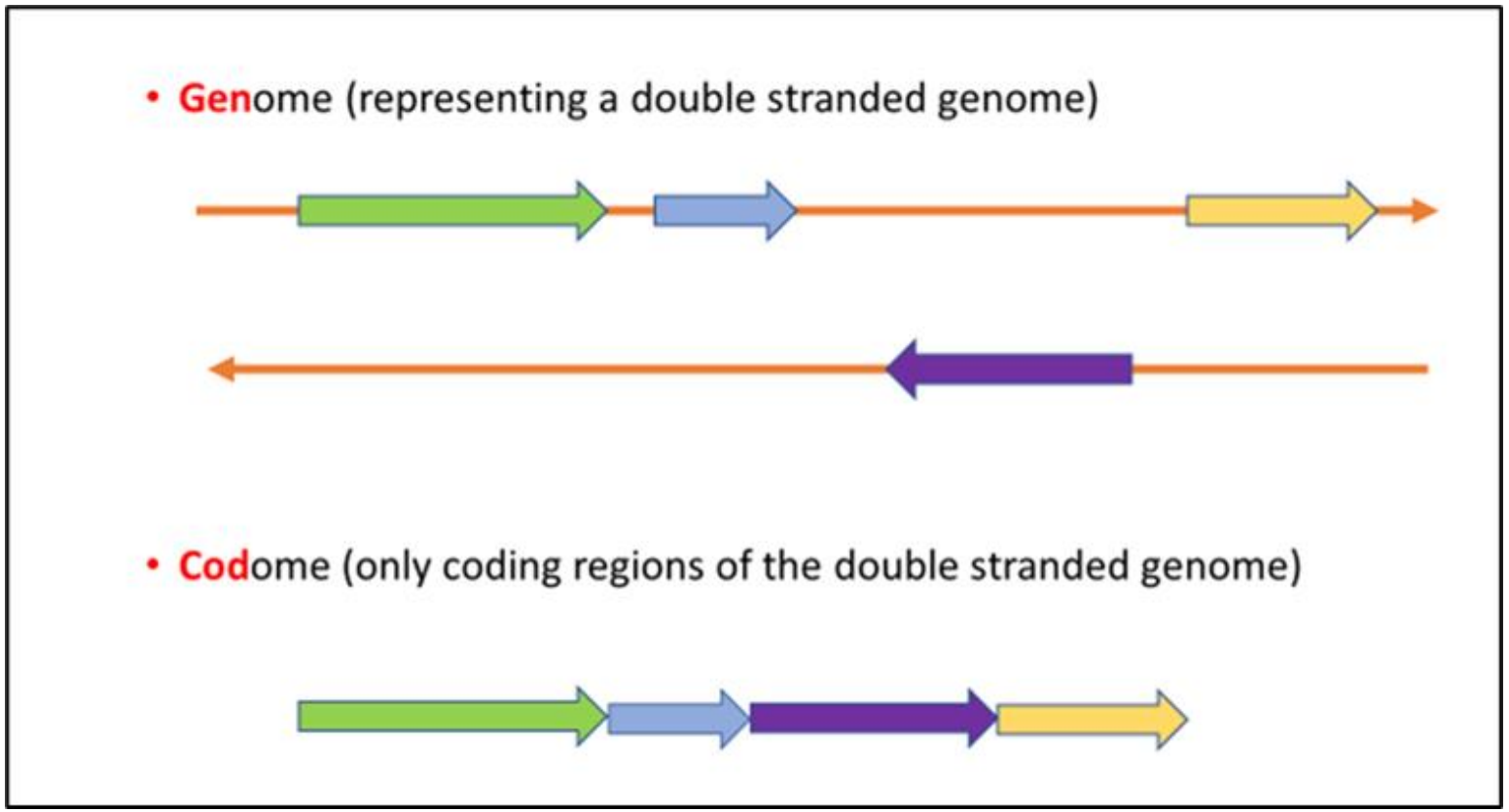

Figure 17: Example illustrating the process of creating “codomes” from viral genomes. First, all coding DNA sequences (CDSs) are collected from the annotated genome. In the case of double-stranded genomes, CDSs may occur on both the forward and reverse strands: forward-oriented genes (green, blue, yellow) are positioned above the genome backbone, while reverse-oriented genes (purple) are positioned below. For codome construction, reverse strand CDSs are reverse-complemented so that all coding regions are oriented in the same (forward) direction. Finally, all CDSs are concatenated in their genomic order to form a continuous sequence containing only coding regions, excluding non-coding intergenic sequences.

2.4 Alignments

2.4.1 Pairwise sequences alignments.

Pairwise sequence alignment is a bioinformatics technique that compares two biological sequences to identify regions of similarity or homology. It involves the systematic arrangement of the two sequences, considering insertions, deletions, and substitutions, to maximize the alignment score and identify conserved regions.

Once the new directory was arranged with considering each genome segment of any segmented viruses as a separate virus, a code was written to automate the pairwise sequence alignments.

- EMBOSS transeq

  EMBOSS (European Molecular Biology Open Software Suite) is a comprehensive bioinformatics software package that provides a wide range of tools and utilities for sequence analysis. One of the tools available in EMBOSS is "transeq", which is used for translating nucleotide sequences into their corresponding amino acid sequences.

  Rice et al. (2000) noted that transeq tool in EMBOSS performs six-frame translation, which means it generates all six possible reading frames (three forward frames and three reverse frames) from a nucleotide sequence. It identifies open reading frames (ORFs) within each frame and translates them into their corresponding amino acid sequences.

  Script #6 was used to automate EMBOSS transeq tool and apply it on the parsed and filtered codomes with their Ref-codomes and the output was saved in FASTA formatted text files, similar to the term given to all filtered and parsed genomes, follows to name the peptide translates as transcodomes and Ref-transcodomes.

- EMBOSS Needle

  In pairwise sequence alignment, there are various algorithms and methods used to calculate alignment score and identify the optimal alignment. One of the commonly used algorithms is Needleman-Wunsch Algorithm; This algorithm performs global

alignment by considering all possible alignments and calculating a dynamic programming matrix to find the optimal alignment. EMBOSS Needle is a method used for comparing amino acid sequences of two proteins computationally. It is feasible to tell from these results whether there is significant homology between the proteins. This data is important to track their potential evolutionary growth (Needleman and Wunsch, 1970).

- Aligning DNA files

  Script #7 used to automate Needle and to perform Needle pairwise alignment for each Ref-codome with its corresponding codome sequences. The output is later saved in a text file that lists the identity scores and similarity percentage for each alignment done.

- Aligning peptide files

  All transcodome sequences were also pairwise aligned with Ref-transcodomes using script #7 and a second text file is saved for protein identity scores.

- Troubleshooting

  Working through pairwise aligning, it was noticed that parsed sequences from GenBank records contain a number of errors that was solved by editing the main parsing script, the most occurring one was:

  - Stop codons

    When needle run through script #7 and start to produce an output, low identity scores percentage appeared in transcodomes with Ref-transcodomes aligning comparing to codomes with Ref-codomes aligning, the reason for these low similarities are caused by the presence of stop codons in peptide translated sequences. This was sorted by considering some points in concatenating CDSs and codome creation, such as:

    a. Presence of N residues in nucleic acid sequence which reflects a gap or a missing region, this was solved by re-parse records and remove codomes with N count more than 25% of the total number of nucleotides.

b. Consider the value of codon start when saving the sequence, which appear in the GenBank record under CDS feature. Going back to each GenBank record and specifically under CDS sub-feature there is term “codon start=” with a numerical value after =, this value is then counted when parsing records and start concatenating CDS to save each codome sequence.
c. Parse and save codome sequences in a multiple of 3 values.

2.4.2 Filters

- Filters prior pairwise alignment

  In order to reduce time consuming computationally with alignment runs, and to guarantee higher identity matches with less gaps at sequences pairwise alignments, additional filters were applied:

  Codome sequences length: a threshold of codome to Ref-codome lengths was choose >=75% and <=100%, this means for any codome to run a pairwise alignment with its Ref-codome it should be 75% or more from its Ref_codome length and not more than the Ref-codome length. This filtration threshold was chosen starting from 75% to ensure that codome is sufficiently complete comparing to its Ref-codome and an upper bound of 100% was chosen to avoid over-extended sequences. This was done using script #5.

  Reference genome size: any "NC_" Ref-codome should not be more than 50kb. This threshold was applied to exclude too long reference genomes that might affect alignment quality. This was done using script #5. Figure 18 shows the final directory arrangement for one viral species after applying filters, each segment will have a separate directory where FASTA files are saved:

  a. Ref-codome FASTA sequence with genome size filter applied as mentioned.

  b. Codomes FASTA sequences after applying sequences length filter as described previously. Figure 18 show species African swine fever virus as an example which has 9 segments.

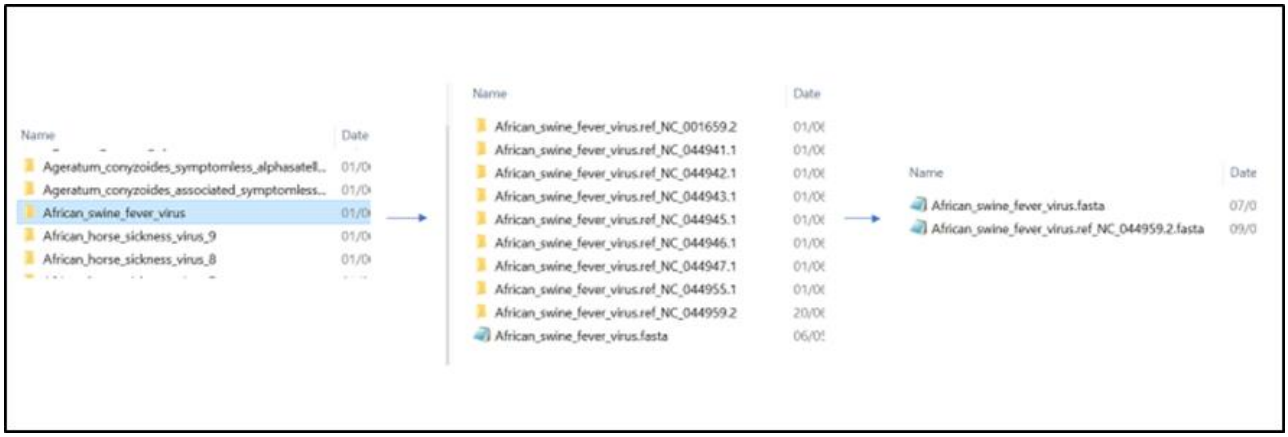

Figure 18: Species directory arrangements after applying filters and before needle alignment takes place. In this directory each subdirectory represents a segment which is considered a separate species and include two FASTA files; Ref-codome sequence and codomes follow length filters.

- Filters after pairwise alignment

  Once pairwise alignment is completed for all codomes with Ref-codomes and transcodomes with Ref-transcodomes, a range of identity and similarity scores was recorded. Additional filters were applied for each sequences data set to remove low identity matches:

  Peptide and DNA level: using script #8, sequences were filtered and chosen according to:

  a. On nucleic acid level codomes to Ref-codomes alignment scores, sequences scored 90% to 100% identity scores on Needle alignment were selected. This filter was applied to keep codomes which are sufficiently similar to their reference, to minimize alignment artifacts and support accurate evolutionary parameters estimation.
  b. Sequences scored between 50 to 90 at nucleic acid alignment, were referred to peptide identity level transcodomes to Ref-transcodomes to only save sequences scored 90% and more for peptide needle alignment. This additional identity scores were added to increase number of sequences with high peptide level identity, assuring they are functionally and structurally equivalent to their reference. see Figure 19.

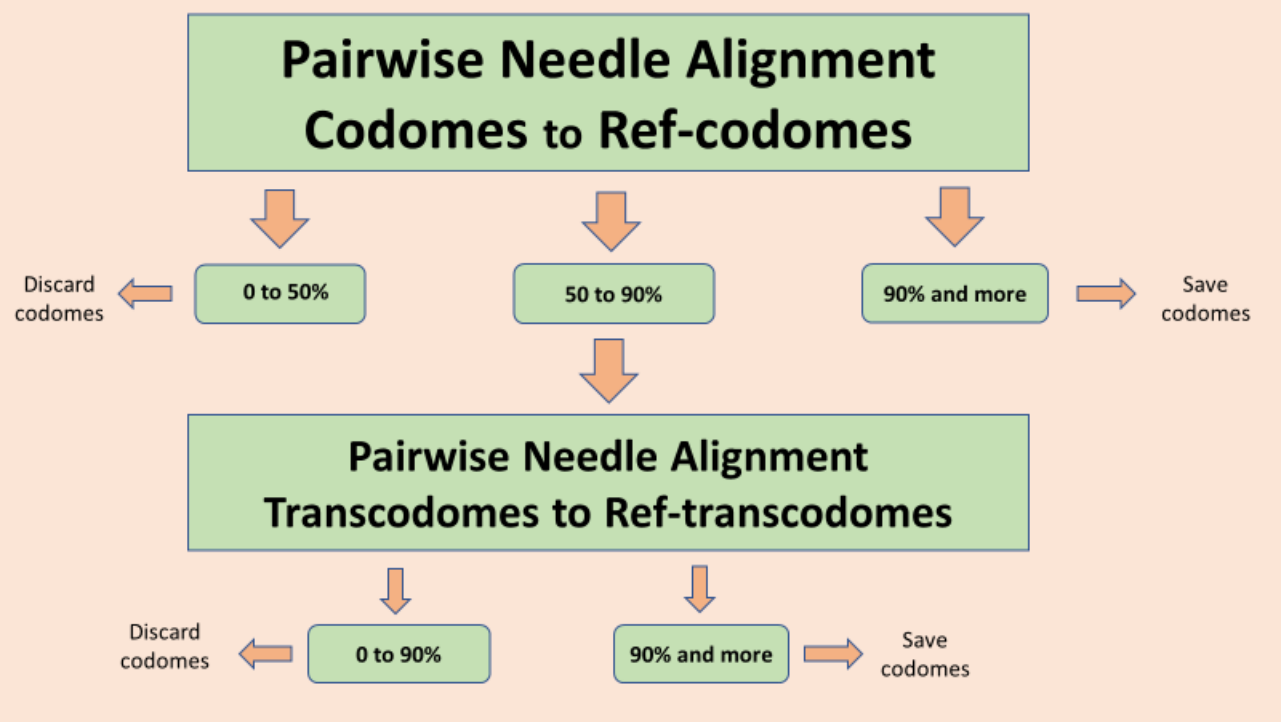

Figure 19: Flowchart illustrating the filtering process for viral sequences based on identity scores obtained from Needle pairwise alignments at both the nucleotide and peptide levels. In the first stage, at codome, sequences with identity scores between **0–50%** are discarded, those between **50–90%** proceed to a second filtering stage, and those with **≥90%** identity are saved directly. In the second stage, the corresponding transcodome sequences are aligned against Ref-transcodome, and only those with **≥90%** identity is retained.

- Date gap and number of sequences

  on this level, number of data set alignments were achieved, each one has a range of number of sequences. Further filters applied on them where:

  a. Date Gap: within parsing script #1 each sequence collection date was saved in an external file, then script #9 was used to save earliest collection date and a latest collection date for each alignment sequences set separately. Following, a date range was calculated for each aligned sequences data set. The filter applied was a date gap of 5 years and more, any dataset years gap less than 5 years was discarded. This age gap was chosen to ensure sufficient temporal signal obtaining reliable substitution rates.

  b. Number of sequences: on each alignment data set sequences number ranges from 10 to 50, any dataset with sequences less than 10 or more than 50 were discarded.

This cut-off was chosen to guarantee that parameters are studied on a phylogenetic reliable and statistic meaningful data.

- Removal of "well-studied" viruses

To avoid overrepresenting heavily studied viral families and ensure balanced taxonomical coverage of alignment sequences analysed, a manual removal step was performed. Some viruses as Orthomyxoviridae (e.g., Influenzae viruses), Caliciviridae, and Coronaviridae (e.g., SARS-CoV-2), appeared in high number of parsed sequences in the initial step. As this project aims to explore evolutionary patterns across diverse virus families, the top two hits (Orthomyxoviridae and Caliciviridae) were excluded to allow more contribution of other families with less sequences submitted on databases. Additionally, retroviruses (e.g., HIV from Retroviridae) and adenoviruses were removed once filtering criteria mentioned previously were applied for the reference genome size threshold, as their Ref-genome exceeded 50kb.

2.4.3 Multiple sequence alignment

Multiple sequence alignment is a computational technique used to align and compare biological sequences typically DNA, RNA, or protein sequences. It involves arranging sequences in a way that maximizes the identification of conserved regions, insertions, deletions, and other sequence variations, providing insights into evolutionary relationships, functional motifs, and structural features (Katoh et al., 2002; Katoh and Standley, 2013).
After pairwise alignment was performed, MAFFT was chosen to perform multiple sequence alignments for all viral codomes in each alignment data set using script #10.

2.4.4 MEGA for quality checking

Once alignments for each viral sequences data set were ready, a manual quality checking on MEGA (Molecular Evolutionary Genetic Analysis) software took place, checking started by examining any presence of stop codons, gap positions, frame shifts and poorly aligned regions.

MEGA provides a range of functionalities for conducting various types of molecular evolutionary analyses, including phylogenetic reconstruction, sequence alignment, evolutionary distance estimation, and hypothesis testing (Kumar et al., 1994; Tamura et al., 2021).

2.5 Recombination

Following sequences alignment manual checking, alignment data sets went through second manual quality checking for the presence of recombinant sequences.

2.5.1 Simplot

Each alignment data set was examined for the presence of recombination sequences using Simplot version 3.5.1 with Bootscan at different window sizes and nucleotide steps. Simplot study and visualize the evolutionary history and recombination events in nucleotide sequences. It is particularly useful for studying the genetic diversity and recombinant origins of viral genomes (Lole et al., 1999).

Simplot Bootscan is a computational method that employs a sliding window approach to analyse nucleotide sequences for evidence of recombination. It examines the similarity between the query sequence and a set of reference sequences along the genome, allowing the detection of potential recombination breakpoints (Rubio et al., 2014). Below are strategies followed on Simplot Bootscan runs:

- Step size=20 bp.
- Strict consensus
- Bootscan was not performed on alignment length less than 250 nt.
- Window size starts from 250 bp and increases according to alignment length. See Figure 20 for an example.

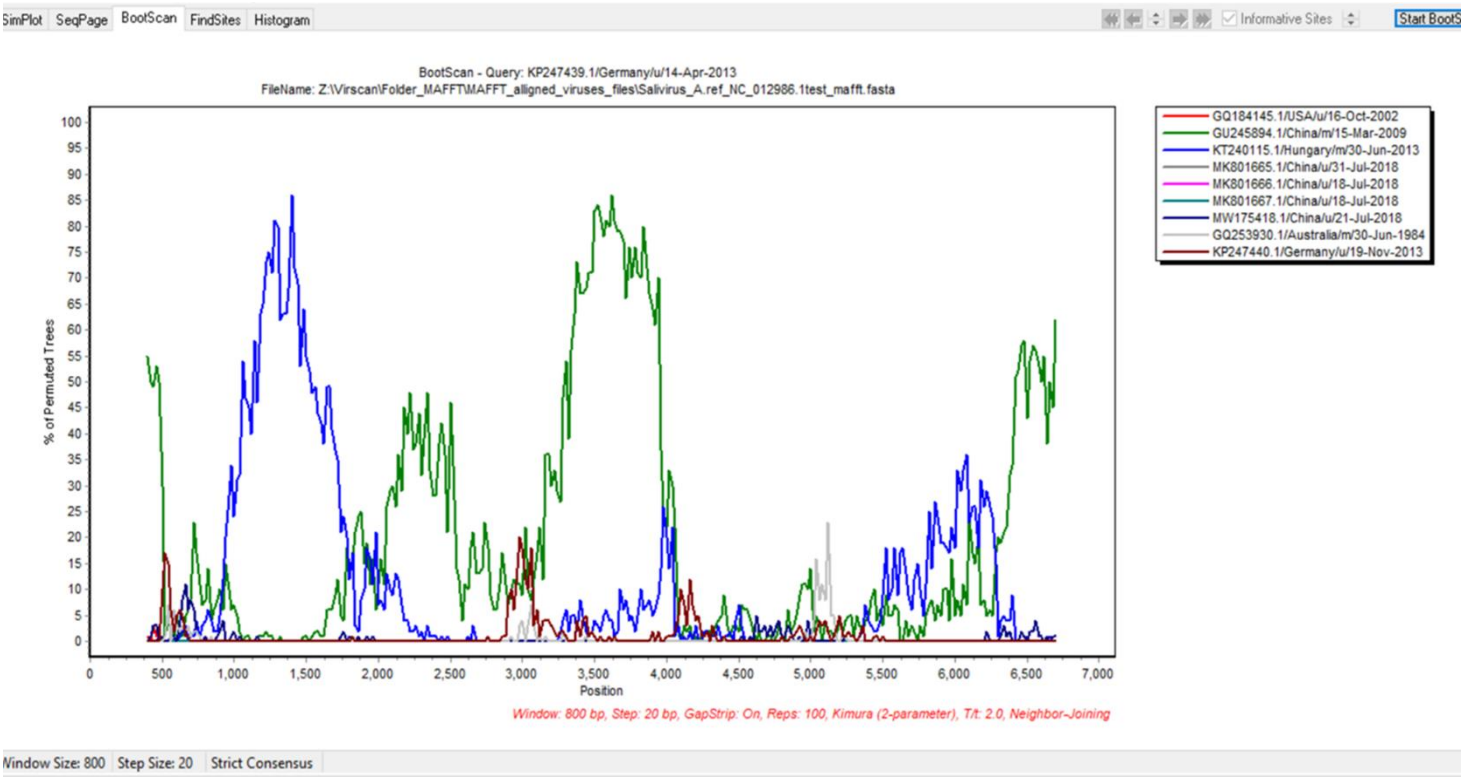

Figure 20: Example output from a Bootscan analysis illustrating potential recombination events in a multiple sequence alignment. The plot shows the percentage of permuted trees supporting different parental sequences across the genome positions of the query sequence (*KP247439.1/Germany/u/14-Apr-2013*). Crossovers between coloured lines indicate potential recombination breakpoints, where sequence similarity shifts from one parental strain to another. Analysis parameters included a window size of 800 bp, step size of 20 bp.

2.6 Clocks

Manual quality checking continues and after recombination, clocklike behaviour, or temporal signal for all sequences in each viral MAFFT aligned data set with known dates, was examined. For this to be done, phylogenetic trees were built and created for the aim of predicting relation between genetic distance and time.

2.6.1 Alignments tree building

Neighbour Joining trees (NJ) were built for each alignment set using MEGA software. Neighbour Joining is a popular algorithm used for constructing phylogenetic trees, based on genetic distance matrices. The Neighbour Joining algorithm follows a

recursive approach to build tree by iteratively joining pairs of sequences or subtrees based on their pairwise distances (Gascuel, 1997).

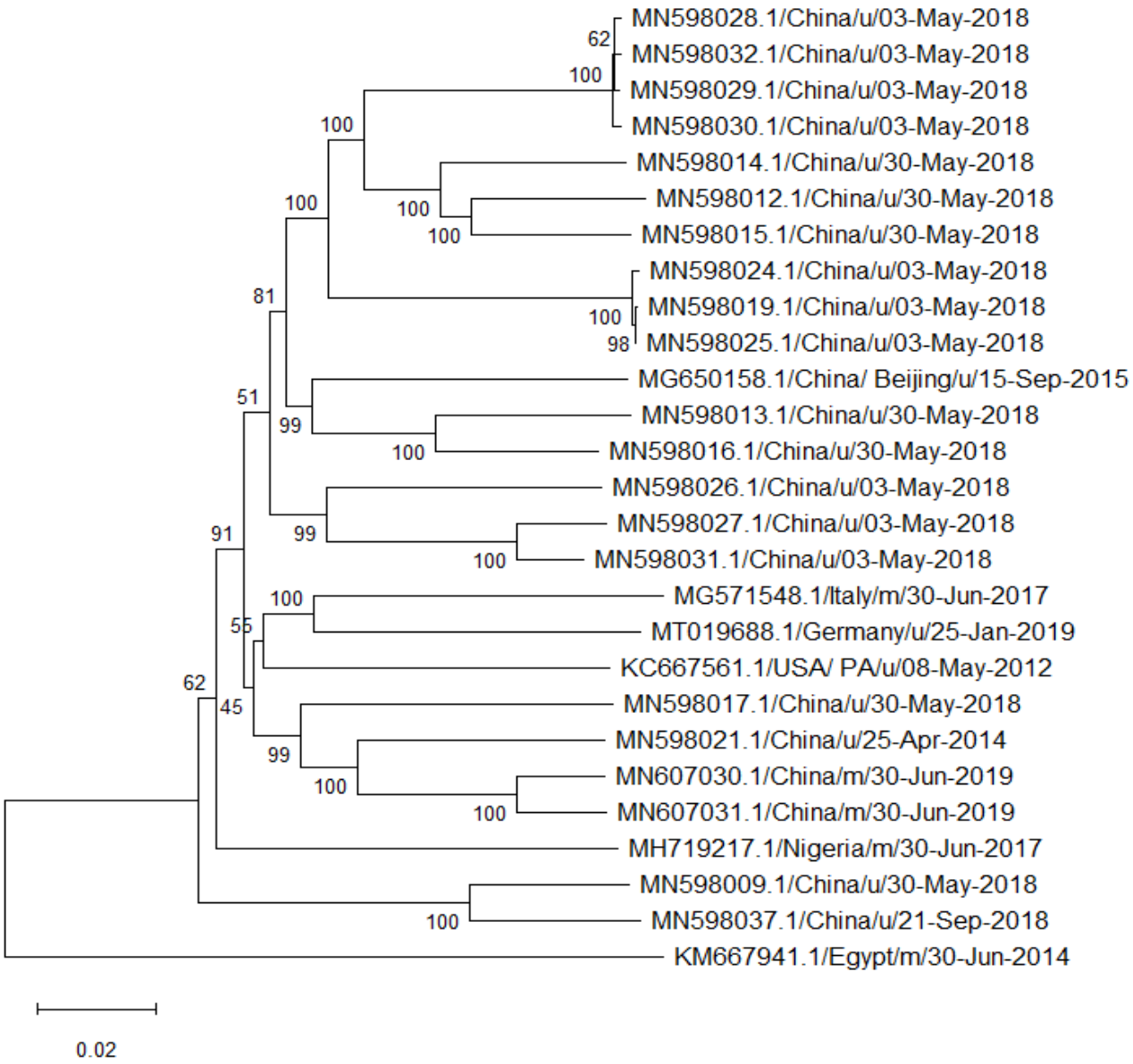

Figure 21 A: Neighbour Joining tree of Enterovirus E codome sequences (27 sequences).

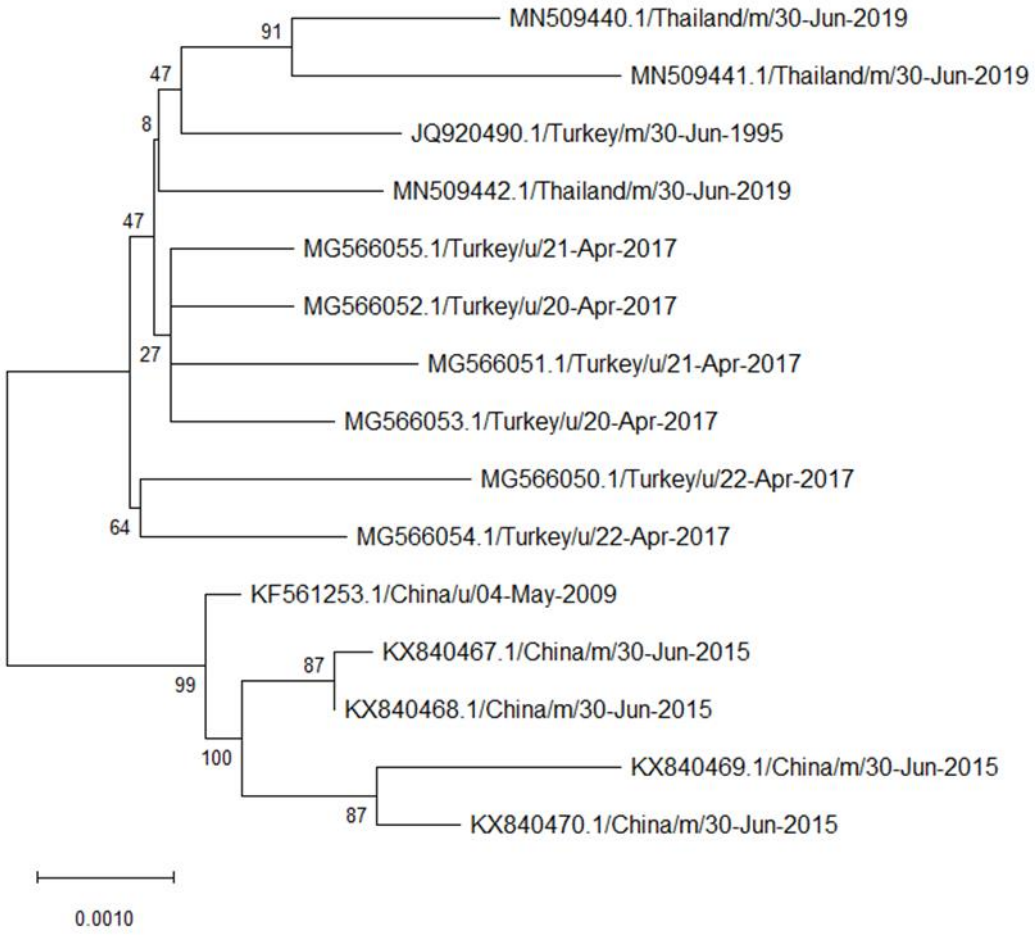

Figure 21 B: Neighbour Joining tree of Citrus chlorotic dwarf-associated virus codome sequences (15 sequences).

Figures 21A and 21B represents a Neighbour Joining phylogeny of two viruses from the filtered dataset alignments, and ready to start tempo and mode analyses. Those trees constructed based on MAFFT aligned nucleotide sequences. Bootstrap values are shown at nodes to indicate branch support, and scale bar represents genetic distance, figure 21 A is the phylogenetic tree for Enterovirus E, 27 codome sequence alignment, and figure 21 B is the phylogenetic tree of 15 codome sequences for Citrus chlorotic dwarf-associated virus

2.6.2 NWK format files created

The constructed phylogenetic trees were converted to Newick format. Newick trees format is used for representing phylogenetic trees in a compact and readable manner. It is a plain text format that represents the hierarchical relationships among taxa or sequences in the tree, as well as branch lengths or support values, if available (Cardona et al., 2008).

2.6.3 TempEst

This is the last part on alignments quality checking done manually. Once Newick formatted tree for each alignment was produced, it works as the input file for TempEst software, where correlation coefficient R values are recorded for each alignment data set separately.

2.7 Rates

Now alignment data sets are ready to study tempo and mode. The first part in the evolutionary analysis will be substitution rate measuring, starting by creating XML files for each single alignment:

2.7.1 XML files

XML (eXtensible Markup Language) is the input file format for BEAST (Lin et al., 2005). XML creation starts by using BEAUTi software, BEAUTi takes one alignment data set as an input and produce .xml output file that stores all data for BEAST software needs to run. Start by importing FASTA format data set, then specifying data and alignment, followed by setting up the evolutionary model to produce XMLs. All selected features in BEAUTi are described in steps below:

a. Tips: parse tip dates as a calendar date format dd-MMM-yyy.
b. Sites:

- Substitution model : GTR (Barba-Montoya et al., 2020; Tavaré, 1986).
- Site heterogenicity model: gamma + Invariant sites.

c. Clocks: uncorrelated relaxed clock with lognormal relaxed distribution.

d. MCMC: chain of length chosen 1,000,000,000.

Later, when XML features are satisfactory over BEAUTi, script #11 was written to automate XMLs formation copying the features and output produced by BEAUti and applied on all alignment data sets available.

2.7.2 BEAST

- Beagle library

  For BEAST recent versions to run over Linux, BEAGLE library is obligatory (Ayres et al., 2012; Baele et al., 2019).

- BEAST 1.10.4

  BEAST run from the command line after successful installation on Linux, script #12 written to run BEAST on all aligned data sets XMLs at once.

- Analysing output through tracer

  While BEAST running, output generated is saved as ".log" file for each alignment and then analysed by tracer software package version 1.7.1 (Rambaut et al., 2018).

2.7.3 Limitations and considerations

Bayesian phylogenetic analysis implemented in BEAST, uses Marcov Chain Monte Carlo (MCMC) to estimate the posterior distribution of model parameters, including evolutionary rates and tree topologies, given the observed sequence data and prior assumptions (Drummond and Rambaut, 2007). However, BEAST is sensitive to number of sequences and temporal range. Datasets with very few sequences, recombinant sequences or insufficient sampling across time can lead to wide posterior distribution or failure to achieve

convergence. Therefore, starting by MCMC chain length of 5 million was tested then increased to 1,000,000,000 in the final analysis, with Effective Sample Size (ESS) >200 as a threshold for reliable estimates.

2.8 Selection

2.8.1 SLR version 1.4.3

For SLR software to run in terminal, number of commands is written and run on each aligned data set subsequently as follows:

- Phylip file creation

  Starting by converting all MAFFT aligned FASTA sequences to PHYLIP format, script #13 was used to automate PHYLIP files generation for each aligned data set. A further modification of PHYLIP format files is then used as SLR input (Retief, 2000).

- Dismat file creation

  Following PHYLIP files creation, distance matrix was calculated by creating dismat files using fdnadist program through script #14.

  The output of fdnadist is a distance matrix, which is a symmetric matrix that shows the pairwise genetic distances between sequences (Joe Felsenstein, 2004).

- Tree file creation

  Once dismat files created, generation of tree files proceeds using fneighbor. The fneighbor program implements the Neighbor-Joining (NJ) algorithm, which construct phylogenetic trees based on genetic distance matrices. After running fneighbor by using script #15 with distance matrix as input, tree files were produced containing the inferred phylogenetic tree in Newick format.

- Sequence files creation

Additionally, using script #16 with an input PHYLIP files previously generated, the code used seqret command to convert PHYLIP files to in files or sequence "seq" files. Seqret program is a part from EMBOSS package used to manipulate and convert sequences between different file formats, facilitating interoperability between various bioinformatics tools and databases (Rice et al., 2000).

- Control files

  Once all file types mentioned above were created and saved, script #17 was written to generate a control "ctl" file for every alignment, each "**.**ctl" file contains path of seq file, tree file and out file for all alignments data sets.

  Later, within script #18 SLR command was applied over all "slr.ctl" files present, and the output file saved as text out files.

2.8.2 SLR out files analysing

The “out” file generated for each alignment data set was saved as a text file in species directory and could be analysed separately or alternatively, script #19 was written and applied on all out files to save results in a spreadsheet that contains:

- Positive selected sites

  Presence of selected sites in every alignment, and number of positive selection if available, also site number of all positive selected sites appeared in peptide sequence.

- Positive sites product and gene

  Knowing the location of positive selected sites in the peptide sequence, script #20 was written to perform:

  a. Peptide location calculation

     GenBank records for all alignments with positive selection output of SLR were collected and script #20 was applied to translate CDS nucleotide bases location to peptide locations, the output for the code is specifying location of coding regions starting and ending numbers in amino acid bases.

  b. Define protein ID for positive sites.

Once peptide translate was determined in starting and ending positions, selected sites protein IDs, gene and more additional information as locus tags and notes were copied from GenBank records.

c. Mat-peptide positive sites locations.

GenBank records with mat-peptide sequence location goes under peptide locations calculation separately in order to get more details for regions with selected sites.

d. Debugging files with join CDS

Some records with joined CDS locations were calculated manually for peptide locations ranges to avoid overlapping.

- Polymorphism analysis

  Positive selected sites identified through SLR run were filtered to include only multiple plus signs (++, +++, and more). Peptide alignments for the selected sites were visualized using MEGA to count variable amino acid presence across sequence.

2.8.3 Manual alignment: ClustW

Noticed minor number of alignments has no out files output after SLR code was applied, this was due to presence of stop codons in the peptide sequence or some out of frame alignments, SLR was repeated after performing manual sequence alignment with ClustalW which then was followed with manual checking (Larkin et al., 2007; Thompson et al., 1994).

2.8.4 Positive selection and Bottleneck effect

Although sitewise dN/dS ratios were used to detect evidence of positive selection, it is recognized that sequence enrichment can also result in from non-selective processes such as population bottlenecks. Bottlenecks reduce genetic diversity by randomly restricting the number of variants that persist, which can mimic the reduced variability seen under strong purifying or directional selection. In contrast, diversifying positive selection actively increases sequence variation, a pattern that cannot be explained by bottlenecks. To distinguish adaptive signals of positive selection from patterns caused by population bottle necks, filtering criteria as pairwise identity threshold and temporal diversity, were applied

to minimize the inclusion of highly enriched datasets. These steps helped reduce the risk of false positives in selection ensuring that the detected signals are more likely to reflect true adaptive evolution.

2.9 Functions

2.9.1 Domains

- InterProscan database protein run

  Later, when positive selected sites have known protein IDs, all selected proteins were listed in a text file and script #21 run to extract all InterProscan data in the output file (Paysan-Lafosse et al., 2023; Zdobnov and Apweiler, 2001).
  The output was saved in a spreadsheet with the following information for each protein ID:
  - Pfam domain ID and name (Mistry et al., 2021; Sonnhammer et al., 1998)
  - Protein length
  - E Value
  - InterPro domin ID and name
  - Gene Ontology IDs

- Protein Ids with no InterProscan output
  Number of the selected proteins list has no data in InterProscan.

2.9.2 Gene Ontology

- QuickGo for function identification
  Based on the output of InterProscan, all selected proteins with GO IDS were collected and checked on QuickGo to access Gene Ontology (GO) annotations. QuickGO retrieves GO annotations for genes and proteins. The Gene Ontology categorizes biological knowledge into three main categories: molecular function, biological process, and cellular component (Gene Ontology, 2012; Gene Ontology et al., 2023).

- Pfam domain name and clan member

  Additionally, Pfam IDs of selected proteins were collected and proceeded to search for Pfam clan and clan members for each ID. A clan is a group of protein families that share a common evolutionary origin and structural or functional characteristics, every clan consists of multiple individual protein families, known as clan members which share significant sequence and/or structural similarities (Finn et al., 2006).

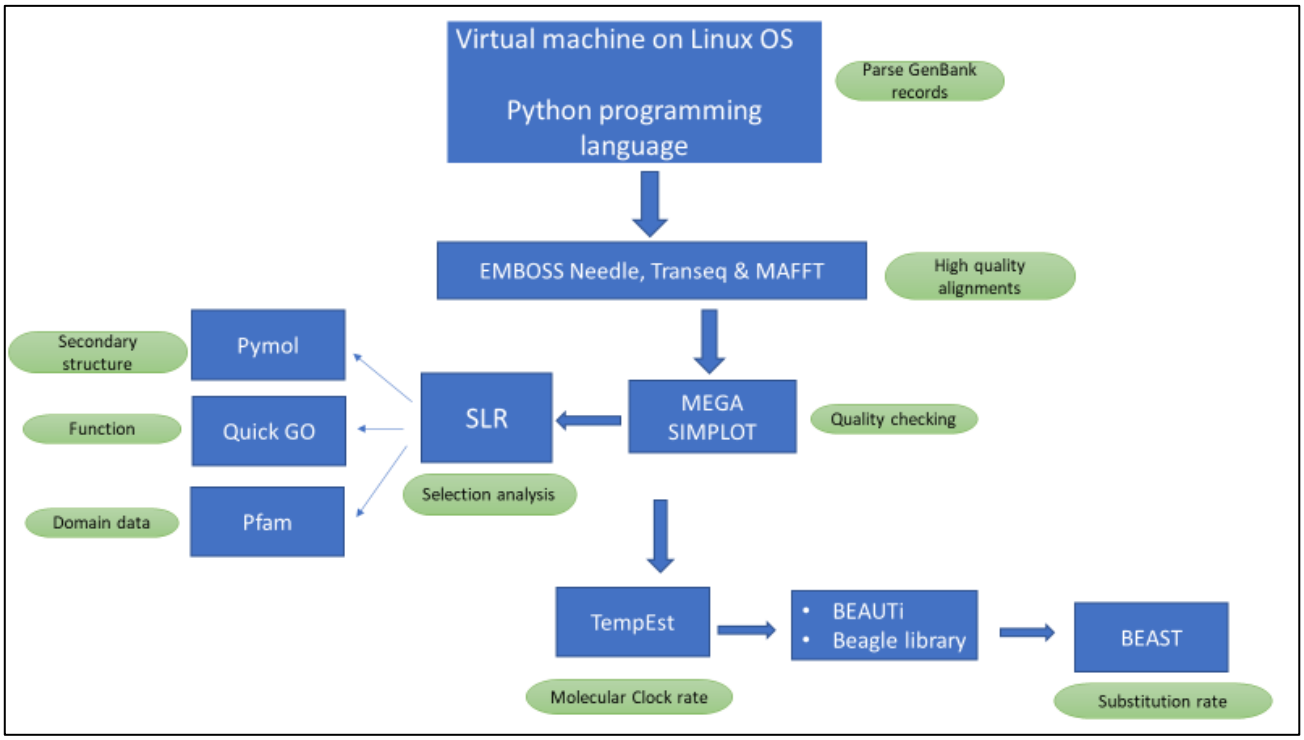

Figure 22: Workflow of software tools used in the study, from data parsing using Python and sequence alignment with (EMBOSS Needle, Transeq, MAFFT) to quality checking (MEGA, SimPlot), selection analysis (SLR with QuickGO, Pfam, MOE), and evolutionary rate estimation (TempEst, BEAST and BEAUti).

2.10 Structure

2.10.1 Blastall

Having the list of proteins IDs with positive selected sites, script # 22 was written to split the FASTA file which has all selected protein sequences to multiple files. Later, script #23 followed to apply blastall command on protein IDs as entries to produce an output for all

PDB local database matches (Berman et al., 2000; Zardecki et al., 2022). A list for each protein ID with blastall high similarity matches is then collected for protein identification in order to visualize some viruses form the alignment data sets with positive selected sites in different proteins functions for tertiary structure.

2.10.2 Molecular Operating Environment

Molecular Operating Environment (MOE) is a molecular modelling, machine learning, and simulations software platform developed by Chemical Computing Group (CCG) (Labute et al., 2002; Vilar et al., 2008). MOE provides tools for bioinformatics, molecular modelling, and computational chemistry, including 3D molecular visualisation, molecular editing, sequence alignment and energy minimization.

Choosing 4 protein IDs with different GO functions and a known solved structure where the positive selected sites are present, for each 100% similarity match protein a PDB format file is downloaded and visualised over Molecular Operating Environment MOE from Chemical Computing Group software as follows:

- Selected residues displayed for visualisation.
- Mutant amino acids are added through protein builder to additional chains which are aligned and superposed with the original chain.
- Each mutant residue is then subjected to energy minimization on a forcefield of Amber10 within 4.5 radius to ensure structure stability.
- Molecular surface created for specific residues for comparison with a colour coding based on Lipophilicity (Heiden et al., 1993).

Chapter 3: Results

3.1 GenBank records parsing and filtering

3.1.1 Number of viral genomes parsed

Following work mention in Methods, script #1 was applied on viral GenBank records, parsed codomes were produced and saved in "Viruses" taxonomic hierarchy directory. Additionally, statistics were produced listing values for viral records number in the initial GenBank input file and number passes the primary filtering criteria for each species. Total number of records are described in Table 3.

Table 3: Represents number of sequences pass the parsing filtering criteria in all viruses GenBank records.

Process No.	Process Description	Number of Sequences	Number of Distinct Species
1	Download viral GenBank v2.44	4,0402,060	183,382
2	GenBank records CDS + collection date	2,649,652	156,482

3.1.2 Number of reference genomes parsed

Furthermore, part 2 of script #1 generated the output of Ref-codomes and saved in the same directory according to taxonomic hierarchy. Number of reference records passed the filtering criteria shown in Table 4.

Table 4: Represents number of sequences pass the filter in all viral Reference genomes GenBank records.

Process No.	Process Description	Number of Sequences	Number of Distinct Species
3	Download viral Ref-genomes	53,696	4,368
4	Ref-genomes + CDS	50,713	3,675

3.1.3 Taxonomic distribution for records passed the filtering criteria

- Taxonomic distribution for codomes species after initial parsing

As a part of script #1 output, a text file for all saved codomes species taxonomy hierarchy was produced, 1502 unique taxonomies were present out of 156,482 distinct species. I proposed to call them unique taxonomies since each taxonomic structure has no repetition, a unique taxonomy can be defined as the specific combination of taxonomic ranks at each level in the hierarchy. For example, the unique taxonomy for Influenza A virus will be "Viruses (superkingdom); Riboviria (clade); Orthornavirae (kingdom); Negarnaviricota (phylum); Polyploviricotina (subphylum); Insthoviricetes (class); Articulavirales (Order); Orthomyxoviridae (Family); Alphainfluenzavirus (Genus); Alphainfluenzavirus influenzae (Species)" which means this unique taxonomy is associated with all records and sequences related to this specific viral species. It is important to understand hierarchy variability and that not all taxonomic classifications follow the same hierarchical structure, and the presence of a species, genus, or other ranks can vary depending on the group of viruses being classified. In many cases, the taxonomic hierarchy may not extend all the way to the species level, some may only be classified up to the family, order, or genus level. Also, taxonomic classifications may include categories labelled as "no rank", this could be due to represented taxa do not comply with standard ranks or uncertain classification like genus or species. For example, the taxonomy hierarchy for Anolis sagrei adenovirus 1 will be "Viruses (superkingdom); Varidnaviria (clade); Bamfordvirae (kingdom); Preplasmiviricota (phylum); Tectiliviricetes (class); Rowavirales (Order); Adenoviridae (Family); unclassified Adenoviridae (No rank)" (Schoch et al., 2020).

Within the 1502 unique taxonomy hierarchies for codomes after parsing, 134 distinct families were recorded. However, the number of families occurrence in these taxonomies varies significantly. At the top end of the spectrum, 'Orthomyxoviridae' stands out with 119,924 hits, indicating its prevalence and clinical importance, followed by 'Caliciviridae' with 15,196 hits. Conversely, there are ten families at the other end of the spectrum that have only one hit e.g.,

'Clavaviridae' and 'Mononiviridae', suggesting a much lower level of representation or occurrence within the dataset. These variations in hit counts across the 134 families provide valuable insights into the diversity and distribution of viral species within GenBank. Figures 23 and 24 are pie charts for family level taxonomy distribution among parsed codomes.

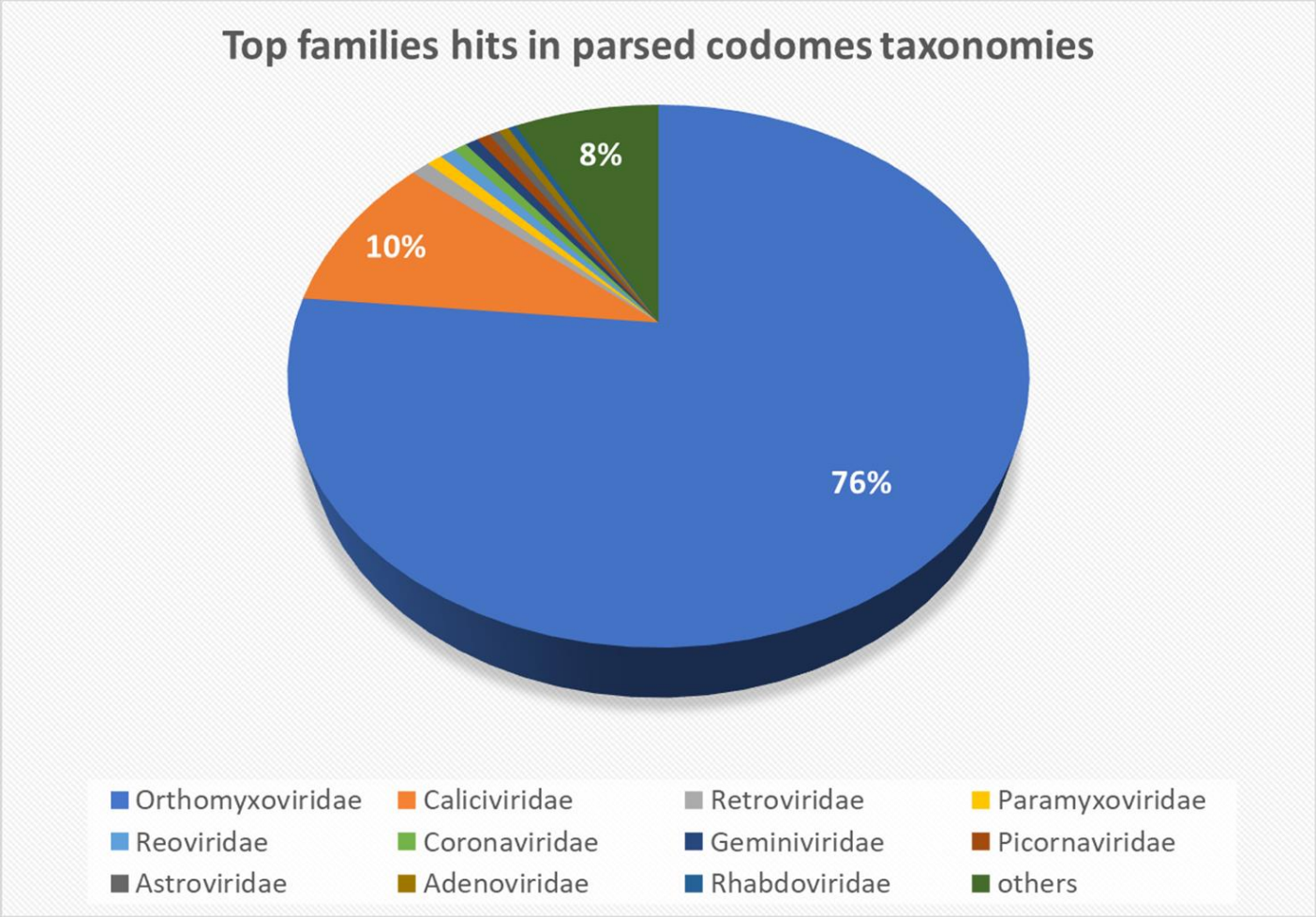

Figure 23: Pie chart showing the family-level taxonomic distribution of all parsed codomes. The chart illustrates the relative proportions of viral families identified after parsing codomes from the GenBank record. Orthomyxoviridae represents the largest proportion (76%), followed by Caliciviridae (10%) and followed by other families as Retroviridae, Paramyxoviridae, Reoviridae, Coronaviridae, Geminiviridae, Picornaviridae, Astroviridae, Adenoviridae, Rhabdoviridae. See supplementary Table S1.

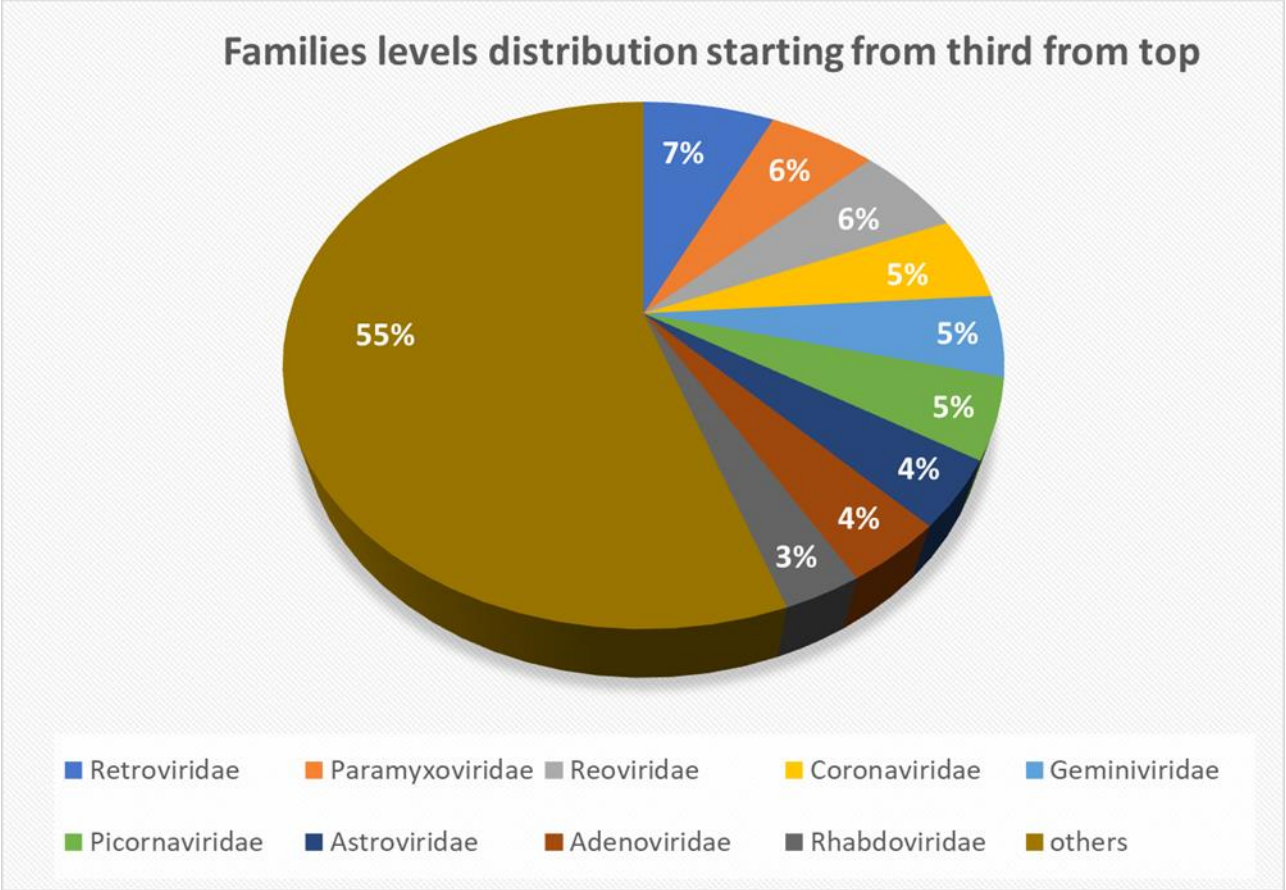

Figure 24: Pie chart showing the family-level taxonomic distribution of parsed codomes after exclusion of the two most abundant families from the dataset. This chart represents the proportional contributions of the remaining viral families, revealing that "others" (a pooled group of low-frequency families) account for the largest proportion (55%). Retroviridae (7%), Paramyxoviridae (6%), Reoviridae (6%), Coronaviridae (5%), Geminiviridae (5%), and Picornaviridae (5%) follow, with smaller contributions from Astroviridae (4%), Adenoviridae (4%), and Rhabdoviridae (3%).

3.1.4 Taxonomic distribution for alignment datasets

Following Methods mentioned previously, 393 alignment data sets passed all stringent criteria to study tempo and mode. Those alignments showed 143 unique taxonomic hierarchy distribution and following similar method in family level calculation, number of 54 families were identified. In those alignment datasets the top end started by 'Reoviridae' with 63 hits, followed by 'Geminiviridae' with 56 hits. While on the other hand 16 families showed only one hit e.g., 'Matonaviridae' and 'Astroviridae'. Table 5 specifies number of species with unique taxonomy hierarchy numbers and family level hits for these taxonomies. Figure 25 is a pie chart for family level taxonomic distribution within 393 viral alignment data sets.

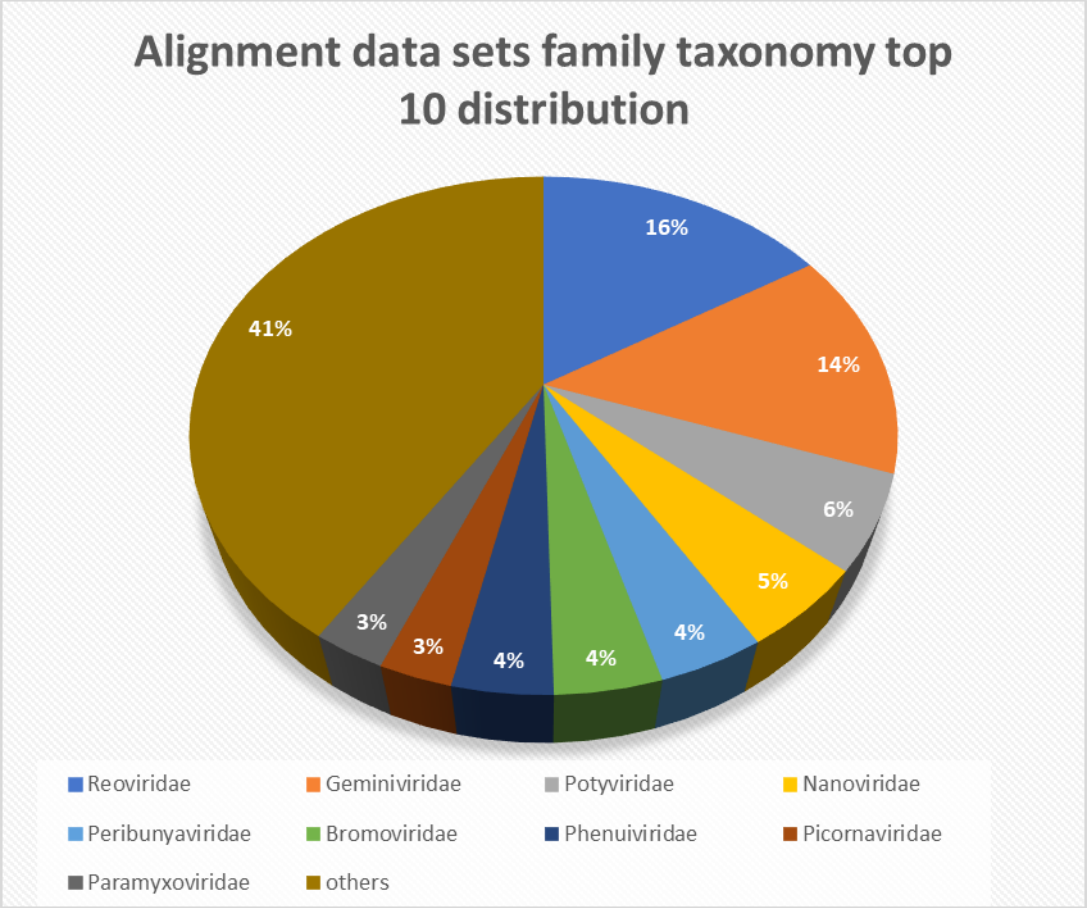

Figure 25: Pie chart showing the family-level taxonomic distribution of 393 alignment datasets, limited to the top 10 most represented viral families. Reoviridae (16%) followed by Geminiviridae (14%). Potyviridae represented 6% of the datasets, Nanoviridae 5%, Peribunyaviridae 4%, Bromoviridae 4%, and Phenuiviridae 4%. Picornaviridae and Paramyxoviridae were the least frequent within the top 10, each representing 3% of the total datasets. Percentages are based on the total number of alignments in the dataset. See Supplementary Table S2 for full data.

Table 5: Represents number of unique taxonomies for parsed codomes and their families.

	Species – Data sets	**Number of Unique Taxonomies**	**Number of Families**
Parsed codomes	156,482	1502	134
Alignment datasets	393	143	54

3.1.5 Flowchart of filtering process

Referring to Methods section, parsed codomes underwent several preparatory processes to facilitate the study of tempo and mode. As mentioned in Table 4, a total of 3,675 distinct species were parsed from GenBank viral reference genomes. Subsequently, the number of alignments was reduced to 393 datasets prior the evolutionary analysis phase. Figure 26 specifies the filtering process steps, and the number of distinct species decrease from 3,675 to 393.

3.1.6 Overview of GenBank parsing section

Section 3.1 represents results and outcome of GenBank records parsing and filtering through multistep pipeline. Initially, a large number of viral genomes were retrieved, from which only those with CDS features and collection dates were saved, out of 4,0402,060 sequences downloaded, 2,649,652 met the initial parsing criteria which represent 156,482 distinct species. Followed by reference genomes parsing for only CDS with a total number of 50,713 sequences which represents 3,675 distinct species. Taxonomic distribution of parsed records showed higher hits for families as Orthomyxoviridae, Caliciviridae and Retroviridae. After applying thresholds including sequence length ratios, identity scores, number of sequences and dates gap, 393 alignment datasets were saved, including number of diverse families as Reoviridae, Geminiviridae, and Potyviridae. A flowchart at figure 26 summarizes the reduction from 3675 species with reference genomes to the final 393 alignments that met the filtering criteria and ready for tempo and mode analyses.

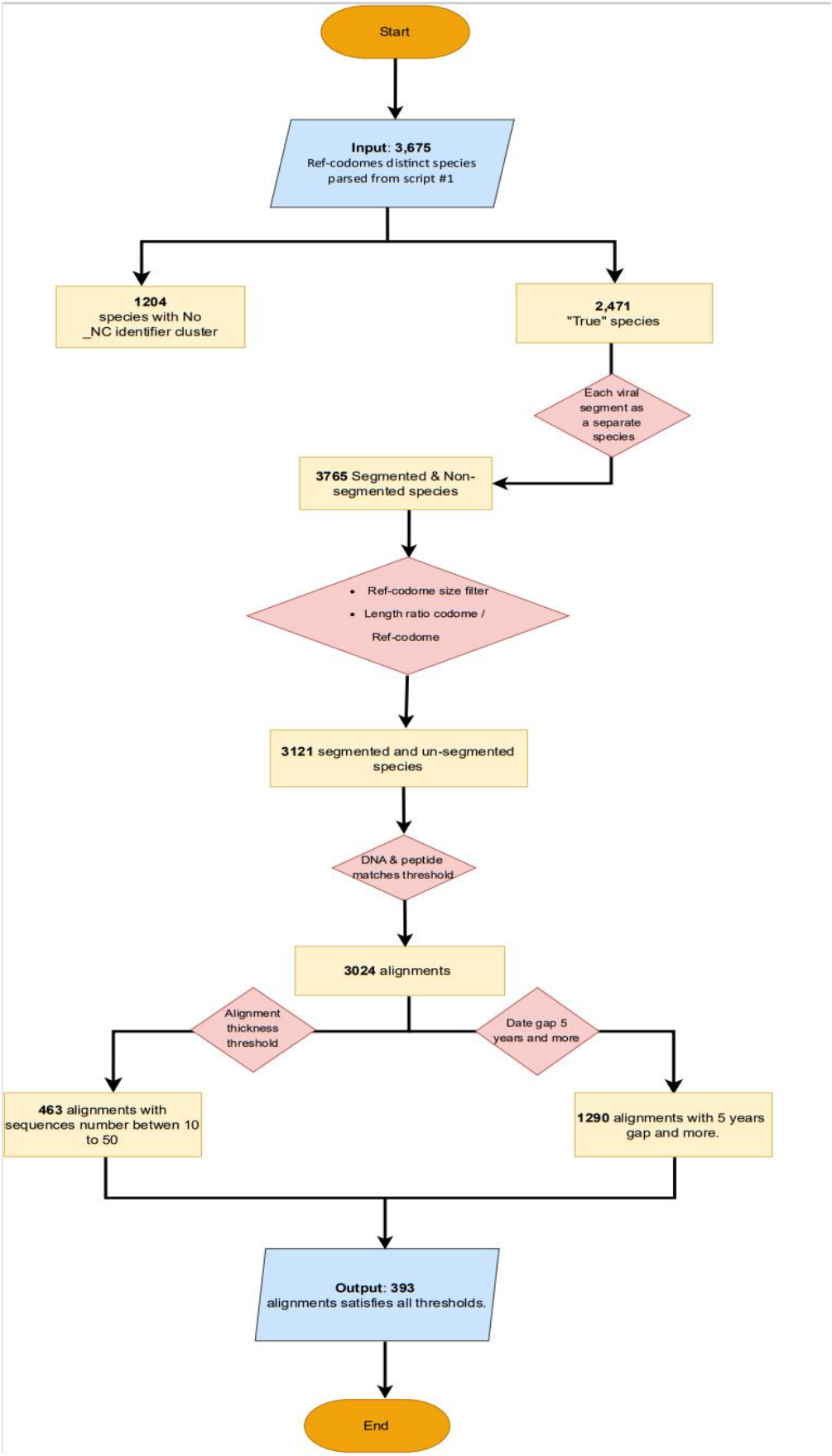

Figure 26: Flowchart illustrates the gradual decrease in the number of distinct species as filters were applied.

3.2 Recombination

3.2.1 Recombination analysis

- Recombinant

  210 alignments showed recombinant sequences, 127 could not be modified for; high number of recombinant signals or short alignment length where window size could not be increased, these 127 alignments were removed from substitution rate analysis. While 83 alignments were modified and saved by removing some recombinant sequences or increasing window size allowing for a broader examination of sequence similarity and evolutionary patterns.

- Non-Recombinant

  183 alignments showed no obvious signal of recombination in Simplot and proceeded to substitution rate analysis. Table 6 summarises number of alignments passed and failed Bootscan before and after modification.

Table 6: Number of alignments pass or fail Bootscan for recombination before removal of recombinant sequences and after modification and removal of recombinant sequences from alignment datasets.

	PASS Bootscan	**FAIL Bootscan**
1st Attempt	183	210
After modification	266	127

3.2.2 Taxonomic distribution for recombinant and after modification

Later, when recombination analysis was complete, taxonomy hierarchy for alignment datasets were divided according to the presence of recombination as follows:

- Recombinant

In the set of sequence alignments containing recombinant sequences, there were a total of 78 unique taxonomies classifications observed. Among these, 37 distinct families were identified. The family 'Potyviridae' was the most frequently encountered, with 19 occurrences, followed closely by 'Geminiviridae' with 18 occurrences. In contrast, 'Caulimoviridae' was the least represented, with only a single occurrence. Figure 27 is a pie chart for family level taxonomy distribution among recombinant viruses.

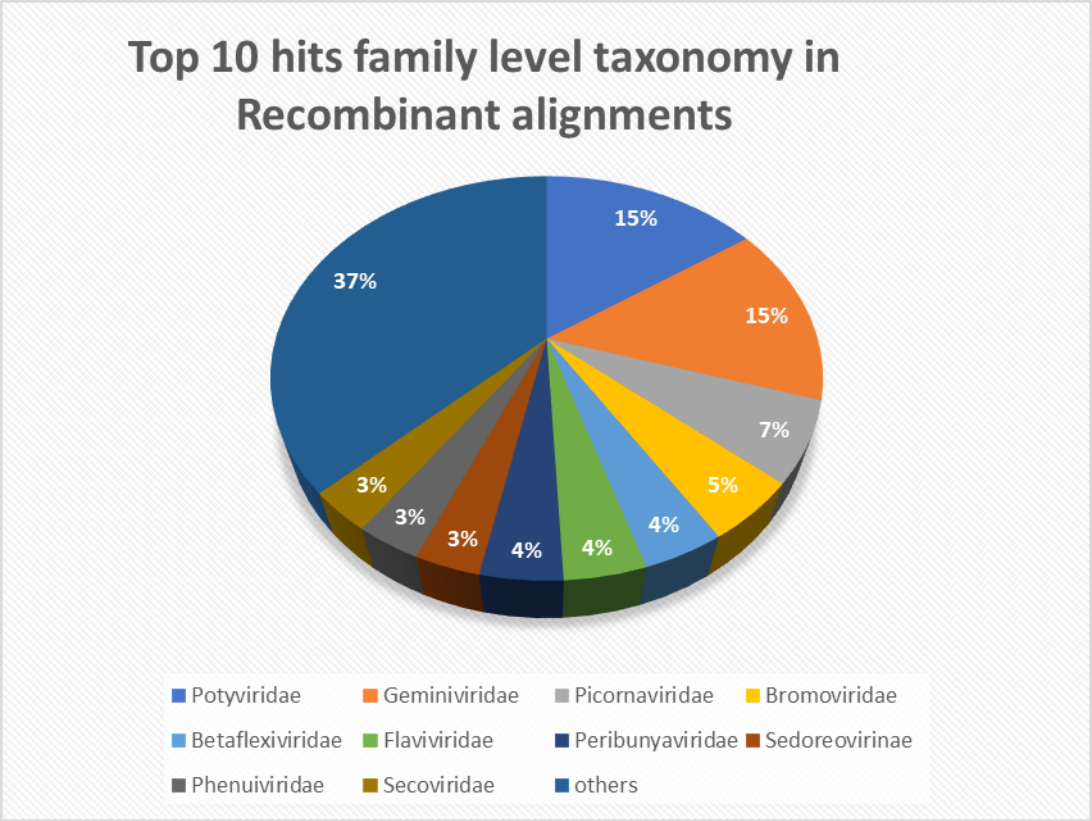

Figure 27: Family level taxonomic distribution for recombinant alignments. In this pie chart Potyviridae and Geminiviridae has the top hits among families for alignments showing recombinant sequences. See supplementary Table S3.

- Non-recombinant

  In recombination free alignments and modified data sets, a total of 96 distinct taxonomic classifications were identified. Among these, there were 49 different families. The family 'Reoviridae' had the highest representation, with 57 occurrences, followed by 'Geminiviridae' with 38 occurrences. In contrast, 'Endornaviridae' and 'Kitaviridae' were the least prevalent, each appearing only

once. Figure 28 is a pie chart for family level taxonomy distribution among non-recombinant viruses.

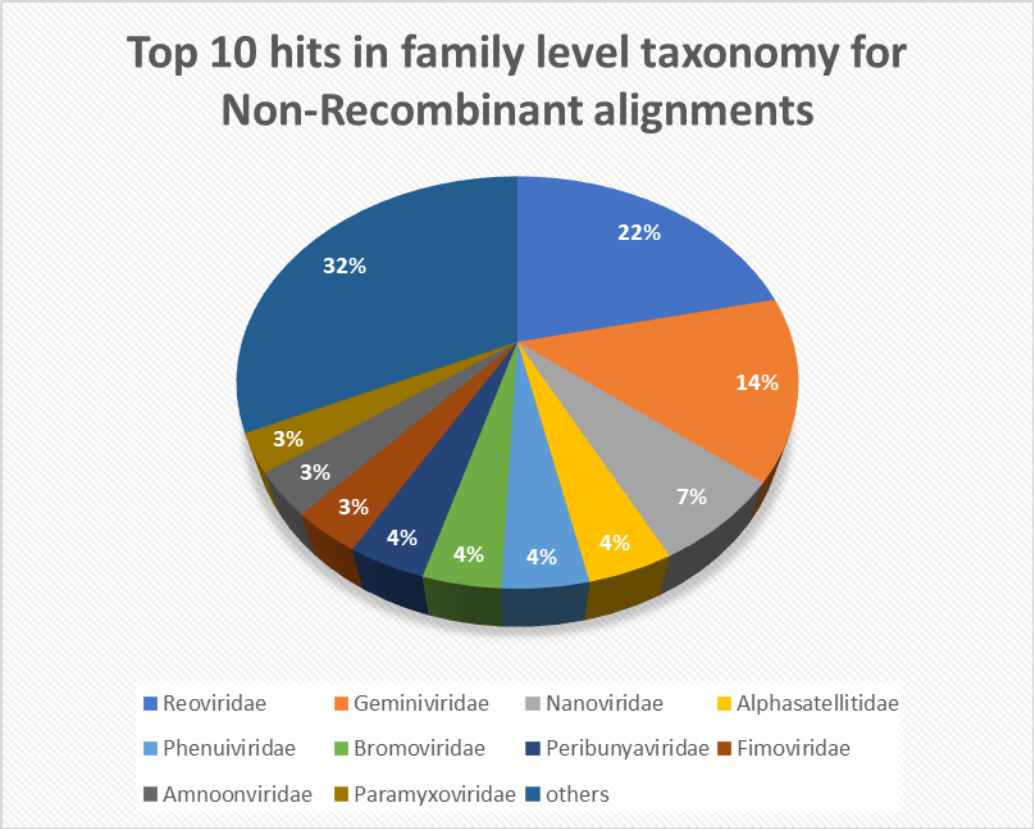

Figure 28: Family level taxonomic distribution for non-recombinant alignments. On this pie chart the top hit family that did not show recombination was Reoviridae, followed by Geminiviridae. See supplementary Table S4.

3.2.3 Recombination in segmented viruses

Among the 393 alignments examined, 199 alignments appeared as a single segment (either it's a non-segmented virus or only one segment passed the stringent filtering criteria), and the remaining 194 alignments were representing 61 segmented viruses' species. During the recombination studies, specific segments of these viral species showed different behaviour in comparison to others. Some examples are detailed in Table 7.

Table 7: Illustrates three segmented viruses recombination behaviour. A. “Faba bean necrotic yellow virus” which has 6 segments where two of them showed recombination that cannot be modified. B. “Peanut stunt virus” which has two segments, and both has recombinant sequences that cannot be modified. C. “Human rotavirus B” which has 10 segments, and all are recombination free by removal some sequences in two out of 10 segments datasets.

Species Segment	**Sequences Number**	**Length of Alignment**	**Recombination/ Comments**
A. Faba bean necrotic yellows virus NC_003559.1	22	500	Yes/ could not modify
Faba bean necrotic yellows virus NC_003560.1	22	850	No
Faba bean necrotic yellows virus NC_003561.1	18	460	No
Faba bean necrotic yellows virus NC_003562.1	39	340	No
Faba bean necrotic yellows virus NC_003563.1	28	500	No
Faba bean necrotic yellows virus NC_003566.1	24	460	Yes/could not modify
B. Peanut stunt virus NC_002038.1	12	3000	Yes/ could not modify
Peanut stunt virus NC_002039.1	12	2800	Yes/ could not modify
C. Human rotavirus B NC_021541.1	17	3480	No
Human rotavirus B NC_021542.1	39	700	No
Human rotavirus B NC_021543.1	23	2250	Yes/modified by removing two sequences
Human rotavirus B NC_021544.1	24	1150	No
Human rotavirus B NC_021545.1	17	2800	No
Human rotavirus B NC_021547.1	17	1000	No
Human rotavirus B NC_021548.1	19	900	No
Human rotavirus B NC_021549.1	17	500	No
Human rotavirus B NC_021550.1	24	640	No
Human rotavirus B NC_021551.1	17	2200	Yes/modified by removing two sequences

3.2.4 Recombination patterns

- Segmented viruses

In the 61 segmented viruses, each virus exhibits a variable number of segments, ranging from 2 to 10 segments per virus which represents number of alignment data sets. When examining the presence of recombination events across these alignments, it was observed that 35 viruses displayed concordant patterns. The definition of concordance is either all segments within a virus exhibited recombination signals, or all segments were lacking any detectable recombination events.

However, in the remaining 26 viruses, we observed discordant patterns. In these, some segments of the viruses displayed clear recombination signals, while other segments showed no evidence of recombination. Table 8 lists recombination patterns concordance and discordance within segmented viruses in detailed numbers.

Table 8: Difference in recombination pattern among 61 segmented viruses, where each one includes 2 to 10 segments alignments dataset, the second column from the left displays number of viruses where all segments show recombinant signals, the third column from the left presents number of viruses where all segments did not show recombination in their alignments sets sequences, last column displays number of viruses where segments showed different recombination behaviour. See supplementary Tables S5 and S6.

Number of Segmented Virus Species	Concordant Pattern Recombination	Concordant Pattern Non-Recombination	Discordant Pattern
61	6	29	26

3.2.5 Multiple taxonomic hierarchy level

While investigating taxonomy hierarchy within viral datasets during the recombination analysis, variations in concordance and discordance were observed across multiple hierarchical levels. To maintain consistency in analysis, an additional specific criterion was added when counting taxa at the order, family, and genus levels, as outlined below:

Order: any order level considered as valid if it was present in the taxonomy list with multiple associated family levels, two and more. For example, 'Geplafuvirales' appeared in 56 species but with only one family 'Geminiviridae', so it was not included in the analysis. While 'Hepelivirales' was included in the concordance-discordance analysis since it appeared with different families 'Alphatetraviridae', 'Benyviridae', 'Hepeviridae' and 'Matonaviridae'.

Family: A family was counted if it appeared with more than one associated genus hit. Here 'Geminiviridae' was counted in the family level since it includes 8 different genera: 'Becurtovirus', 'Begomovirus', 'Capulavirus', 'Citlodavirus', 'Maldovirus', 'Mastrevirus', 'Opunvirus' and 'Turncurtovirus'.

Genus: A genus also was counted if it was associated with more than one species hit, for example 'Mammarenavirus' appeared in 5 hits with 3 different species datasets;

Lymphocytic choriomeningitis mammarenavirus, Guanarito mammarenavirus and Machupo mammarenavirus. Table 9 listing order, family, and genus levels with the number of every pattern appears for each taxonomical level.

Table 9: Number of different taxonomic hierarchy levels and their concordance-discordance pattern of recombination. Second column from the left shows number of hierarchy level included in the analysis. Third column from the left shows number of concordant patterns with recombination in all entries for three levels; order, family and genus. Fourth column from the left shows number of concordant patterns with no recombination in all entries for each level. The last column shows number of discordant patterns.

	Number	**Concordant Pattern Recombinant**	**Concordant Pattern Non-Recombinant**	**Discordant**
Order	9	0	0	9
Family	29	3	3	22
Genus	36	5	9	22

3.2.6 Host association in concordant species

Viral species for the concordant genera from Table 9 were identified and linked to their hosts. Table 10A is listing the 14 genera with their species studied in the analysis, among 43 species examined 4 belonged to mammalian hosts, 4 were associated with insects, one with fish and the remaining were plant hosts. Table 10B lists concordant families with their corresponding genera, species, and relevant hosts.

Table 10: A. Lists concordant genera with their associated species, and the respective host for each species. For discordant genera, see supplementary Table S7. B. Lists concordant families with their associated genera, species, and the respective host for each species.

Genus	Species	Host
Colecusatellite (concordant non-Recombinant)	Melon chlorotic mosaic alphasatellite	Plant
	Chilli leaf curl alphasatellite	Plant
	Gossypium darwinii symptomless alphasatellite	Plant
	Ageratum enation alphasatellite	Plant
	Ageratum yellow vein India alphasatellite	Plant
	Tomato leaf curl alphasatellite	Plant
unclassified Begomovirus-associated alphasatellites (Concordant Non-Recombinant)	Nanovirus-like particle	Plant
	Ageratum conyzoides symptomless alphasatellite	Plant
	Guar leaf curl alphasatellite	Plant
Mastrevirus (concordant non-Recombinant)	Chickpea chlorosis Australia virus	Plant
	Panicum streak virus	Plant
	Paspalum striate mosaic virus	Plant
	Sweet potato symptomless virus 1	Plant
Orthoreovirus (concordant non-Recombinant)	Mammalian orthoreovirus 3	Mammalian
	Piscine orthoreovirus	fish
Tobamovirus (concordant non-Recombinant)	Pepper mild mottle virus	Plant
	Tomato mosaic virus	Plant
	Tomato mottle mosaic virus	Plant
Potexvirus (concordant Recombinant)	Bamboo mosaic virus	Plant
	Citrus yellow vein clearing virus	Plant
Carlavirus (concordant non-Recombinant)	Garlic common latent virus	Plant
	Potato virus M	Plant
Foveavirus (concordant Recombinant)	Apple stem pitting virus	Plant
	Grapevine rupestris stem pitting-associated virus	Plant
Trichovirus (concordant Recombinant)	Apple chlorotic leaf spot virus	Plant
	Grapevine Pinot gris virus	Plant
Mammarenavirus (concordant non-Recombinant)	Guanarito mammarenavirus	Mammalian
	Lymphocytic choriomeningitis mammarenavirus	Mammalian
	Machupo mammarenavirus	Mammalian
Iflavirus (concordant Recombinant)	Deformed wing virus	Insect
	Sacbrood virus	Insect
Ipomovirus (concordant Recombinant)	Cassava brown streak virus	Plant
	Cucumber vein yellowing virus	Plant
	Ugandan cassava brown streak virus	Plant
Negevirus (concordant non-Recombinant)	Piura virus	Insect
	Wallerfield virus	Insect
Betasatellite (concordant non-Recombinant)	Cotton leaf curl betasatellite	Plant
	Papaya leaf curl betasatellite	Plant
	Ageratum yellow leaf curl betasatellite	Plant
	Cotton leaf curl Gezira betasatellite	Plant
	Cotton leaf curl virus betasatellite	Plant
	Croton yellow vein mosaic betasatellite	Plant
	Cotton leaf curl Burewala betasatellite	Plant

Table10 B

Family	Genus	Species	Host
Polyomaviridae (concordant non-Recombinant)	Alphapolyomavirus	Trichodysplasia spinulosa-associated polyomavirus	Mammalian
	Deltapolyomavirus	Human polyomavirus 6	Mammalian
		Human polyomavirus 7	Mammalian
	Gammapolyomavirus	Goose hemorrhagic polyomavirus	Avian
Virgaviridae (concordant non-Recombinant)	Furovirus	Japanese soil-borne wheat mosaic virus	Plant
	Pomovirus	Potato mop-top virus	Plant
	Tobamovirus	Pepper mild mottle virus	Plant
		Tomato mosaic virus	Plant
		Tomato mottle mosaic virus	Plant
Alphaflexiviridae (concordant Recombinant)	Allexivirus	Garlic virus B	Plant
	Potexvirus	Bamboo mosaic virus	Plant
		Citrus yellow vein clearing virus	Plant
Dicistroviridae (concordant Recombinant)	Aparavirus	Israeli acute paralysis virus	Insect
	Cripavirus	Aphid lethal paralysis virus	Insect
	Triatovirus	Black queen cell virus	Insect
Iflaviridae (concordant Recombinant)	Iflavirus	Deformed wing virus	Insect
		Sacbrood virus	Insect
	Unclassified Iflaviridae	La Jolla virus	Insect
Alphasatellitidae (concordant non-Recombinant)	Colecusatellite	Melon chlorotic mosaic alphasatellite	Plant
		Chilli leaf curl alphasatellite	Plant
		Gossypium darwinii symptomless alphasatellite	Plant
		Ageratum enation alphasatellite	Plant
		Ageratum yellow vein India alphasatellite	Plant
		Tomato leaf curl alphasatellite	Plant
	unclassified Begomovirus-associated alphasatellites	Nanovirus-like particle	Plant
		Ageratum conyzoides symptomless alphasatellite	Plant
		Guar leaf curl alphasatellite	Plant

3.2.7 Overview of recombination analysis results

This section investigates recombination events across 393 high quality viral alignments. 210 alignments showed evidence of recombination based on Simplot analysis, with 127 not suitable for modification, and 183 alignments showed no evidence for recombination. Followed by investigating taxonomic distribution of recombinant and non-recombinant alignments, with Potyviridae family as the top hit in presence of recombinant sequences while Reoviridae the family with the highest frequent occurrence among non-recombinant alignments, and Geminiviridae present in both. Moving to study recombination on segmented viruses, a total of 61 segmented viruses, where recombination was assessed across individual segments showed different behaviours. With 35 viruses displaying concordant patterns across segments, and 26 exhibited discordant segmental signals. When recombination concordance was examined across taxonomic levels, no consistent pattern was observed at order level, and only 6 out of 29 families and 14 out of 36 genera showed concordance. Furthermore, species displaying concordant recombination patterns were more frequently associated with plant hosts.

3.3 Molecular clock estimation

3.3.1 Temporal signal analysis for molecular clock

Data sets were examined for “R” values using TempEst, for each data set root to tip plot on TempEst were examined for the presence of temporal signal. These plots measure the genetic distance of sequences from the root of the tree against the time distance, suggesting the presence or absence of temporal signal, indicating sequences evolving in a clock-like manner over time. The correlation coefficient value in a root-to-tip plot indicates the presence and strength of a temporal signal in a dataset. Figures 29 and 30 are screenshots for root to tip plot from TempEst output with high and low correlation coefficient “R” values.

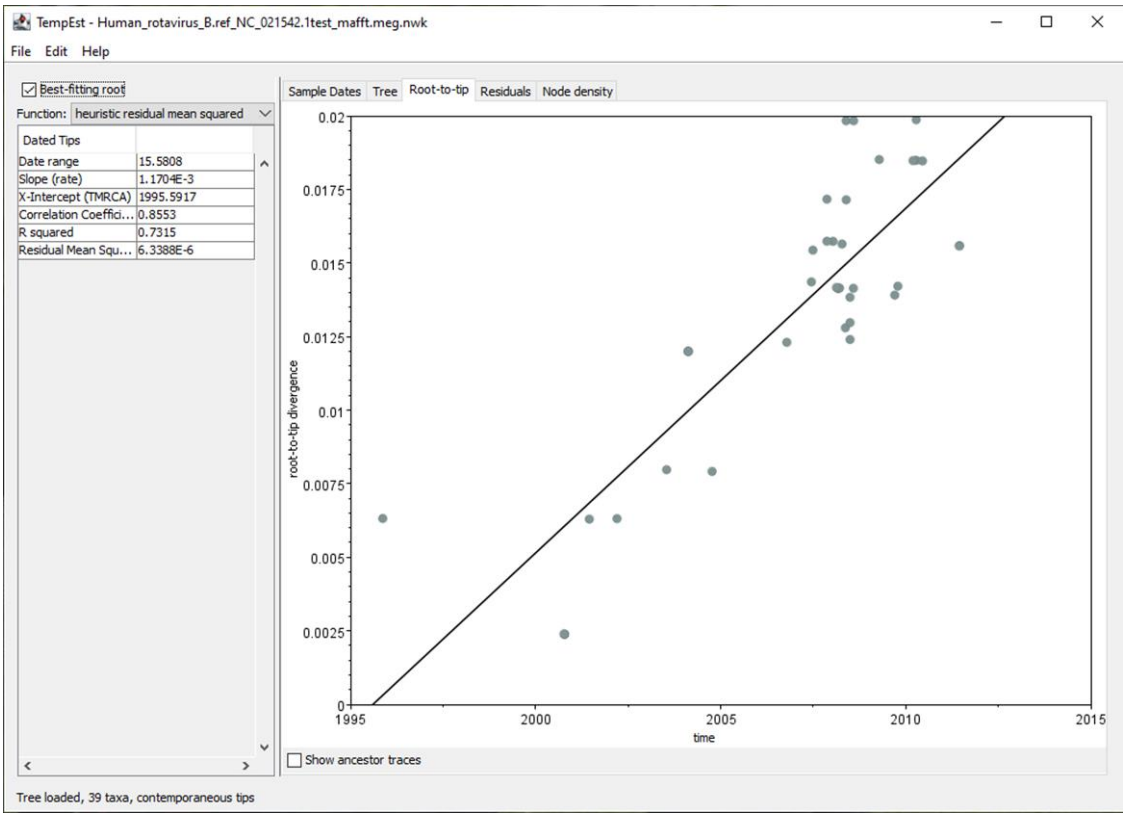

Figure 29: Screenshot of root to tip plot on TempEst for “Human rotavirus B” with high R value.

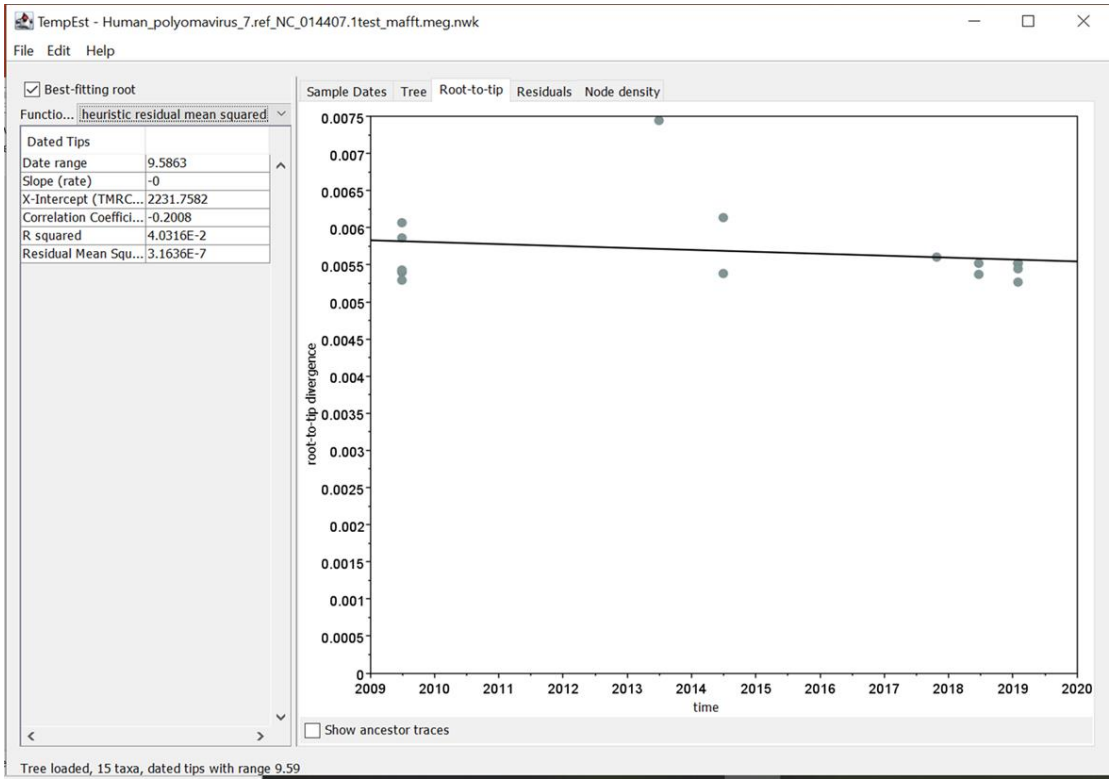

Figure 30: Screenshot of root to tip plot on TempEst for "Human polyomavirus 7" with low R value.

3.3.2 Correlation Coefficient "R" values by TempEst

- Wider range for TempEst data sets

  Additional to the 393 alignments, temporal signal was estimated in a wider range of data. Some filtrating restrictions were eased in the second dataset as follows: number of sequences was expanded from 50 to 200 sequences per data set with no change in years age gap, allowing more 121 datasets were added to the temporal signal analysis.

- Pie chart for "R" values

  514 alignments were examined for temporal signal; 272 alignments have correlation coefficient values "R" more than 0.5 and proceed to substitution rate analysis. Figure 31 is a pie chart for "R" values percentages.

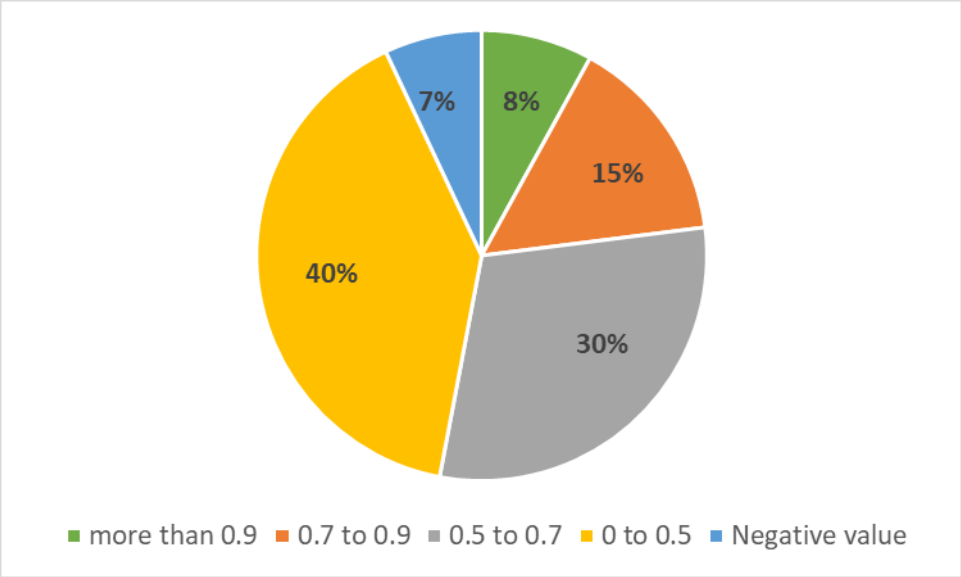

Figure 31: Pie chart showing the distribution of TempEst correlation coefficient (R) values for 514 alignments used in molecular clock analysis. The majority of alignments (40%) had R values between 0 and 0.5, followed by 30% with values between 0.5 and 0.7. Alignments with higher correlation values were less frequent, with 15% between 0.7 and 0.9 and 8% above 0.9, indicating stronger temporal signal. A small proportion (7%) had negative R values, suggesting no meaningful correlation between genetic divergence and sampling time. See supplementary Tables S8 and S9 for highest and lowest R values genera and families.

3.3.3 Taxonomic distribution according to temporal signal

Later, once temporal signal was studied in 514 alignments, taxonomy hierarchy for datasets were divided according to correlation coefficient values as follows:

- Taxonomy for viral datasets with high correlation coefficient values

  In the set of sequence alignments showed evidence of stronger temporal signal with "R" value >= 0.5, there were a total of 122 unique taxonomies observed. Among these, 49 distinct families were identified. The family 'Reoviridae' was the most frequently encountered, with 40 occurrences, followed closely by 'Geminiviridae' with 34 occurrences. In contrast, 'Adenoviridae' was the least represented, with only a single occurrence.

- Taxonomy for viral datasets with low correlation coefficient values

  On the other hand, in the collection of sequence alignments where temporal signal "R" value was <=0.5, a total number of 105 unique taxonomic classifications was observed.

with 42 distinct families, 'Geminiviridae' was the most prevalent, appearing 32 times, closely followed by 'Reoviridae' with 23 occurrences. While the 'Adenoviridae' family was the least frequent, with only a single hit. Figures 32 and 33 are pie charts for family's distribution according to "R" values.

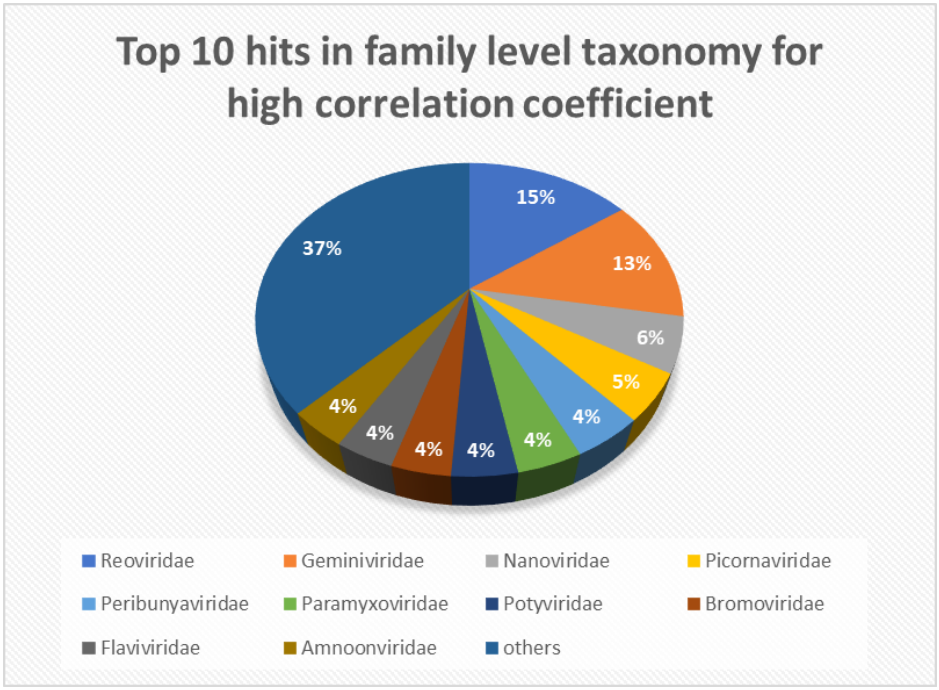

Figure 32: Pie chart showing the family-level taxonomic distribution of datasets with TempEst correlation coefficients (R) ≥ 0.5, indicating moderate to strong temporal signal. See supplementary Table S10.

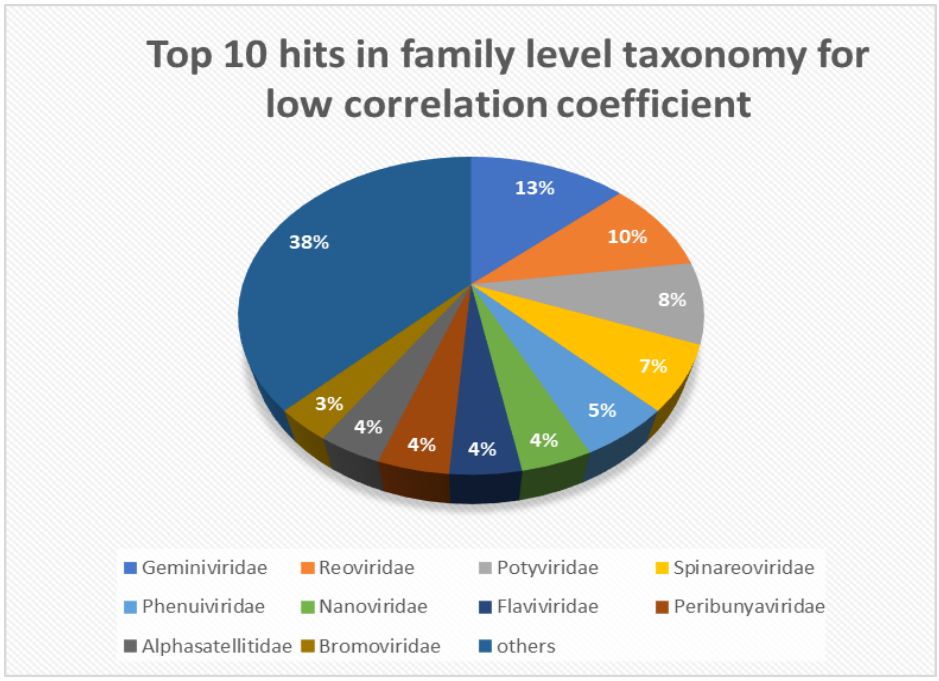

Figure 33: Family level taxonomic distribution for "R" values < 0.5. See supplementary Table S11.

3.3.4 Temporal signal patterns

- Segmented viruses

  Following analysis shown in section 3.2.4, concordance and discordance patterns were examined within datasets of segmented viruses. Among 514 alignments, 78 segmented viruses were represented. When temporal signal measured within each of these segmented datasets, the following patterns appeared:

  a. Concordant high "R" values: among the segmented viruses, 27 exhibited a concordant pattern, where all segments displayed strong temporal signals with high R values exceeding 0.5.
  b. Concordant low "R" values: on the other hand, 20 segmented viruses demonstrated a concordant pattern with low R values less than 0.5.
  c. Discordant patterns: the remaining segmented viruses showed a discordant pattern, showing a mixture of high and low R values within the segments of these viral species.

Table 11: Difference in molecular clock among 79 segmented viruses, where each one includes 2 to 10 segments alignments dataset, the second column from the left displays number of viruses where all segments "R" values are higher or equal to 0.5, the third column from the left presents number of viruses where all segments "R" values are lower than 0.5, last column displays number of viruses where segments showed high and low "R" values. See supplementary Tables S12 to S14.

Number of Segmented Viruses	**Concordant Pattern "R" >= 0.5**	**Concordant Pattern "R" < 0.5**	**Discordant Pattern**
79	27	20	32

- Multiple taxonomic hierarchy level

  Also, taxonomical levels patterns were studied for concordance and discordance in molecular clock analysis. Considerations in order, family and genus levels counting continues similar to section 3.2.5. Table 12 lists order, family and genus levels with patterns appear for each taxonomical level.

Table 12: Number of different taxonomic hierarchy levels and their concordance-discordance pattern of molecular clock. Second column from the left shows number of hierarchy level included in the analysis. Third column from the left shows number of concordant patterns with R value >= 0.5 in all entries for three levels; order, family and genus. Fourth column from the left shows number of concordant patterns with R value < 0.5 in all entries for each level. The last column shows number of discordant patterns.

	Number	**Concordant pattern "R" >= 0.5**	**Concordant pattern "R" < 0.5**	**Discordant Pattern**
Order	10	0	0	10
Family	29	4	1	24
Genus	49	6	10	33

3.3.5 Host association in concordant species

Studying host association continues as previously done in section 3.2.6. For temporal signal analysis, viral species for the 16 concordant genera and 5 concordant families were identified and linked to their hosts. Table 13A is listing concordant genera with their corresponding species, among 37 species examined, 2 were found with avian hosts, 2 with insect hosts, 13 with fish hosts and the remaining 20 were mammalian hosts. Table 13B lists concordant families with their genera, species, and relevant hosts.

Table 13: A. Lists concordant genera with their associated species, and the respective host for each species. For discordant genera, see supplementary Table S15. B. Lists concordant families with their associated genera, species, and the respective host for each species.

Genus	Species	Host
Gyrovirus (concordant low "R")	Chicken anemia virus	Avian
	Avian gyrovirus 2	Avian
Betapolyomavirus (concordant low "R")	Human polyomavirus 1	Mammalian
	WU Polyomavirus	Mammalian
	JC polyomavirus	Mammalian
Babuvirus (concordant low "R")	Cardamom bushy dwarf virus	Plant
	Banana bunchy top virus	Plant
Foveavirus (concordant low "R")	Apple stem pitting virus	Plant
	Grapevine rupestris stem pitting-associated virus	Plant
Capillovirus (concordant low "R")	Apple stem grooving virus	Plant
	Cherry virus A	Plant
Trichovirus (concordant low "R")	Grapevine_pinot_gris_virus	Plant
	Apple chlorotic leaf spot virus	Plant
Respirovirus (concordant high "R")	Porcine respirovirus 1	Mammalian
	Human respirovirus 1	Mammalian
Orthorubulavirus (concordant high "R")	Mumps orthorubulavirus	Mammalian
	Human orthorubulavirus 2	Mammalian
	Mammalian orthorubulavirus 5	Mammalian
Lyssavirus (concordant high "R")	Australian bat lyssavirus	Mammalian
	European bat 1 lyssavirus	Mammalian
Vesiculovirus (concordant high "R")	Chandipura virus	Mammalian
	Vesicular stomatitis New Jersey virus	Mammalian
	Vesicular stomatitis Indiana virus	Mammalian
Orthonairovirus (concordant low "R")	Kasokero virus	Mammalian
	Crimean-Congo hemorrhagic fever orthonairovirus	Mammalian
Orthotospovirus (concordant low "R")	Groundnut ringspot virus	Plant
	Capsicum chlorosis virus	Plant
Iflavirus (concordant low "R")	Sacbrood virus	Insect
	Deformed wing virus	Insect
Kobuvirus (concordant high "R")	Aichi virus 1	Mammalian
	Porcine kobuvirus	Mammalian
	Canine kobuvirus	Mammalian
Parechovirus (concordant high "R")	Parechovirus A	Mammalian
	Human parechovirus 1	Mammalian
Ipomovirus (concordant low "R")	Cucumber vein yellowing virus	Plant
	Ugandan cassava brown streak virus	Plant
	Cassava brown streak virus	Plant

Table13B

Family	Genus	Species	Host
Tombusviridae (concordant high "R")	Umbravirus	Pea enation mosaic virus 2	Plant
	Luteovirus	Barley yellow dwarf virus	Plant
	Gammacarmovirus	Soybean yellow mottle mosaic virus	Plant
Filoviridae (concordant high "R")	Ebola virus	Sudan ebolavirus	Mammalian
	Orthomarburgvirus	Marburg marburgvirus	Mammalian
Rhabdoviridae (concordant high "R")	Lyssavirus	Australian bat lyssavirus	Mammalian
		European bat 1 lyssavirus	Mammalian
	Sprivivirus	Carp sprivivirus	Fish
	Vesiculovirus	Chandipura virus	Mammalian
		Vesicular stomatitis New Jersey virus	Mammalian
		Vesicular stomatitis Indiana virus	Mammalian
	Novirhabdovirus	Infectious hematopoietic necrosis virus	Fish
Solemoviridae (concordant high "R")	Enamovirus	Citrus vein enation virus	Plant
	Sobemovirus	Rice yellow mottle virus	Plant
Dicistroviridae (concordant low "R")	Aparavirus	Israeli acute paralysis virus	Insect
	Cripavirus	Aphid lethal paralysis virus	Insect
	Triatovirus	Black queen cell virus	Insect

3.3.6 Overview of molecular clock results

Temporal signal analysis was performed using TempEst to evaluate the correlation coefficient "R" values as a measure of molecular clock strength. Among 514 alignments assessed, 272 datasets with R >=0.5 proceeded for substitution rate estimation, indicating sufficient temporal signal. Taxonomic distribution analysis showed that both high and low correlation alignments were taxonomically diverse, with Reoviridae and Geminiviridae frequently presented across both categories. Moving to segmented viruses, 78 distinct segmented species were identified, of which 27 showed complete concordance with high R values across all segments, 20 showed concordances with low R values, and 32 exhibited discordant segmental patterns. Further analysis across multiple taxonomic levels showed limited concordance in temporal signal, as only 5 out of 29 families and 16 out of 49 genera showed concordance in R values across datasets. While majority displayed discordant patterns. Lastly, host analysis indicated that concordant families and genera were distributed across both plant and mammalian viruses.

3.4 Substitution rate analysis

3.4.1 BEAST estimation for substitution rates

Easing filtration restriction continues with a second attempt, additional to the 393-alignments in the first batch, the data set was expanded by relaxing the stringent filtration criteria, as follows: a second batch was added with pairwise aligned data sets containing number of sequences from 50 to 80, also a year age gap of 4 years. Selecting alignments with R => 0.5 from TempEst output values, BEAST run on total 350 alignment sets.

Once XML files were processed by BEAUti and followed by BEAST software run, output "log" files are later examined by Tracer software, for each alignment data set, number of traces with summary statistics were generated, below are traces highlighted for Bayesian analysis output:

- MeanRate values

  In this study, one of the parameters analysed in the output file is the meanRate. The meanRate, expressed as substitutions per site per year, quantifies the average rate at which genetic changes occur in the dataset. The analysis yielded meanRate values from 7.52×10^{-6} to 0.0991 substitutions/site/year (s/s/y). According to this range, alignments were divided to 5 categories for evolution speed, each category corresponding approximately to an order of magnitude. Table 14 lists the 5 categories and number of alignments for each speed.

Table 14: Lists the five categories separating viral alignments data sets according to meanRate values. See supplementary Tables S16 and S17.

	MeanRate Range (s/s/y)	**Number of Alignments**
Very Slow	7.52×10^{-6} to 9.99×10^{-6}	2
Slow	1.91×10^{-5} to 9.93×10^{-5}	31
Moderate	1.05×10^{-4} to 9.91×10^{-4}	150
Fast	1.00×10^{-3} to 9.70×10^{-3}	145
Very Fast	1.00×10^{-2} to 9.91×10^{-2}	22

3.4.2 Coefficient of variation values

The second parameter analysed over tracer is coefficient of variation. Coefficient of variation can be defined as the measure of variation in evolution rates among different lineages or branches in phylogenetic trees, it measures evolutionary rates variation according to meanRate. The coefficient of variation is calculated as the standard deviation of rates divided by the mean rate.

Values of coefficient of variation out of Bayesian analysis run on alignment datasets ranged from 0.052 to 7.3033, with value range lower bound ranges from 1.54×10^{-7} to 1.648.

The threshold for the coefficient of variation was determined based on the clock model suggested by Drummond; BEAST runs using relaxed clock models, the coefficient of variation indicates the clock-like nature of the input data. Values 0 to 0.1 indicate a strong clock-like behaviour, and a strict clock model may be more suitable. Conversely, if the coefficient of variation is high (e.g., greater than 0.1), it suggests a larger standard deviation, requiring the use of a relaxed molecular clock model. If the coefficient of variation exceeds 1, the data exhibit significant non-clock-like behaviour and are generally unsuitable for estimating divergence times (Drummond and Bouckaert, 2015). Table 15 divided data sets according to coefficient of variation values, and Table 16 dividing each category in Table 15 according to their coefficient of variation values. Moreover, family level distribution were studies according to coefficient of variation values as displayed in Figures 34 & 35 for values 0 to 1 and >=1.

Table 15: Number of alignments for all coefficient of variation value range. See supplementary Tables S18 and 19.

Coefficient of Variation	**0 to 0.1**	**0.1 to 1**	**>=1**
Number of Alignments	9	185	156

Table 16: Number of alignments for each coefficient of variation range according to meanRate categories.

Evolution Speed	MeanRate Range (s/s/y)	Number of Alignments	Coefficient of Variation		
			0 to 0.1	0.1 to 1	>=1
Very slow	7.52×10^{-6} to 9.99×10^{-6}	2	0	2	0
Slow	1.91×10^{-5} to 9.93×10^{-5}	31	3	18	10
Moderate	1.05×10^{-4} to 9.91×10^{-4}	150	3	88	59
Fast	1.00×10^{-3} to 9.70×10^{-3}	145	2	69	74
Very fast	1.00×10^{-2} to 9.91×10^{-2}	22	1	7	14

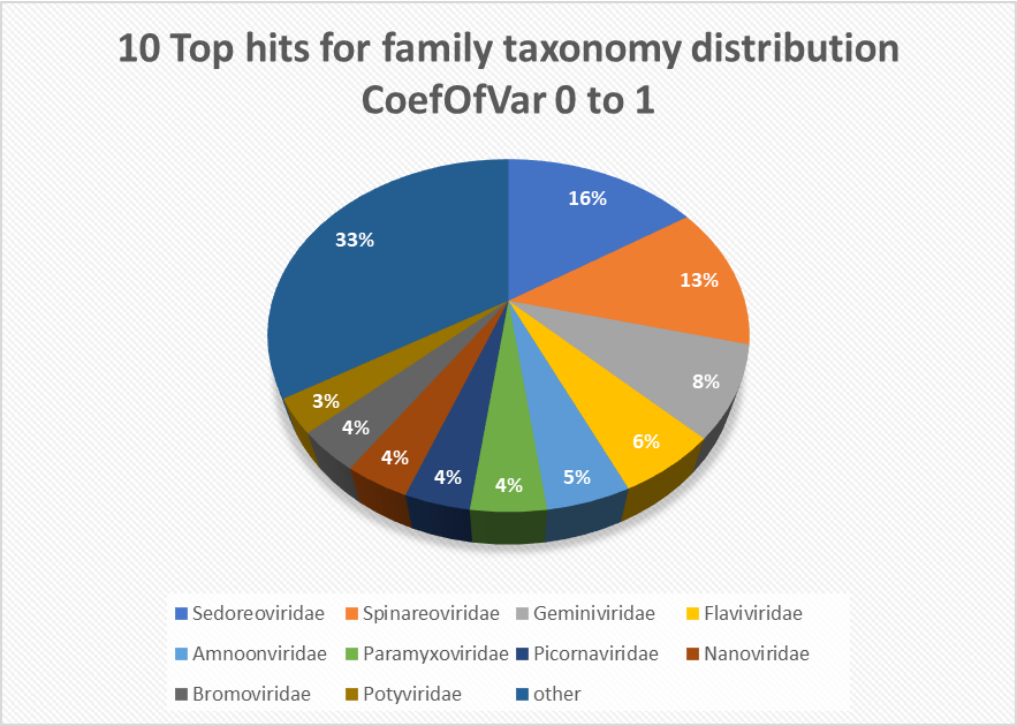

Figure 34: Pie chart showing the family-level taxonomic distribution for datasets with coefficient of variation between 0 and 1, see supplementary Table S20.

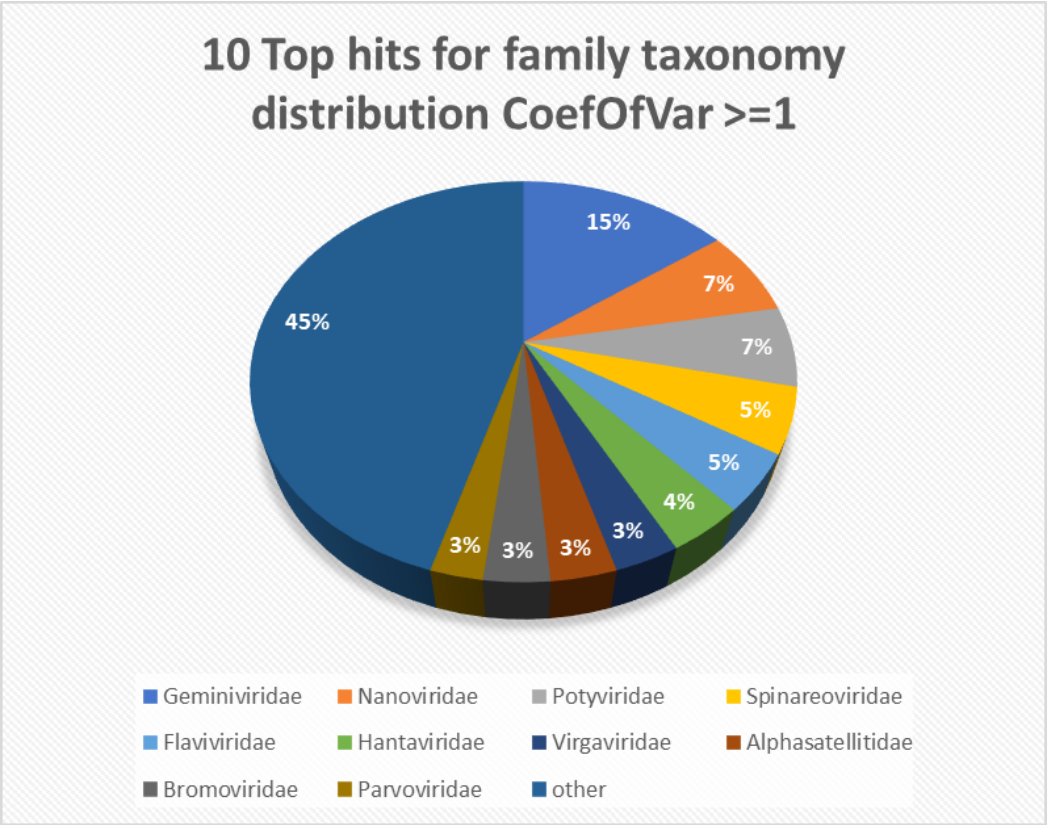

Figure 35: Pie chart showing the family-level taxonomic distribution for datasets with coefficient of variation between equal or more than 1, see supplementary Table S21.

To validate that high coefficient of variation values >=1 was not produced by artifacts, coefficient of variation values was plotted against alignment length. Figure 36 represents the scatter plot, which revealed the absence of a linear relationship between these variables. Moreover, the absence of a significant correlation in the data indicated that not all short alignments exhibited high coefficient of variation values.

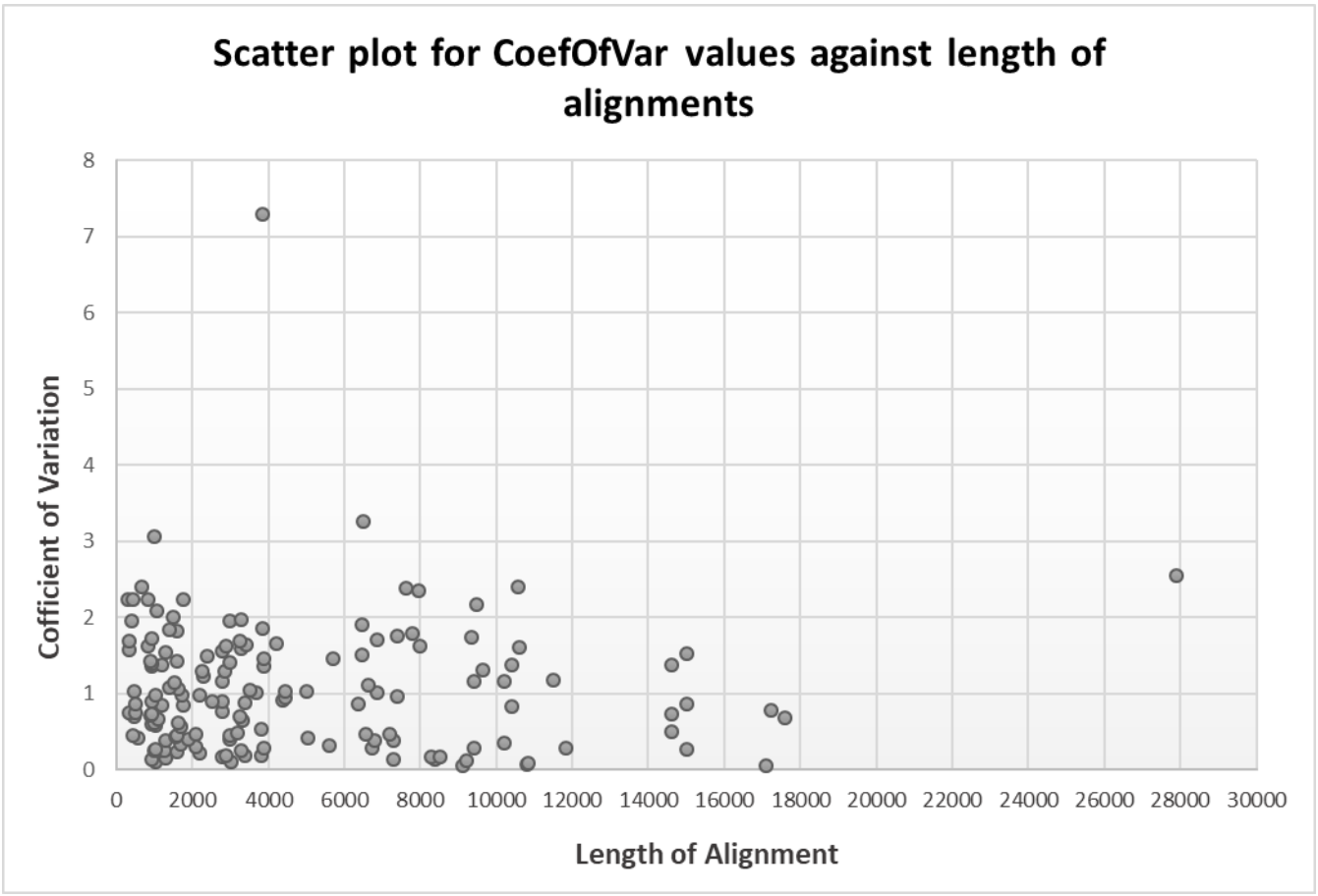

Figure 36: Scatter plot showing the relationship between coefficient of variation values from BEAST output and the length of viral sequence alignments. Each point represents one dataset. No strong linear correlation is observed, indicating that variability in evolutionary rate estimates is not directly dependent on alignment length.

3.4.3 Examples of BEAST analysis parameters

Figures 37 to 40 provide illustrative examples of BEAST analysis output as viewed in Tracer software. The Tracer interface provides an overview of the analysis results, with distinct components:

- Left Panel - Traces: Traces appear, visually represent the evolution of model parameters over time. These traces are essential for assessing parameter convergence and behaviour.

- Top Right Panel - Summary Statistics: A summary of important statistics is presented for the currently highlighted trace. This includes key statistical measures that offer insights into the parameter estimates and their uncertainty.

- Bottom Panel - Frequency Plot: At the bottom, there is frequency plot corresponding to the selected trace. The frequency plot provides a graphical representation of parameter values and their distribution.

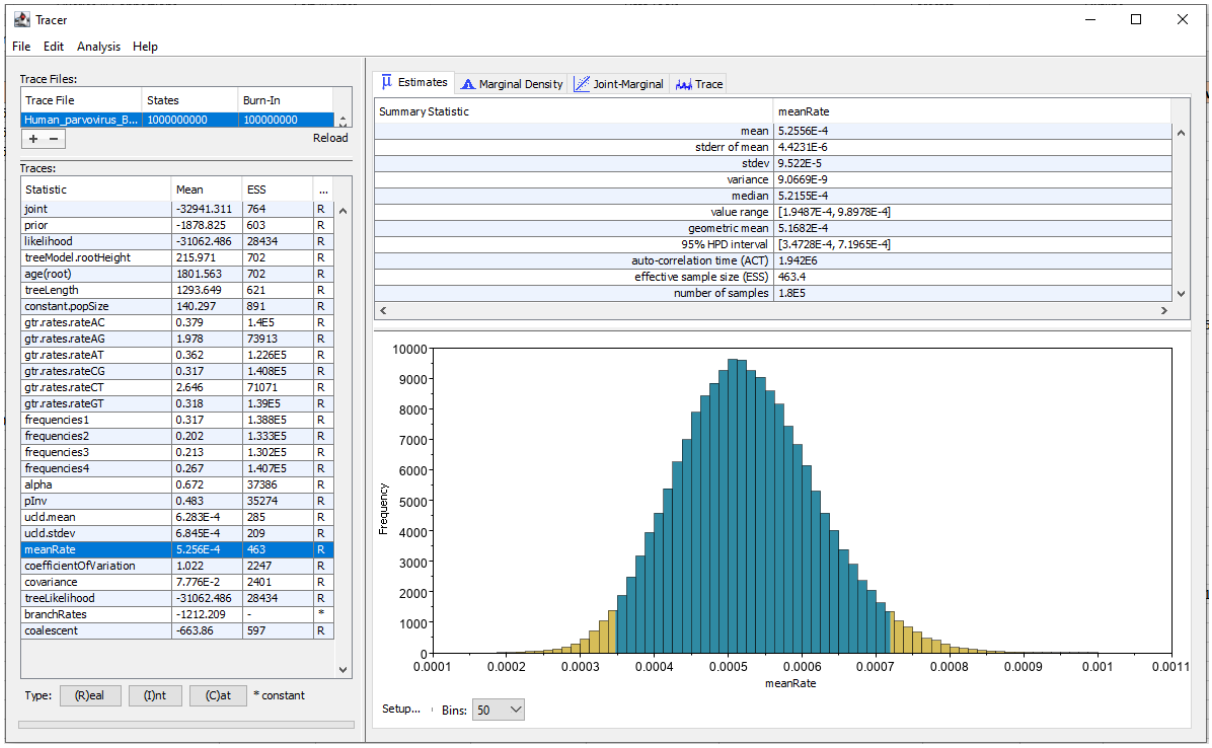

Figure 37: BEAST estimation of substitution rate (meanRate; substitutions/site/year) visualized in Tracer. The traces panel (left) lists all model parameters with their mean values and effective sample sizes (ESS). The highlighted **meanRate** parameter shows the average estimated evolutionary rate, with associated variance, 95% highest posterior density (HPD) interval, and other summary statistics displayed in the Summary Statistic panel (top right). The histogram (bottom right) depicts the posterior distribution of meanRate. The frequency plot indicates that the MCMC chain for this parameter has converged and is approximately normally distributed. An ESS > 200 is generally considered indicative of sufficient sampling; here, the ESS of 463 confirms reliable estimates.

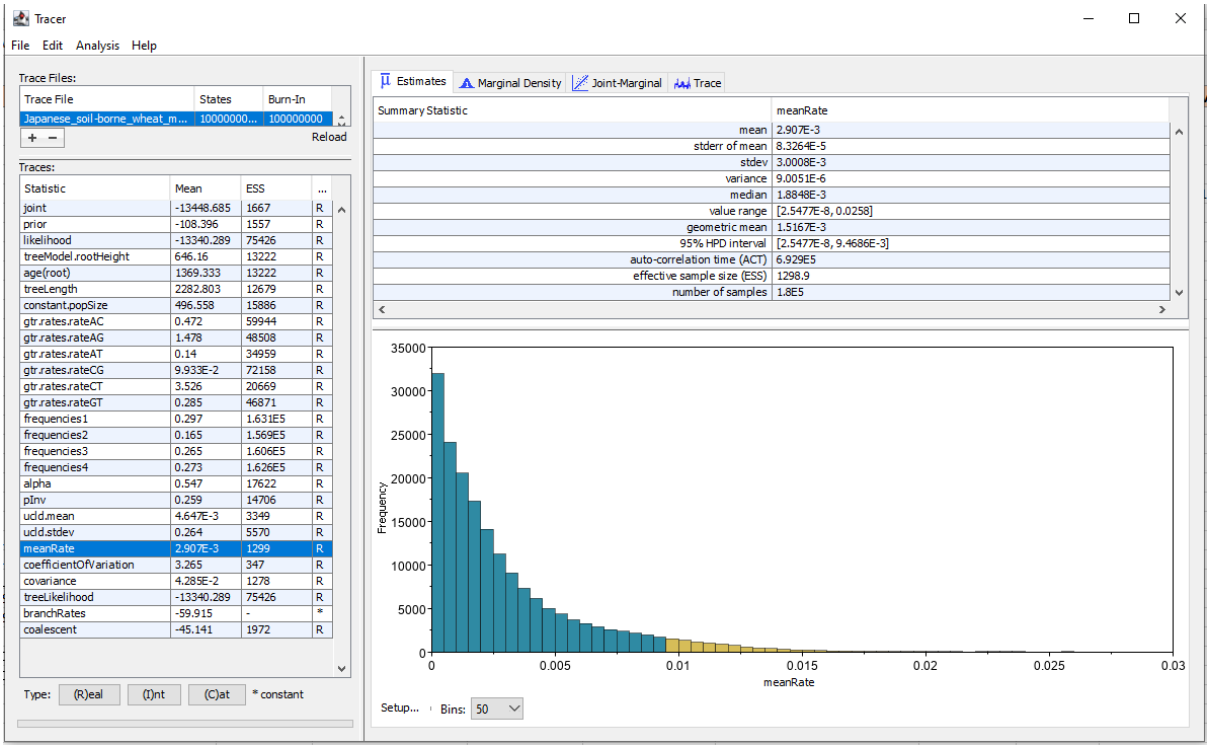

Figure 38: BEAST estimation of substitution rate (meanRate; substitutions/site/year). In some cases, such as shown here, the frequency plot displays an uneven peak rather than a fully converged, normally distributed parameter. This indicates that meanRate convergence was incomplete, which may arise from rate heterogeneity, limited temporal signal, or model fit limitations

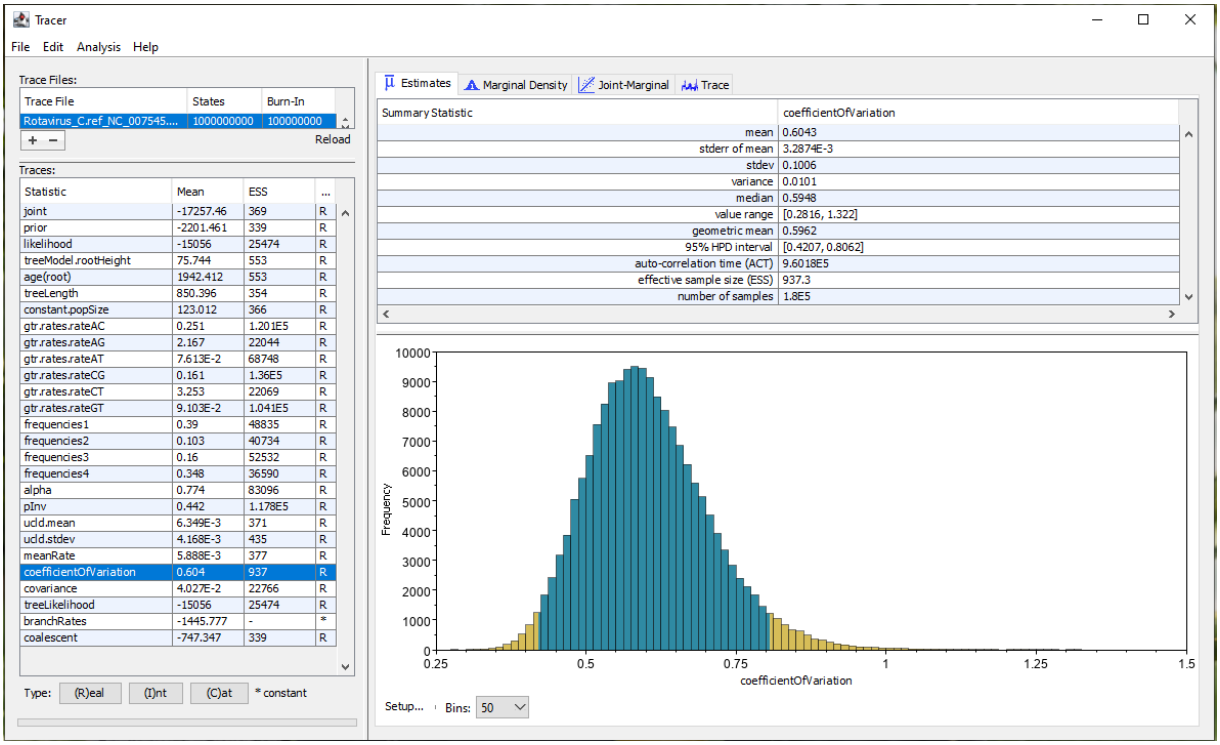

Figure 39: BEAST estimation of the coefficient of variation for Rotavirus C visualized over tracer. The frequency plot shows a clear central peak, indicating stable convergence and reliable parameter estimation across MCMC sampling.

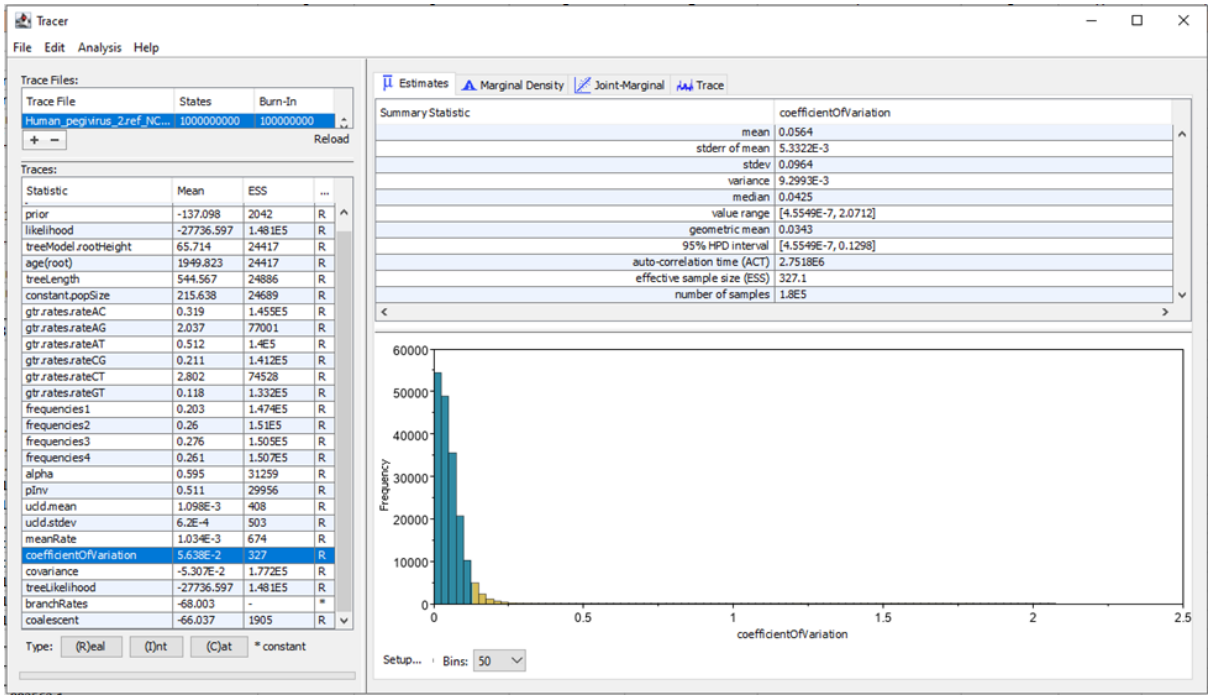

Figure 40: BEAST estimation of the coefficient of variation for Human_pegivirus_2 over tracer. The distribution on frequency plot appears narrowly concentrated toward low values, reflecting limited among-branch rate variation and restricted parameter spread.

3.4.4 Taxonomy distribution according to meanRate

- Taxonomy distribution for very slow evolving alignments

  Referring to Table 14, alignment datasets were divided to five categories according to mean rate values. In the first category which includes two alignments only, family taxonomy levels for the two alignments are 'Arenaviridae' and 'Hantaviridae'.

- Taxonomy distribution for slow evolving alignments

  For the second category that includes 31 alignments, 19 distinct families were identified. The top hit was 'Spinareoviridae' with 7 occurrences followed by 'Polyomaviridae' with 4 hits. In contrast, 'Retroviridae' and 'Secoviridae' are two examples from families with only a single occurrence, Figure 41 is a pie chart displaying distinct families distribution for this specific category.

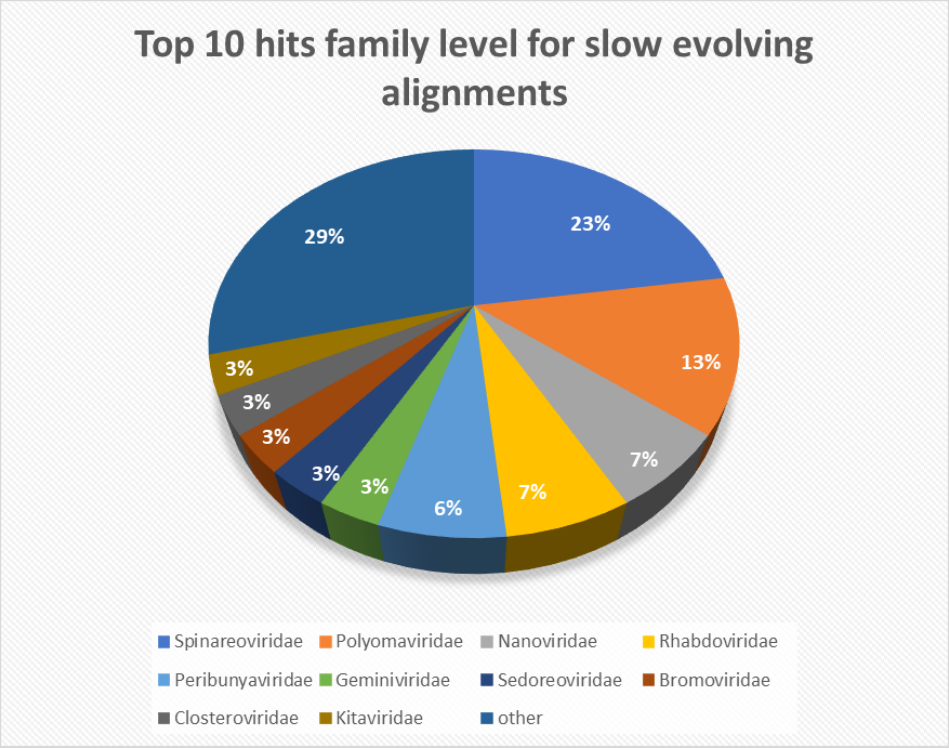

Figure 41: Pie chart illustrating the family-level taxonomic distribution of the top 10 hits for slow-evolving viral alignments. These alignments fall within the second evolutionary rate category, classified according to meanRate values. The chart highlights the proportional representation of each viral family, with Spinareoviridae and Polyomaviridae showing the highest contributions within this category, see supplementary Table S22.

- Taxonomy distribution for moderate evolving alignments

  For the third category that includes 150 alignments, 30 distinct families were identified. The top hit was 'Spinareoviridae' with 17 occurrences followed by 'Geminiviridae' with 16 hits. In contrast, 'Polyomaviridae' and 'Orthomyxoviridae' are two examples from families with only a single occurrence, Figure 42 is a pie chart displaying distribution of distinct families for this specific category.

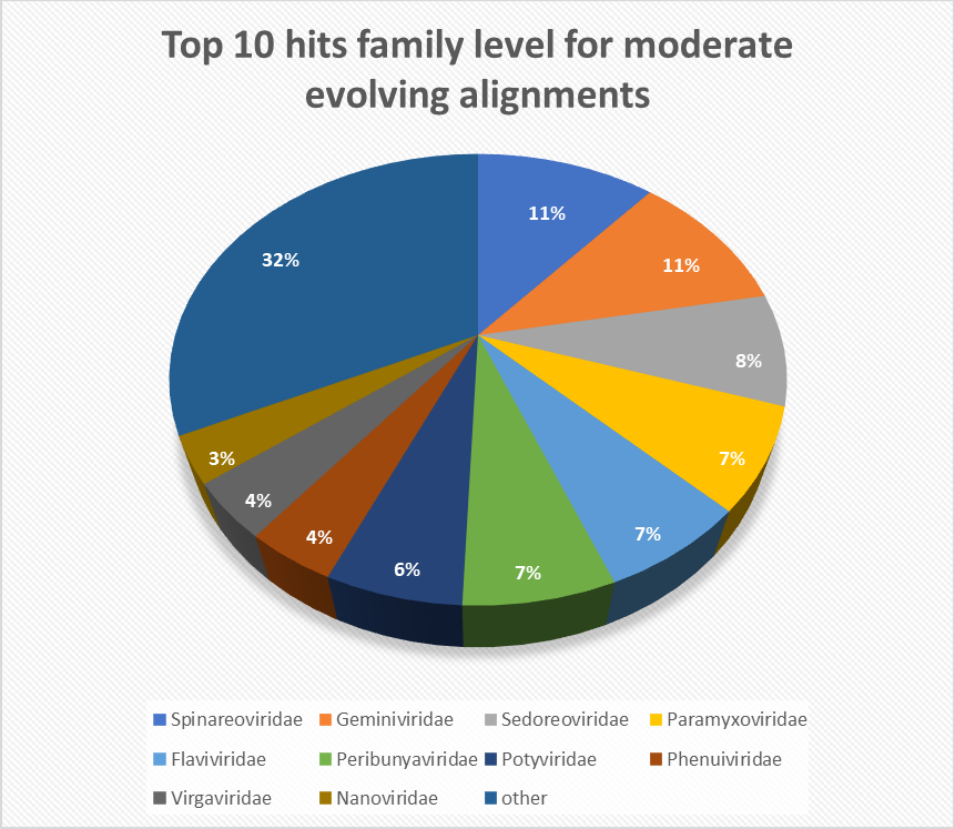

Figure 42: Pie chart illustrating the family-level taxonomic distribution of the top 10 hits for moderate-evolving viral alignments. These alignments correspond to the third evolutionary rate category, as defined by meanRate values. The chart shows a relatively balanced representation among several viral families, with Spinareoviridae and Geminiviridae contributing the largest proportions, see supplementary Table S23.

- Taxonomy distribution for fast evolving alignments

  For the fourth category that includes 145 alignments, 28 distinct families were identified. The top hit families recorded were 'Geminiviridae' with 19 occurrences followed by 'Sederoviridae' with 18 occurences. While other families have only single occurrence, e.g., 'Secoviridae' and 'Virgaviridae'. Figure 43 is a pie chart displaying distinct families distribution for this specific category.

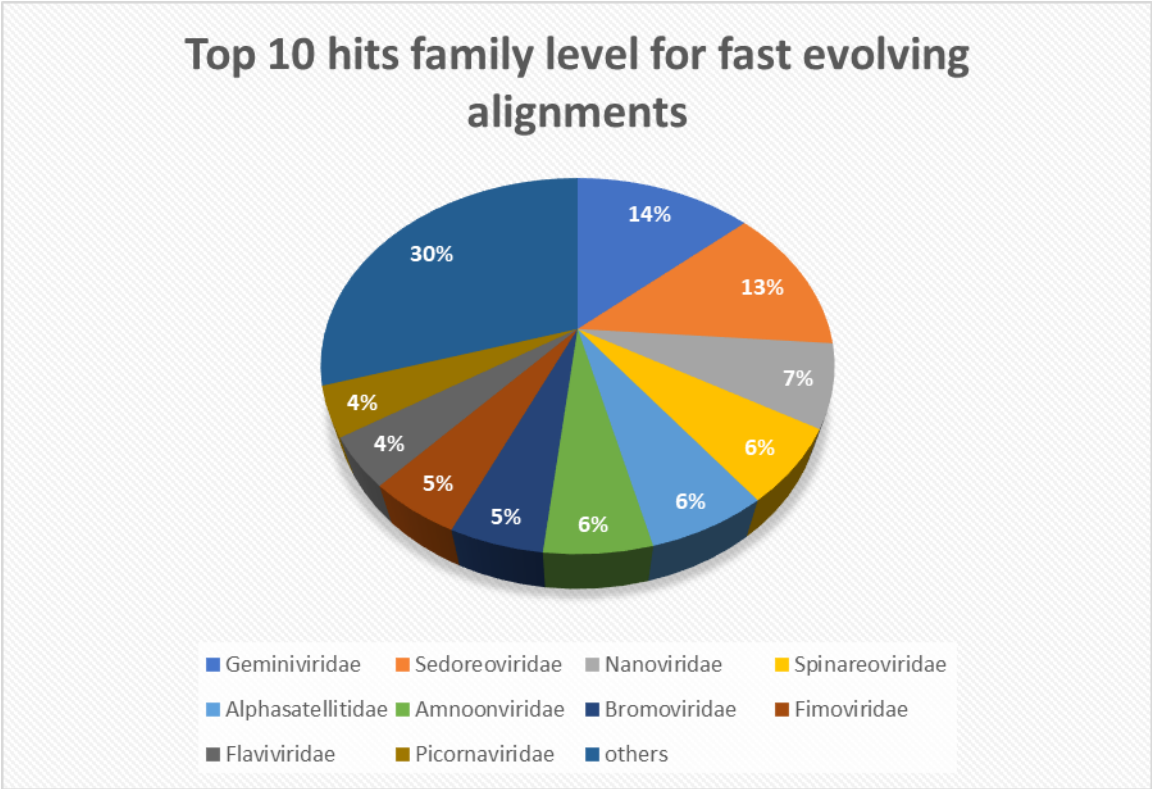

Figure 43Pie chart showing the family-level taxonomic distribution of the top 10 hits for fast-evolving viral alignments. These alignments correspond to the fourth evolutionary rate category, based on meanRate values estimated from BEAST analysis. Geminiviridae and Sedoreoviridae dominate this group, see supplementary Table S24.

- Taxonomy distribution for very fast evolving alignments

  For the fifth and last category that includes only 22 alignments, 11 distinct families were identified. The top hit families recorded were 'Picornaviridae' and 'Geminiviridae'. Other families with only single occurrence as; 'Flaviviridae' and 'Noraviridae'. Figure 44 is a pie chart displaying distinct families distribution for this specific category.

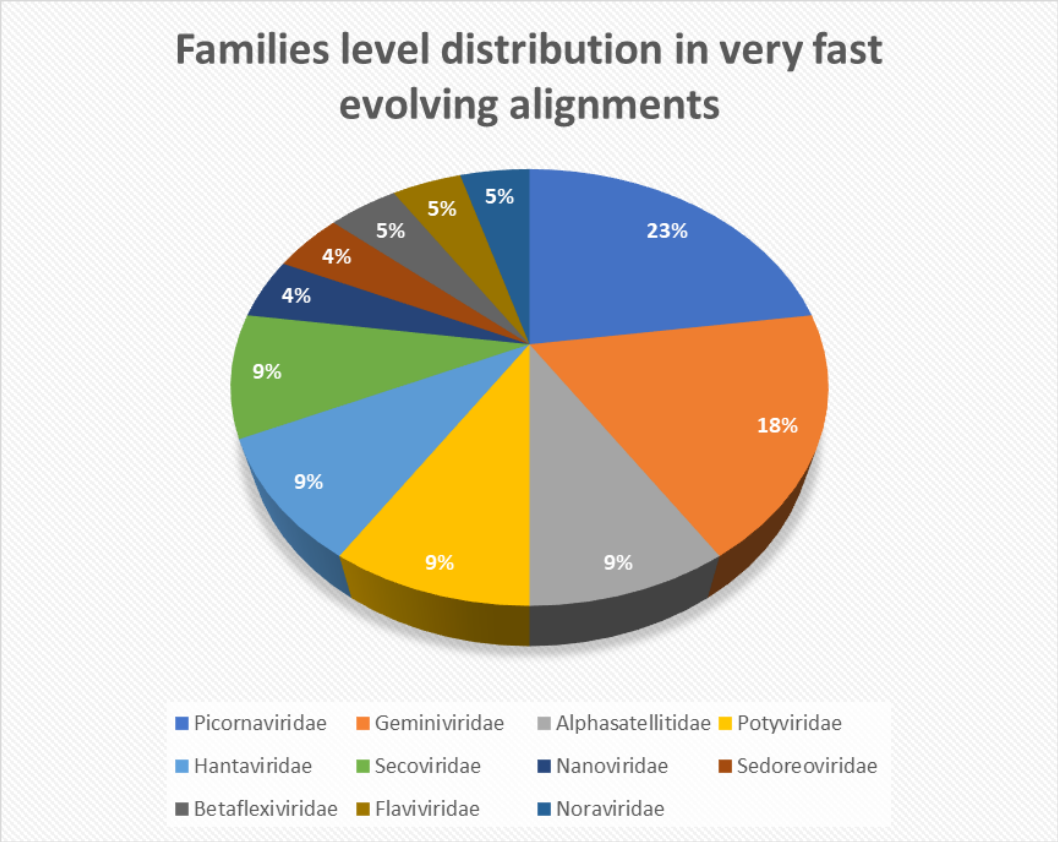

Figure 44: Pie chart showing the family-level taxonomic distribution for very fast-evolving viral alignments, representing the fifth and highest evolutionary rate category based on meanRate values. Picornaviridae and Geminiviridae account for the largest proportions, followed by Alphsatellitidae, Potyviridae, Hantaviridae, and Secoviridae, each contributing equally at 9%, see supplementary Table S25.

3.4.5 Patterns in substitution rate analysis

- Segmented Viruses

  According to coefficient of variation threshold mentioned in section 3.4.2, values >=1 were removed from analysis and were not included in segmented viruses' concordance-discordance study. The remaining alignments represented 27 segmented viruses where each one includes 2 to 7 segments as alignments datasets. Two meanRate patterns appeared for each of these segmented datasets, which are:

  a. Concordant pattern: among segmented viruses 14 showed concordant pattern, where all segments meanRate falls in the same speed category. For example, Rotavirus C had 6 segments which passed the filtering criteria, and they all scored meanRate values between (1.01×10^{-3} to 5.9×10^{-3} s/s/y) for the fast-

evolving viruses. While African horse sickness virus falls in the moderate evolving category with 6 segments, meanRate values were between (2.39×10^{-4} to 2.96×10^{-4} s/s/y).

b. Discordant pattern: on the other hand, 13 segmented viruses showed discordant pattern, where segments meanRate has more than one category speed. For example: Mammalian orthoreovirus 3 has 9 segments took part in the analysis where 2 falls in the moderate category and the remain 7 where fast evolving. Avian orthoreovirus has a discordant pattern, with 10 segments falls between slow, moderate, and fast evolving alignments. Table 17 displays number of segmented viruses for each pattern.

Table 17: Concordance and discordance patterns for meanRate values in 27 segmented viruses, see supplementary Table S26.

Number of segmented viruses	**Concordant Pattern**	**Discordant Pattern**
27	14	13

- Multiple taxonomy hierarchy level

  Following similar analysis performed in sections 3.2.5 and 3.3.4, taxonomical levels patterns were studied for concordance and discordance in substitution rate according to meanRate and coefficient of variation values. Order, family and genus levels counting, and consideration continues as previously done, but with additional alignments datasets number. Table 18 lists order, family and genus levels with the number of patterns appears for each taxonomical level with original number of each level accepted in analysis.

Two meanRate patterns appeared for each of these segmented datasets, which are:

a. Concordant pattern: any taxonomic level considered in a concordant pattern if only one meanRate category appeared.
b. Discordant pattern: any taxonomic level considered in a discordant pattern if more than one meanRate category appeared.

Below are numbers of concordant and discordant patterns corresponding to each taxonomy level with examples.

Order level: 8 order levels took part in substitution rate concordance study. 'Hepelivirales' was the only order showed concordant pattern with all entries have moderate evolving speed, while the other 7 orders showed discordant pattern for more than one meanRate category. For example, 'Reovirales' has 3 different speeds, slow, moderate and fast. Also 'Tymovirales' has 3 speeds moderate, fast and very fast.

Family level: Out of 13 families only 2 showed concordant pattern in evolution speed which are 'Phenuiviridae' and 'Solemoviridae' with all entries showed moderate speed alignments. While the other 11 families showed discordant pattern as 'Paramyxoviridae' that has 3 different speeds, slow, moderate, and fast, also 'Picornaviridae' showed slow, moderate, fast, and very fast evolution speeds.

Genus level: Out of 27 genera, only 8 showed concordant patterns as 'Rotavirus' were all species fall in the fast speed category. On the other hand, the other 19 genera showed discordant pattern, for example, 'Begomovirus' had 3 different speeds, moderate, fast and very fast. And 'Orbivirus' showed slow, moderate, fast and very fast evolving alignments. Table 18 displays number of every taxonomy level included in analysis and number of each pattern studied.

Table 18: Concordance and discordance patterns for meanRate values in taxonomy different levels.

	Number	**Concordant Pattern**	**Discordant**
Order	8	1	7
Family	13	2	11
Genus	27	8	19

3.4.6 Host association in concordant species

For substitution rate analysis concordant genera were linked to viral hosts. Similar to previous parameters analysed, host association continues for substitution rates. Viral species for the 8 concordant genera and 2 concordant families were identified and linked to their hosts. Table 19A is listing concordant genera with their corresponding species, among 17 species examined, 9 were found with mammalian hosts and 8 with plant hosts, no appearance of any other host types among concordant genera. Table 19B lists concordant families with their genera, species, and relevant hosts.

Table 19: A. Lists concordant genera with their associated species, and the respective host for each species. For discordant genera, see supplementary Table S27. B. Lists concordant families with their associated genera, species and the respective host for each species.

Genus	Species	Host
Babuvirus (concordant Fast evolving)	Cardamom bushy dwarf virus	Plant
	Banana bunchy top virus	Plant
Pestivirus (concordant moderate evolving)	Bovine viral diarrhea virus	Mammalian
	Atypical porcine pestivirus 1	Mammalian
	Classical swine fever virus	Mammalian
Mastrevirus (concordant Fast evolving)	Panicum streak virus	Plant
	Paspalum striate mosaic virus	Plant
Rotavirus (concordant fast evolving)	Human Rotavirus B	Mammalian
	Rotavirus C	Mammalian
Respirovirus (concordant moderate evolving)	Porcine respirovirus 1	Mammalian
	Human respirovirus 1	Mammalian
Tenuivirus (concordant moderate evolving)	Rice stripe tenuivirus	Plant
	Rice grassy stunt tenuivirus	Plant
Lyssavirus (concordant slow evolving)	Australian bat lyssavirus	Mammalian
	European bat 1 lyssavirus	Mammalian
Emaravirus (concordant fast evolving)	Fig mosaic emaravirus	Plant
	European mountain ash ringspot-associated emaravirus	Plant

Table 19B

Family	Genus	Species	Host
Solemoviridae (concordant moderate evolving)	Enamovirus	Citrus vein enation virus	Plant
	Sobemovirus	Rice yellow mottle virus	Plant
Phenuiviridae (concordant moderate evolving)	Phasivirus	Phasi Charoen-like phasivirus	Plant
	Tenuivirus	Rice stripe tenuivirus	Plant
		Rice grassy stunt tenuivirus	Plant

3.4.7 Overview of Substitution rate analysis results

Substitution rate analysis was conducted on 350 alignments after a second filtration easing in age gap with high temporal signal R >= 0.5. BEAST estimated meanRate values categorized alignments into five evolutionary speed classes: very slow, slow, moderate, fast and very fast. The majority of alignments fell within moderate and fast categories, representing 45% and 50% respectively. Coefficient of variation was also assessed revealing that 185 datasets exhibited rate heterogenicity requiring a relaxed clock model, while 156 datasets were unsuitable for divergence time estimation. Visual inspection of BEAST output through tracer confirmed convergence for most cases. Taxonomic distribution across speed categories showed distinct pattern. In slow evolving alignments Spinareoviridae and Polyomaviridae were the most represented families. For moderate evolving alignments Spinareoviridae and Geminiviridae were the highest. In the fast-evolving alignments Geminiviridae and Sedoreviridae were at the top hits. Following are the very fast evolving alignments with Picornaviridae and Geminiviridae being the highest among other families. Segmental concordance analysis showed that approximately half of segmented viruses had all segments falling within the same evolutionary category and the other half showed discordance in rates pattern. Taxonomic concordance patterns across order, family and genus levels indicated limited consistence, with concordant rate categories found in only 1 out of 8 orders, 2 out of 13 families, and 8 out of 27 genera. Host association analysis revealed balanced representation of plants and mammalian viruses among concordant families and genera.

3.5 Positive selection

3.5.1 SLR detection for positive selected sites

The SLR (Sitewise Likelihood Ratio) software run on all alignment data sets. Unlike the BEAST analysis, R values were not used as an additional filter, therefor SLR was applied on 448 alignments. For each alignment, an output file was generated containing positive selected sites and their location. Subsequently, these output files were transferred to a spreadsheet, enabling the extraction of summary statistics for the presence of positively selected sites within each alignment dataset.

- Number of alignments with positive selected sites (%)
    - 60% of total alignments have at least one positively selected site. Alignments species has different positive selected sites number ranging from (1 to 203) for each alignment. Figure 45 is a staggered scatter plot with logarithmic scale Y axis representing the distribution of positively selected sites among viral alignment data sets. Alignments were ranked in ascending order from 1 positively selected to 203 and the number in each category plotted on a logarithmic scale.

    - A scatter graph for number of selected sites against alignment lengths was plotted in Figure 46. The absence of linear correlation means that length of viral peptide sequence does not consistently predict the presence or absence of positively selected sites.

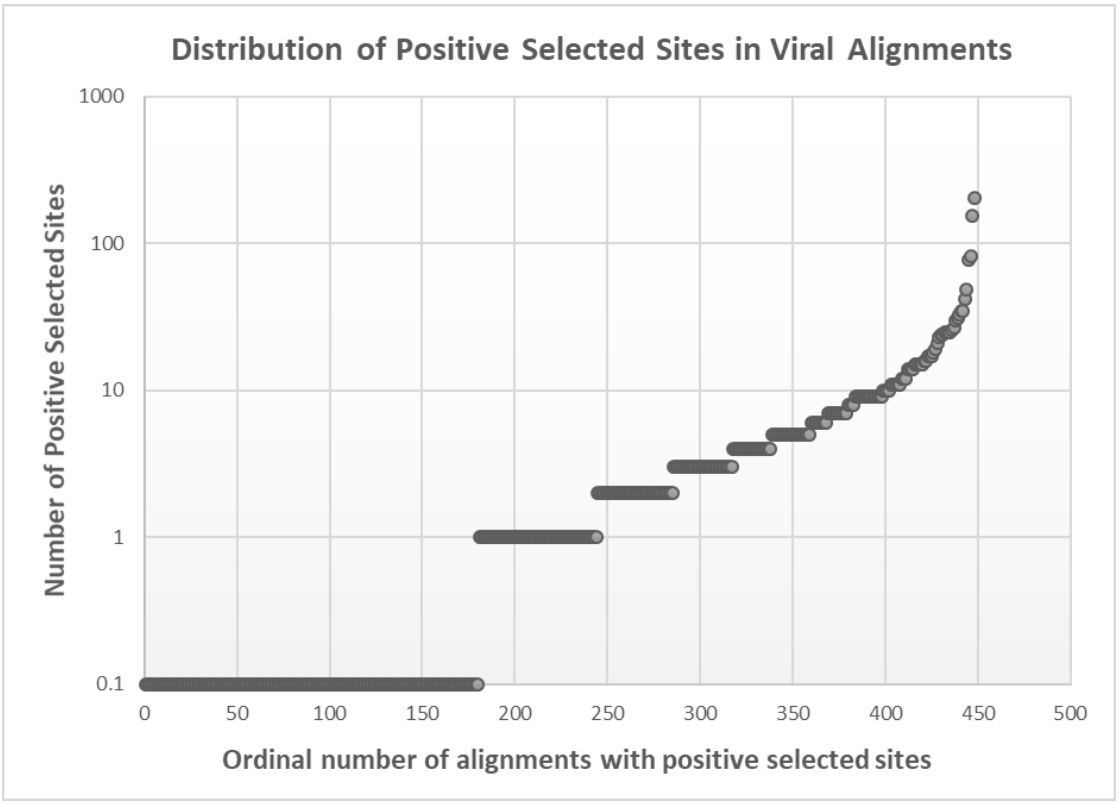

Figure 45: A single dimensional scatter plot for positive selected sites distribution in viral alignments, showing number of positive selected sites in Y-axis represented in a logarithmic scale (starts with 0.1=0 site), against X-axis for alignments with positive sites ordinal number. See supplementary Table S28.

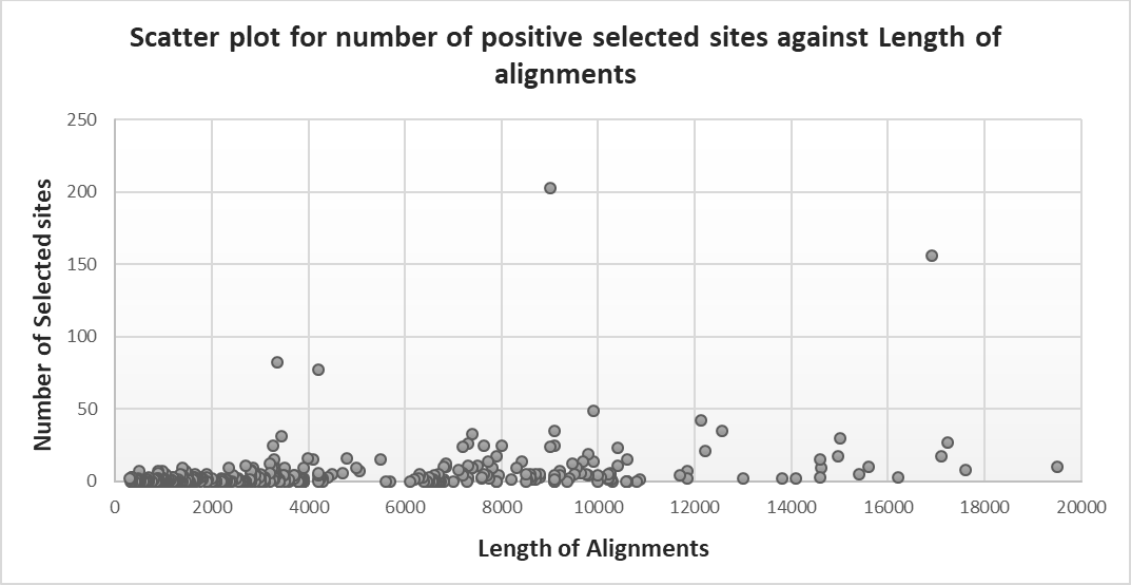

Figure 46: Scatter plot illustrating the relationship between the number of positively selected sites and the length of viral sequence alignments. Each point represents an individual alignment, with the *x*-axis showing alignment length (in nucleotides) and the *y*-axis showing the total number of sites under positive selection as detected by SLR. No strong linear correlation is apparent, suggesting that alignment length alone does not determine the extent of positive selection.

- Overall Kappa and Omega values against length of alignment

  Kappa and Omega are additional numerical parameters included in the SLR run output files.

  - Kappa (κ): is a parameter which express ratio of the transition/transversion rate.
  - Omega (ω): is a parameter that represents the nonsynonymous over synonymous substitution rate, dN/dS ratio. Omega is the primary indicator of average positive selection across the entire alignment as its value reflects strength of positive selection.

  To have a better understanding if the evolutionary rate variation is influenced by alignment length, Kappa values were plotted against length of alignment in a scatter plot represented in Figure 45. The plot aimed to study correlation between the transition/transversion rate ratio and the length of alignments. As displayed in Figure 47 there is no linear correlation between values and alignment lengths.
  In a similar manner, to determine whether alignment length has an influence on the selective pressure in the context of sequence evolution, a second scatter plot was created to explore the relationship between alignment length and Omega values in Figure 48. However, as with Kappa values, Omega values also displayed a scattered pattern with no linear correlation with alignment lengths.

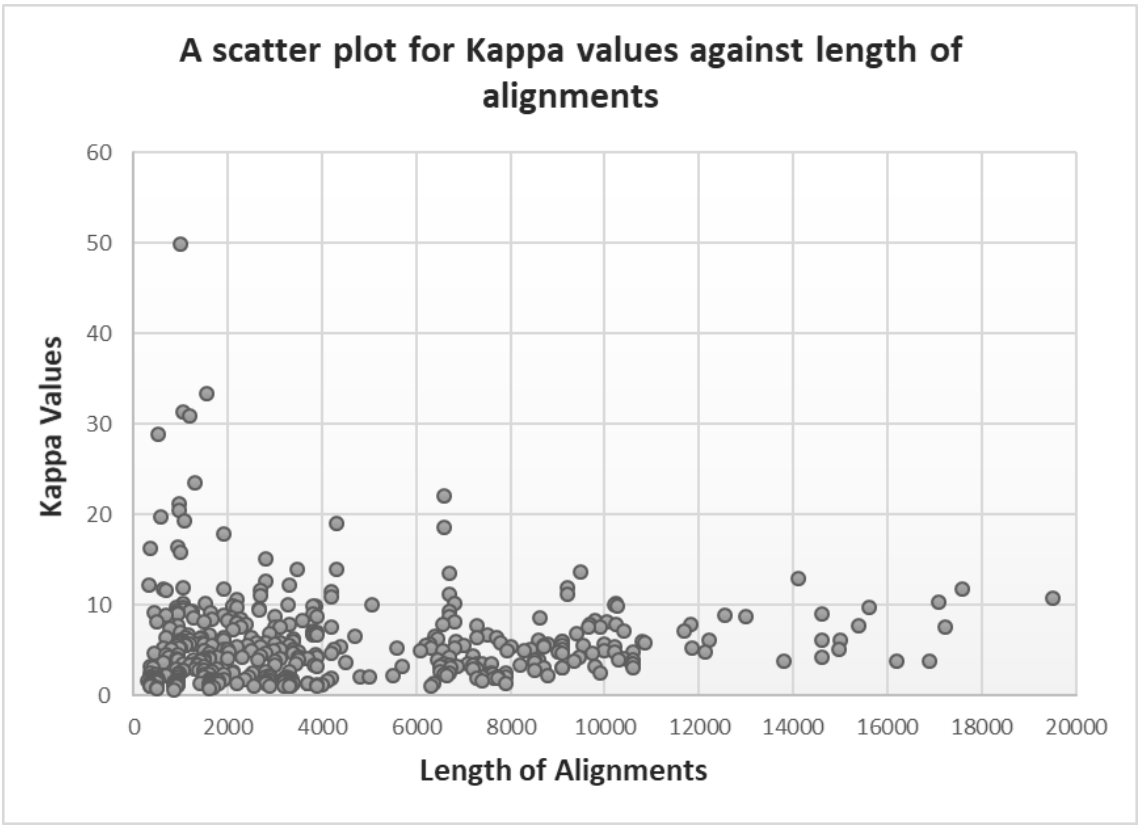

Figure 47: Scatter plot illustrating the distribution of kappa values, against the length of viral sequence alignments that contain positively selected sites. The x-axis indicates alignment length, and the y-axis shows the corresponding kappa values. No clear relationship is observed, suggesting that substitution bias is not strongly influenced by sequence size within this dataset. See supplementary Tables S29 and S30 for the highest and lowest alignments.

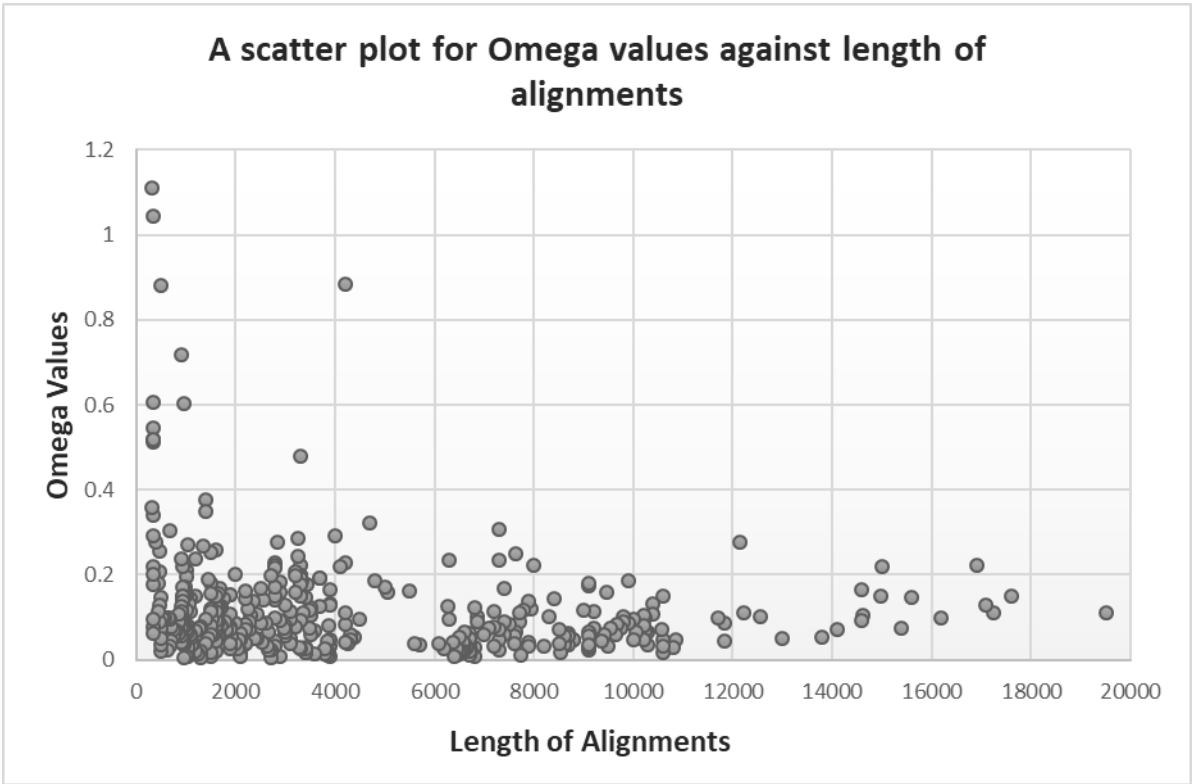

Figure 48: Scatter plot showing the distribution of Omega values, against the length of viral sequence alignments containing positively selected sites. The x-axis represents alignment length, while the y-axis shows Omega values. There is no clear correlation between Omega and alignment size, suggesting that the overall selection pressure is independent of sequence length in this dataset. See supplementary Tables S31 and S32 for the highest and lowest alignments.

3.5.2 Taxonomy distribution according to positive selection

- Taxonomy distribution for alignments with positive selected sites

  According to SLR analysis, within the 268 alignments with positive selected sites 139 unique taxonomies classification were observed, among them 48 distinct families were identified. The top hit was 'Geminiviridae' with 45 occurrences followed by 'Potyviridae' with 18 hits, while 'Solemoviridae' and 'Pneumoviridae' recorded only one hit, Figure 49 is a pie chart displaying distribution of distinct families for alignments with positive selected sites.

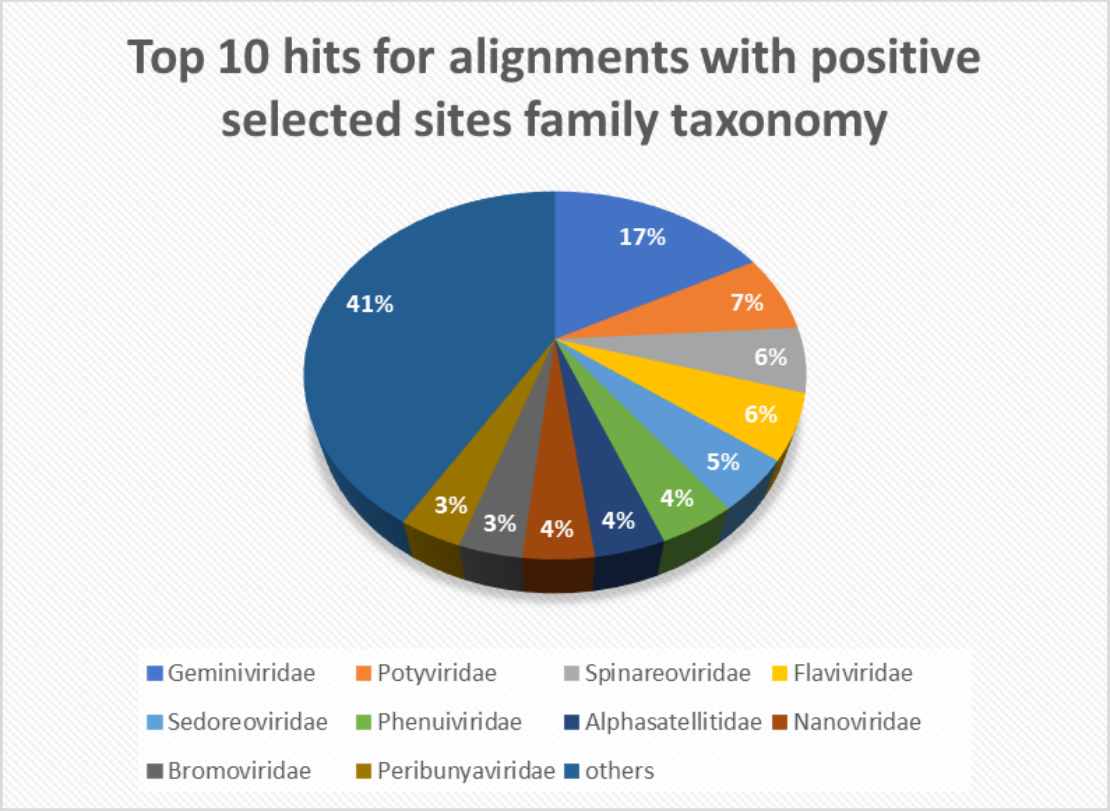

Figure 49: Pie chart illustrating the family-level taxonomic distribution of viral species alignments that contained positively selected sites, as identified by the Sitewise Likelihood Ratio (SLR) method. The largest proportion corresponds to the Geminiviridae family (17%), followed by Potyviridae (7%) and Spinareoviridae (6%). See supplementary Table S33.

- Taxonomy distribution for alignments with No positive selected sites

  The remainder 180 alignments did not show any positive selected sites in peptide sequence by SLR output. Within these alignments 64 unique taxonomies were

observed, among them 30 distinct families were identified. The top hit was 'Sedoreoviridae' with 38 occurrences followed by 'Spinareoviridae' with 25 hits, while 'Alphaflexiviridae' and 'Tospoviridae' recorded only one hit, Figure 50 is a pie chart displaying distinct families distribution for alignments with No positive selected sites evidence at SLR run.

9

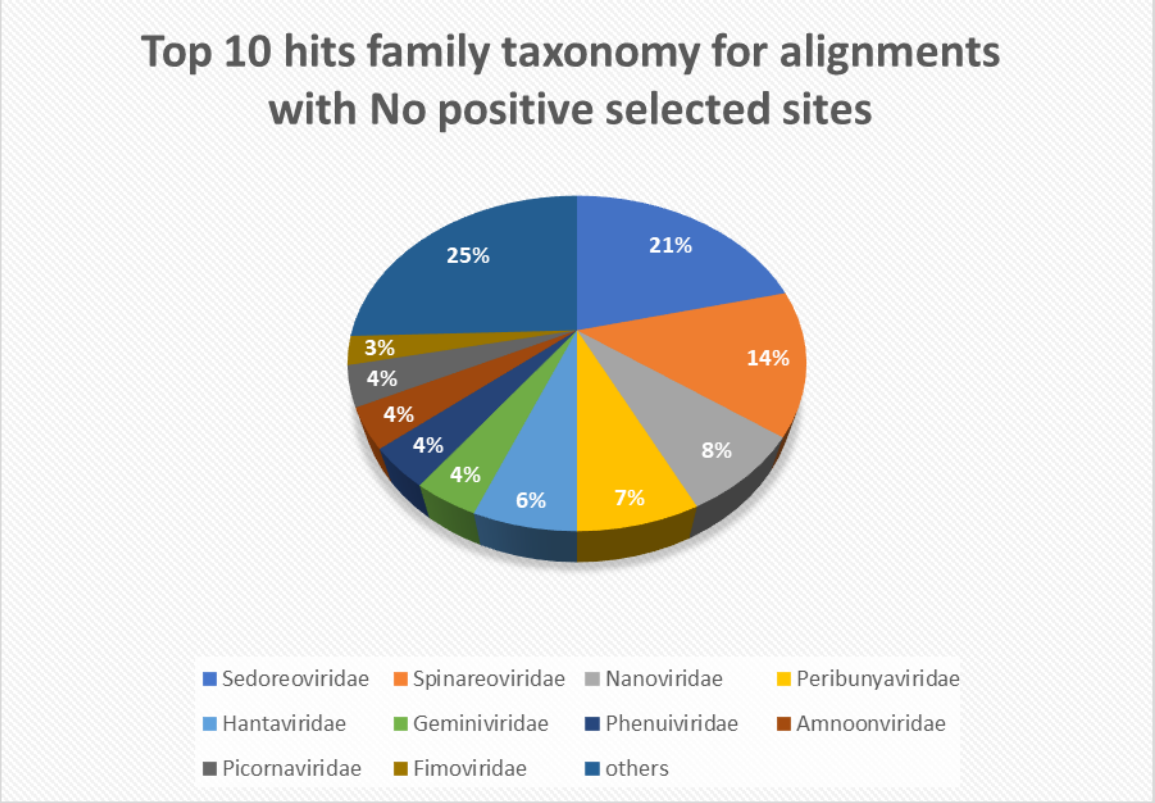

Figure 50: Pie chart showing the family-level taxonomic distribution of viral species alignments without positively selected sites, as measured by the Sitewise Likelihood Ratio (SLR) method. The largest proportion is represented by the Sedoreoviridae family (21%), followed by Spinareoviridae (14%) and Nanoviridae (8%). See supplementary Table S34.

3.5.3 Selected sites polymorphism

- Polymorphism amino acids in alignments with positive selected sites

  Out of 2137 positive selected sites identified, only those with two or more plus signs ++ (SLR's indicator of statistical significance of selection at a specific site) were chosen for polymorphism analysis. Following, 965 sites were picked based on the presence of ++, +++, and +++ to proceed with the analysis.

  Alignments peptide sequences were visualized over MEGA software to check polymorphism amino acids for each positive selected site, number of polymorphisms among alignment data sets ranges from 2 to 13 amino acids.

  For the 965 positive selected sites, the correlation coefficient for polymorphism number with Omega values was -0.00813 and with thickness of alignments was 0.0858. Furthermore, correlation coefficient values were calculated for viral hosts, separating the 965 entries into five categories. Table 20 lists the five categories for each host type with number of species, number of positive selected sites and correlation coefficient against site wise Omega values.

Table 20: Five viral hosts yielded from polymorphism analysis with their corresponding number of species, number of positive sites, average site-wise Omega per selected site, average amino acid diversity per selected site and correlation coefficient values for each host type polymorphism (diversity) number against Omega values.

	Number of Species	**Number of Positive Sites**	**Average Site-wise Omega**	**Average Diversity**	**Correlation Coefficient**
Avian	9	37	6.4	3	0.05
Fish	8	70	10.2	3	-0.16
Insect	6	13	11.6	3	-0.35
Mammalian	65	513	12.13	4	0.11
Plant	111	332	11.5	3	-0.21

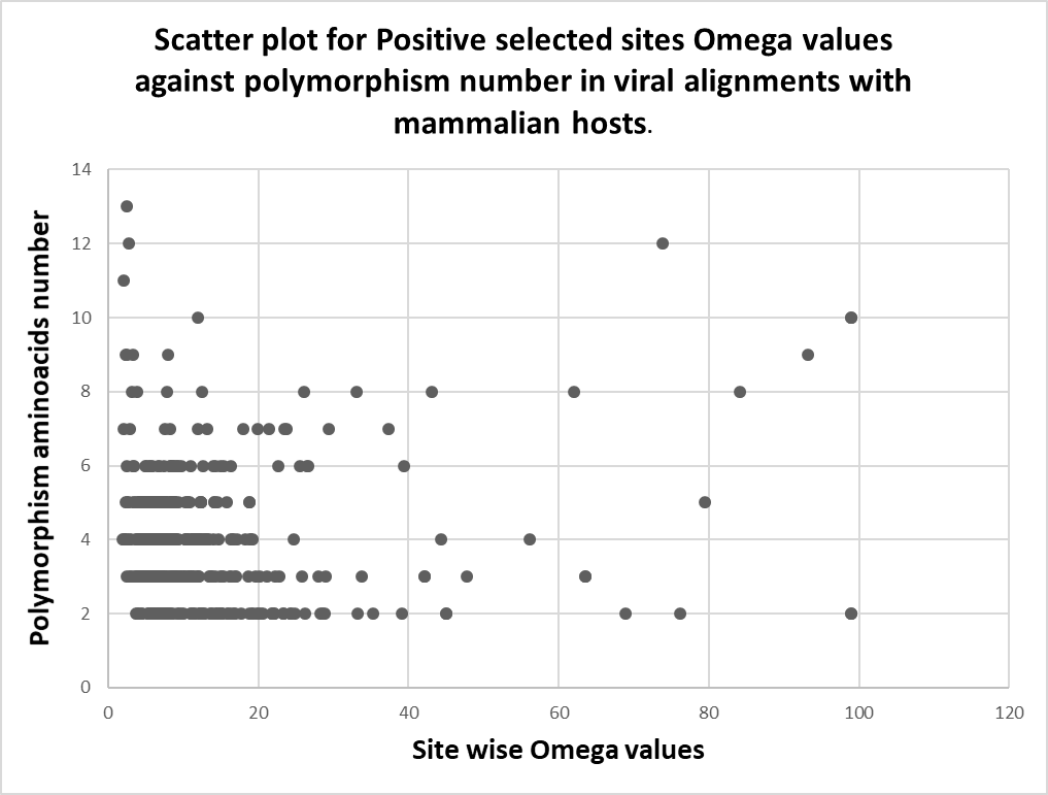

Figure 51: Scatter plot illustrating the relationship between sitewise omega values and the number of polymorphic amino acids at positively selected sites in viral alignments with mammalian hosts only, as mammalian has the highest correlation value between diversity and Omega values among other hosts. With the x-axis for Omega values, and the y-axis showing the count of amino acid polymorphisms observed at that site.

- Selected sites in phylogenetic trees

  Following polymorphism analysis performed, phylogenetic trees were chosen considering Omega values and polymorphism number to check if positive sites appear in directional or diversifying selection. Cases were studied as follows:

  - High Omega – low diversity

    Zucchini yellow mosaic virus alignment was chosen with selected sites number 1856 on peptide sequence. The selected site recorded Omega value 99 (SLR has 99 as its ceiling for omega) with 2 polymorphism amino acids. Figure 52 is the bootstrap tree highlighting branches for the location of the positive site with the amino acids appeared in polymorphism.

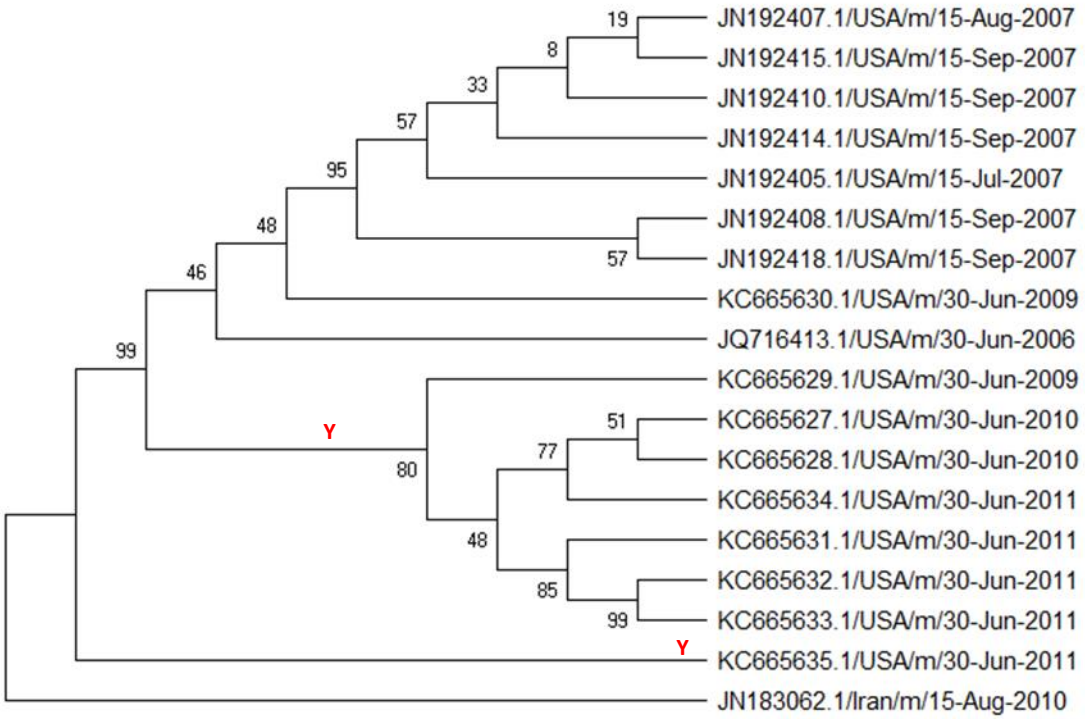

Figure 52: Bootstrap cladogram for Zucchini yellow mosaic virus showing two branches with two polymorphism amino acids Tyrosine (Y) and the original amino acid in the remaining sites was Leucine (L).

- High Omega – high diversity

  Kibale red colobus virus alignment was chosen with selected site number 5194 on peptide sequence, that recorded Omega value 99 with 12 polymorphism amino acids. Figure 53 is the bootstrap tree highlighting branches for the location of the positive site with the amino acids appeared in polymorphism.

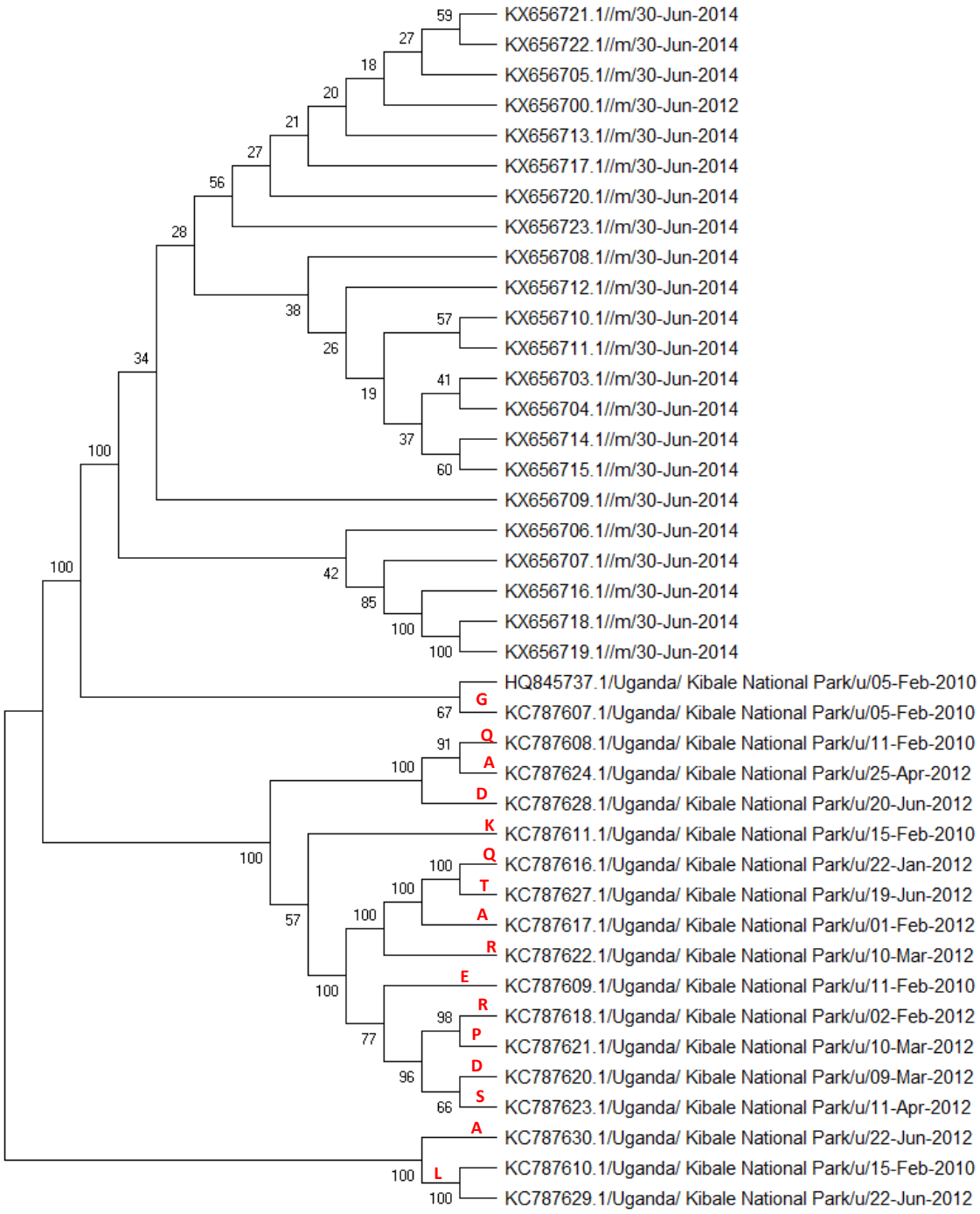

Figure 53: Bootstrap cladogram for Kibale red colobus virus showing branches with 12 polymorphism amino acids Glycine (G), Glutamine (Q), Aspartic acid (D), Alanine (A), Lysine (K), Threonine (T), Glutamic acid (E), Arginine (R), Proline (P), Serine (S) and Leucine (L) while the original amino acid in the remaining site was Asparagine (N).

- Low Omega – high diversity

  Nanovirus-like particle alignment was chosen with selected site number 286, that recorded Omega value 2.6 with 8 polymorphism amino acids. Figure 54 is the bootstrap tree highlighting branches for the location of the positive site with the amino acids appeared in polymorphism.

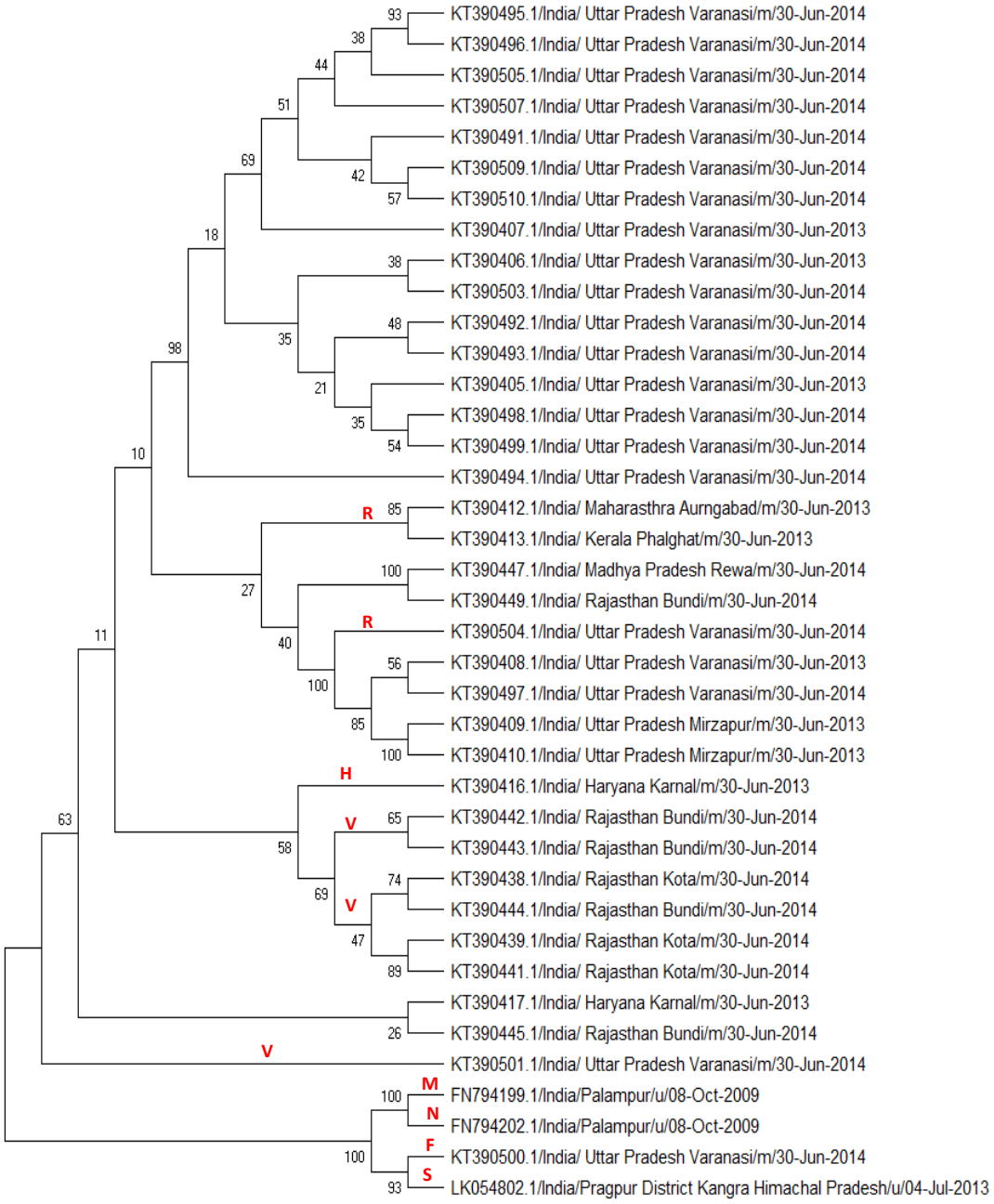

Figure 54: Bootstrap cladogram for Nanovirus-like particle showing branches with 8 polymorphism amino acids Arginine (R), Serine (S), Histidine (H), Valine (V), Methionine (M), Phenylalanine (F) and Asparagine (N) while the original amino acid in the remaining site was Glutamine (Q).

- Low Omega – low diversity

  Infectious pancreatic necrosis virus alignment was chosen with selected site number 482 that recorded Omega value 8 with 2 polymorphism amino acids. Figure 55 is the bootstrap tree highlighting the location of the positive site with the amino acid appeared in polymorphism.

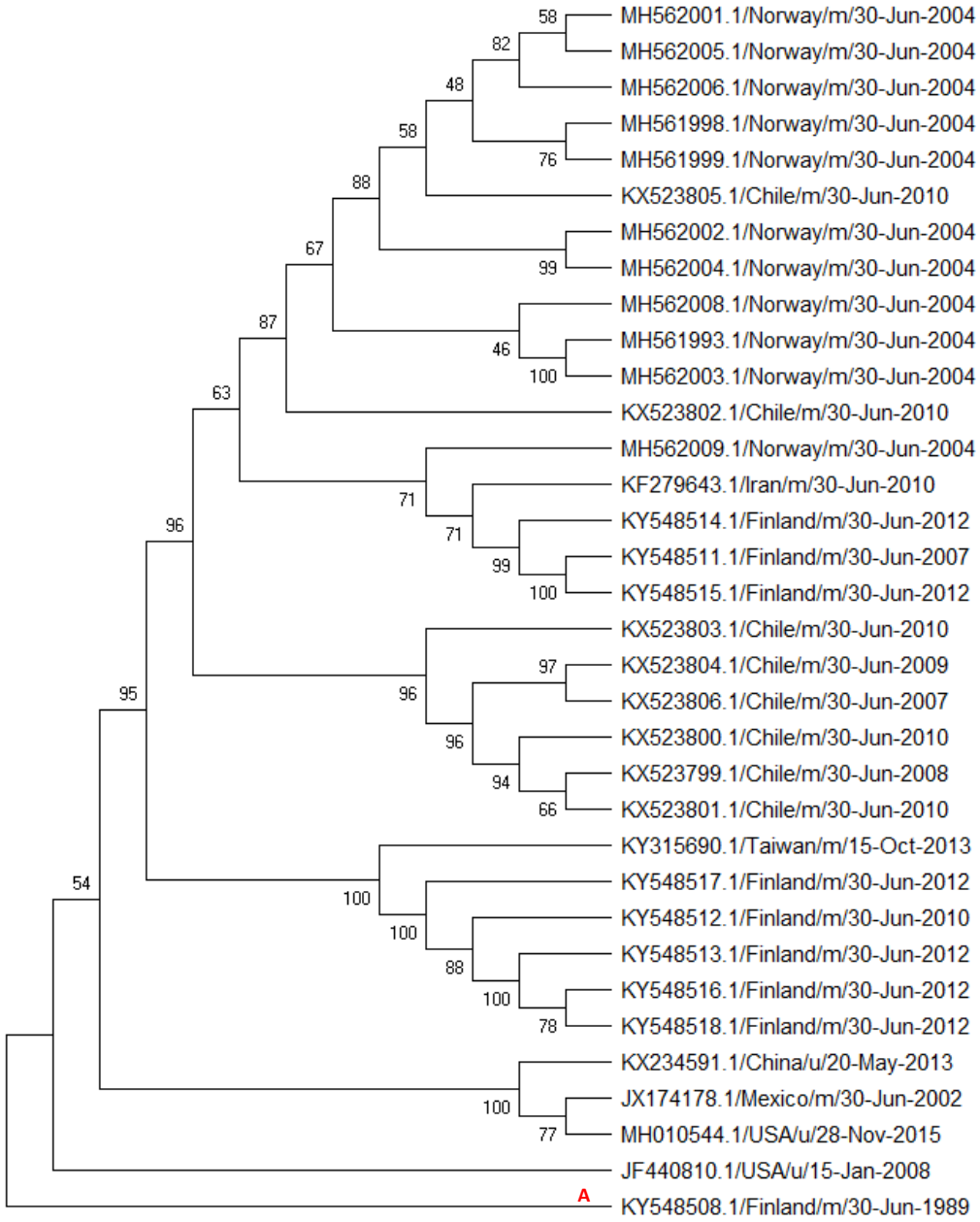

Figure 55: Bootstrap cladogram for Infectious pancreatic necrosis virus showing two polymorphism amino acids Alanine (A) on the sequence from the common ancestor and the original amino acid for the remain sequences was Serine (S).

3.5.4 Selected sites proteins functions

- Gene Ontology

As previously mentioned in Methods chapter sections 2.9.2 and 2.9.3, InterProscan data were collected for proteins with positive selected sites. Out of 353 proteins 300 had data on InterProscan domain while the remaining 53 has no data appeared in the output. Furthermore, number of 79 GO ids were identified as an output with several hits in InterProscan domain. The top hit GO term was GO:0005198 related to "structural molecular activity" function in QuickGo, appeared with 99 protein IDs, followed by GO:0003723 "RNA binding" function appeared with 82 hits. While terms GO:0046812 "host cell surface binding" function and GO:0007165 "signal transduction" function recorded only one hit in the InterProscan domain for selected proteins. Figure 56 is a pie chart for Gene Ontology top hits in InterProscan with their corresponding functions.

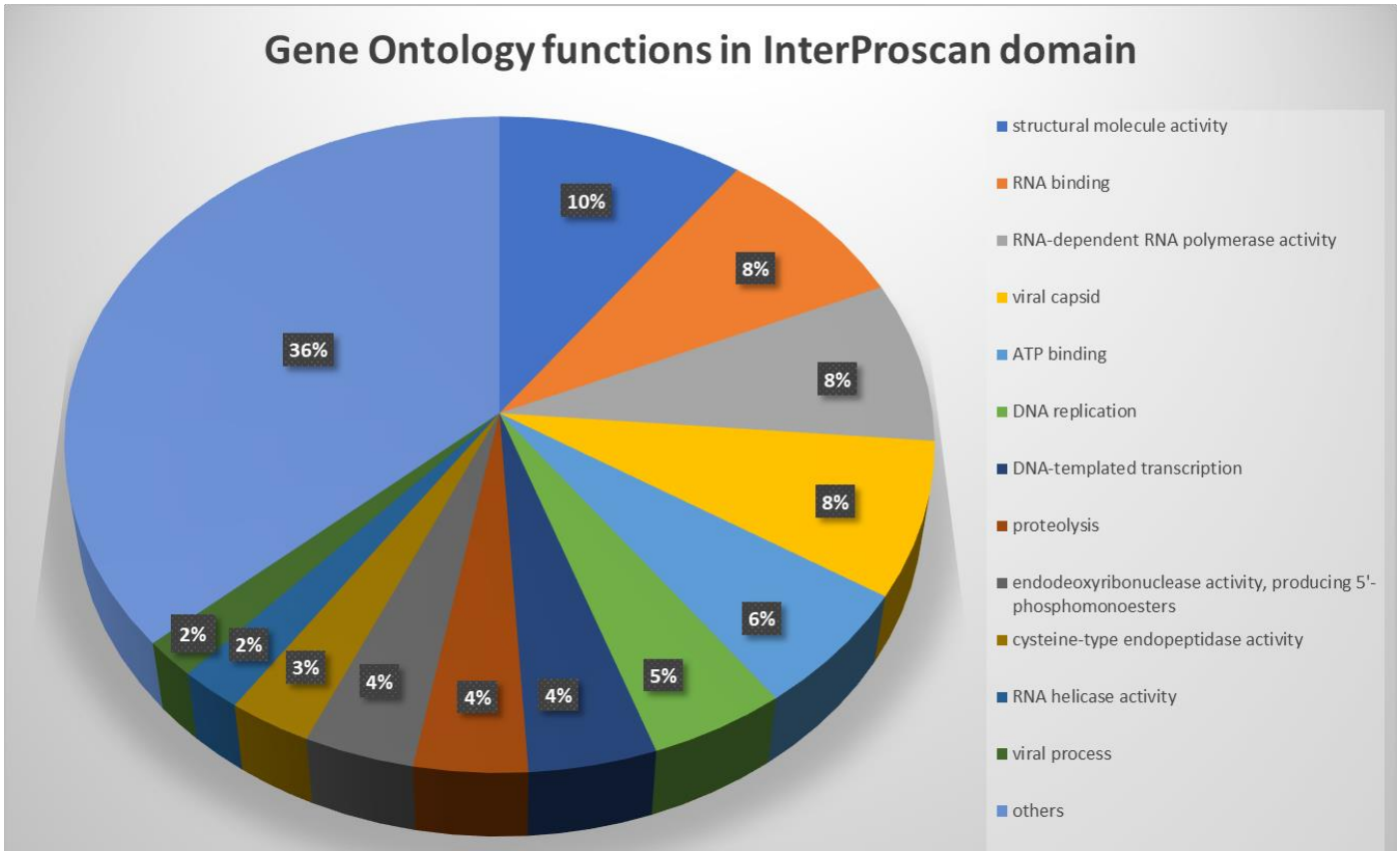

Figure 56: Pie chart illustrating the top 12 Gene Ontology (GO) functional categories for proteins under positive selection, as identified using InterProScan domain annotations. The largest proportion corresponds to structural molecule activity, followed by RNA binding, RNA-dependent RNA polymerase activity, viral capsid, ATP binding and DNA replication. See supplementary Table S35.

For the top 12 hits highlighted in Figure 56, further search was done in order to check for any overlap in ancestor charts in Quick GO function identifier page, it was found that some of GO functions terms meet at some points in ancestor charts which indicates a hierarchical relationship in their Gene Ontology. For example, Gene Ontology terms GO:0003723 "RNA binding" and GO:0005524 "ATP binding" are nested within the broader functional category of GO:0005488, which represents "organic cyclic compound binding". In this hierarchical arrangement, GO:0005488 serves as a parent term that includes a range of specific binding activities. Figures 57 and 58 are screenshots for the two ancestor charts, highlighting the GO term where both meet at.
Also "DNA replication", "DNA-templated transcription" and "proteolysis" overlap in ancestor charts where all three starts by "biological process" and meet at "metabolic process". While "RNA helicase activity" and "cysteine-type endopeptidase activity" both start with molecular function in their ancestor charts and meet at "catalytic activity" and "catalytic activity acting on protein" only.

The last overlapping noticed in 12 top hits ancestor charts was on the GO:0016888 for "endodeoxyribonuclease activity, producing 5'-phosphomonoesters" which has a double function starting GOs in "molecular function" and "biological process", so it overlaps with "catalytic activity" and "metabolic process" with the previous two examples.

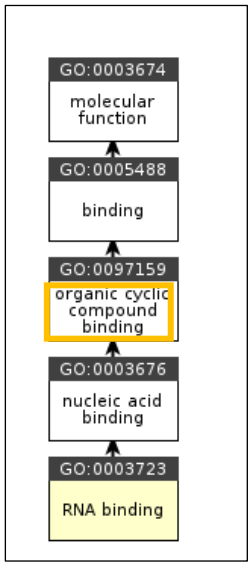

Figure 57: Screenshot of RNA binding GO Ancestor chart where the hierarchy meets with "ATP binding" at Figure 58 in the third level "organic cyclic compound binding".

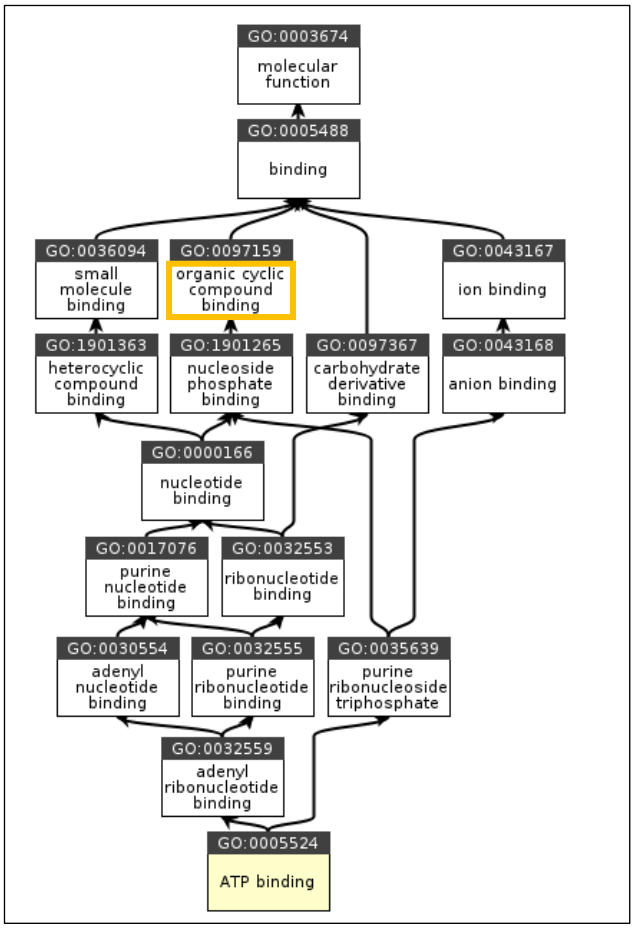

Figure 58: Screenshot of ATP binding Ancestor chart, the hierarchy has the same starting GO term at "molecular function" and meets at "organic cyclic compound binding".

- Pfam domain names and clans
    - Similarly, Pfam IDs from InterProscan output were collected and proceeded to Pfam-legacy to search for Pfam names. For the 300 proteins with positive selected sites, 187 Pfam domain names were identified with range of hits in InterProscan domain ranging from 1 to 28. The top Pfam ID was PF08283 and PF00799 both had 28 hits in InterProscan for "Geminivirus rep protein central domain" and "Geminivirus Rep catalytic domain", followed by PF00680 "Viral RNA-dependent RNA polymerase" with 22 hits. On the other hand, PF05750 for "Rubella capsid protein" and PF08456 for "Viral methyltransferase C-terminal" had only one hit in InterProscan domain. Figure 59 is a pie chart for top 12 Pfam names appeared in InterProscan domain for selected proteins.

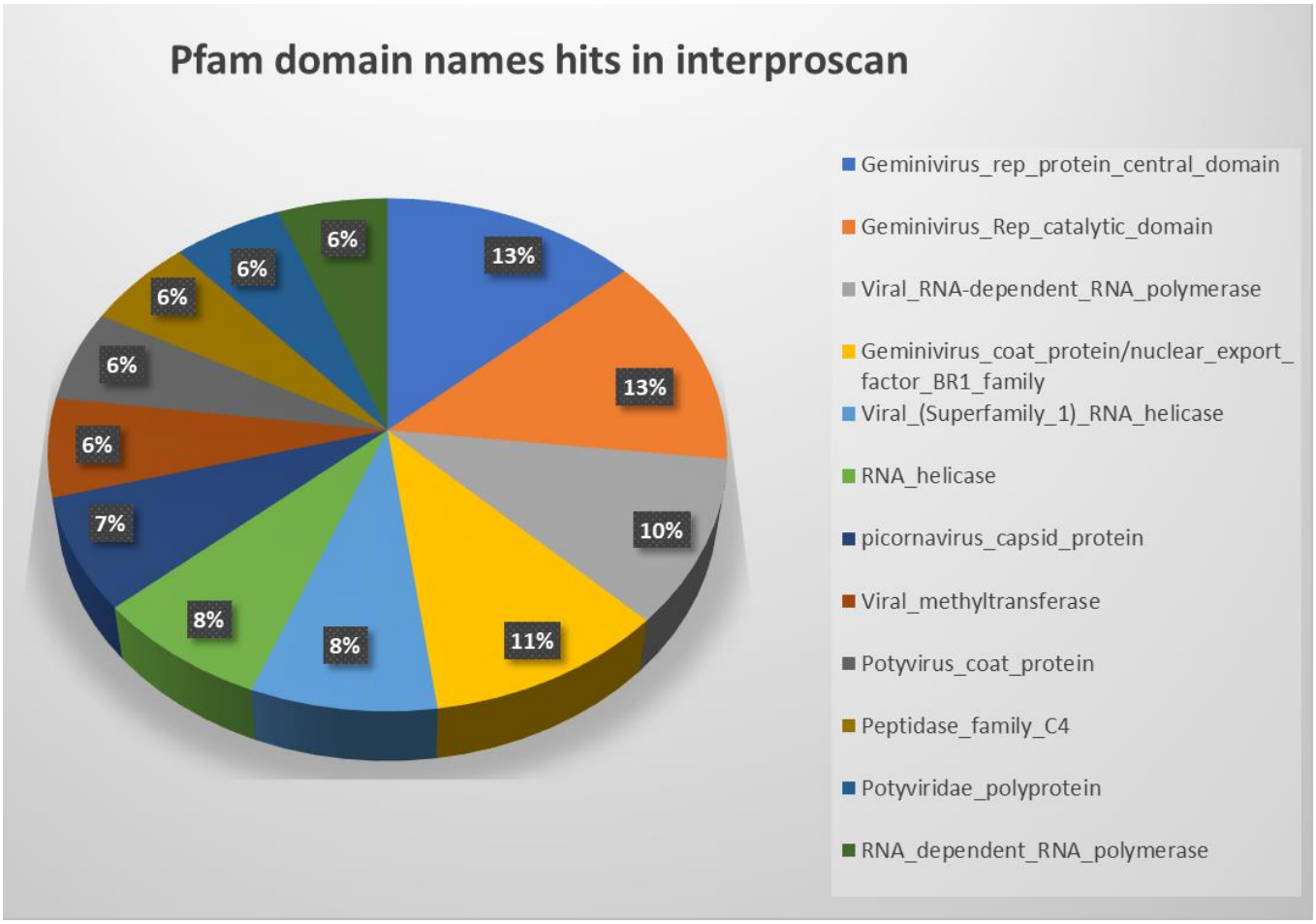

Figure 59: Pie chart showing the top Pfam domain names identified in proteins under positive selection based on InterProScan annotations. The two most frequent domains are Geminivirus rep protein central domain and Geminivirus Rep catalytic domain, followed by Geminivirus_coat_protein/nuclear_export_factor_BR1_family and Viral RNA-dependent RNA polymerase.

- Each Pfam ID was associated with a specific clan and corresponding clan members found using pfam-legacy. Out of the proteins with positive selected sites, which represented 187 distinct Pfam domain names, 91 were found to have no associated Pfam clan. For the remaining Pfam IDs, several shared clan names, resulting in a total of 33 unique clans for the 96 Pfam names and IDs. The top clan which appeared with 12 Pfam names was "Nucleoplasmin-like/VP (viral coat and capsid proteins) superfamily", followed by "RNA dependent RNA polymerase" which appeared with 9 Pfam names. While some clans appeared only with one Pfam ID as "Sialidase superfamily" and "Actin-like ATPase Superfamily". Figure 60 is a pie chart for top 10 Pfam clans in pfam-legacy.

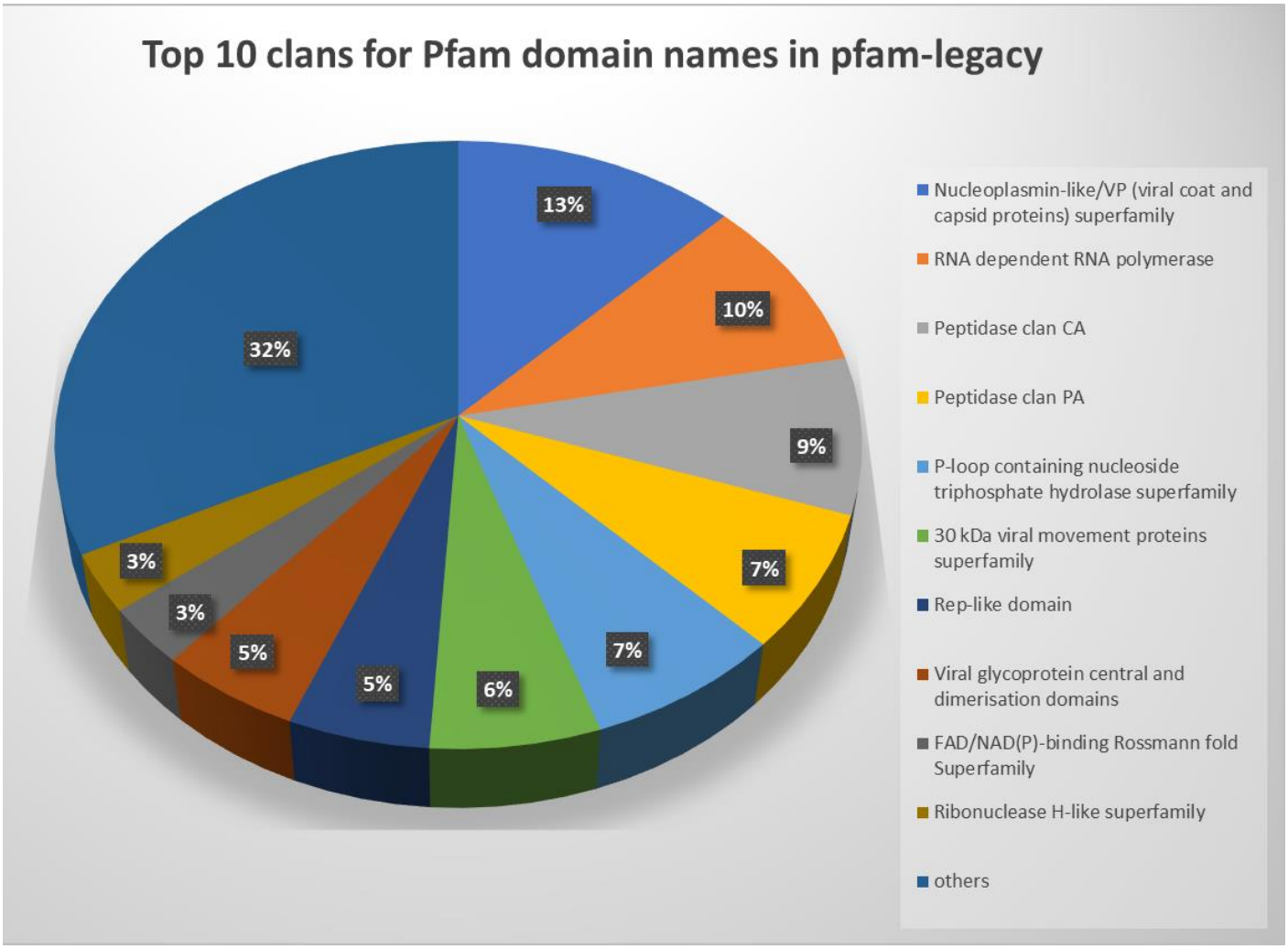

Figure 60: Pie chart showing the top 10 Pfam clan names linked to Pfam IDs through the Pfam-legacy database. The most represented clan is Nucleoplasmin-like/VP (viral coat and capsid proteins) superfamily, followed by RNA dependent RNA polymerase, Peptidase clan CA and Peptidase clan PA. See supplementary Table S36.

3.5.5 Domains and GO terms analysis with Omega calculations

Once Pfam domain names and GO terms associated with proteins with positive selection were identified, the average omega values for their corresponding selected sites were calculated. This was performed to understand the functional and evolutionary aspects of positive sites. Tables 21 and 22 are listing the top 12 hits for each Pfam domains and GO terms taken from InterProscan output with number of proteins and selected sites from SLR output associated to each domain ID and at the end average of Omega values were calculated.

Table 21: Number of positive selected sites with sitewise average Omega for each Pfam domain in the 12 top hits in InterProscan. The first column from the right is sitewise Omega average refers to the sitewise omega average for selected sites present.

Pfam Domain name	InterProscan Domain name	Number of hits in Inter-Proscan	Number of Selected Proteins	Number of Positive Sites	Sitewise Omega average
1. Geminivirus Rep catalytic domain	Geminivirus AL1 replication-associated protein, catalytic domain	28	28	103	9.8
2. Geminivirus rep protein central domain	Geminivirus AL1 replication-associated protein, central domain	28	28	103	9.8
3. Viral RNA-dependent RNA polymerase	RNA-directed RNA polymerase, C-terminal domain	22	22	117	6.7
4. Geminivirus coat protein/nuclear export factor BR1 family	Geminivirus AR1/BR1 coat protein	22	22	37	6.7
5. Viral (Superfamily 1) RNA helicase	(+) RNA virus helicase core domain	17	16	64	8.1
6. RNA helicase	Helicase, superfamily 3, single-stranded DNA/RNA virus	16	16	55	10.5
7. Picornavirus capsid protein	Picornavirus capsid	15	7	29	6.1
8. Viral methyltransferase	Alphavirus-like methyltransferase (MT) domain	13	13	57	7.2
9. Potyvirus coat protein	Potyvirus coat protein	12	12	75	6.7
10. Peptidase family C4	Potyvirus NIa protease (NIa-pro) domain	12	12	75	6.7

11. Potyviridae polyprotein	Polyprotein, Potyviridae	12	12	75	6.7
12. RNA dependent RNA polymerase	Tymovirus, RNA-dependent RNA polymerase	12	12	39	6

Table 22: Number of selected sites with site wise average Omega for each Gene Ontology term in 12 top hits in InterProscan. Similar to Table number 21, the first column from the right is sitewise Omega average refers to the sitewise omega average for selected sites present.

Gene Ontology ID	GO Function	Number of hits in Inter-Proscan	Number of Selected Proteins	Number of Positive Sites	Sitewise Omega Average
1. GO:0005198	Structural molecule activity	99	84	294	7.9
2. GO:0003723	RNA binding	82	64	304	7.4
3. GO:0003968	RNA-dependent RNA polymerase activity	78	60	281	6.6
4. GO:0019028	viral capsid	75	72	260	7.4
5. GO:0005524	ATP binding	59	46	270	8.4
6. GO:0006260	DNA replication	46	42	173	12.14
7. GO:0006351	DNA-templated transcription	43	42	172	6.6
8. GO:0006508	proteolysis	38	21	166	6.1
9. GO:0016888	endodeoxyribonuclease activity, producing 5'-phosphomonoesters	37	37	127	11.06
10. GO:0004197	cysteine-type endopeptidase activity	28	16	90	6.7
11. GO:0003724	RNA helicase activity	21	21	72	9.3
12. GO:0016032	viral process	18	15	45	5.2

3.5.6 GO host-virus relations

Selected proteins for the top GO terms listed in Table 22 were linked to their respective viruses and their hosts in alignment datasets studied. Percentages of hosts related to each GO term were calculated and listed in Table 23.

Table 23: Top hits GO functions with their linked viruses for proteins under positive selection and hosts percentages for each category.

GO Function	Number of Viral Alignments	Host Percentages				
		Mammalian	Plant	Avian	Fish	Insect
Structural molecule activity	84	33%	56%	7%	2%	2%
RNA binding	64	30%	60%	2%	2%	6%
RNA-dependent RNA polymerase activity	60	44%	40%	4%	6%	6%
Viral capsid	72	26%	68%	6%	0	0
ATP binding	46	40%	48%	4%	4%	4%
DNA replication	42	11%	85%	4%	0	0
DNA-templated transcription	42	30%	59%	7%	2%	2%
Proteolysis	21	38%	57%	5%	0	0
Endodeoxyribonuclease activity, producing 5'-phosphomonoesters	37	3%	97%	0	0	0
Cysteine-type endopeptidase activity	16	61%	33%	6%	0	0
RNA helicase activity	21	53%	33%	5%	0	9%
Viral process	15	34%	53%	13%	0	0

3.5.7 Selected proteins structure

- Blastall for selected proteins

  In order to determine tertiary structure for proteins with selected sites, 100% matches to PDB were needed. After applying the blastall command on proteins under positive selection, protein matches in the PDB were identified for 163 proteins. Out of these, only 38 exhibited a 100% identity score. For these 38 proteins, a total of 290 PDB structure IDs were recorded.

- Selected protein studied cases

  For a better understanding of proteins functions, it is important to study their structural details. For some selected proteins having one or more identified positive selected sites with distinct amino acid positions, a 100% match with a known protein solved structure in the Protein Data Bank (PDB) was chosen. MOE software was used to visualize protein's 3D structure. This step was performed to study the possibility of interactions between amino acids and for potential effect of selected sites and residues on structure.

  Out of 38 selected proteins with 100% matches some cases was selected for solved structure analysis:

  a. NP_757372.1 showed positive selection for Porcine parvovirus peptide sequence. 1K3v is a monomer solved structure for Porcine parvovirus, solved by using X-ray crystallography with 3.5 Å resolution (Simpson et al., 2002). Polymorphism analysis showed four different amino acids in the selected site within alignment dataset. Figure 61A, highlighted the three mutants Threonine (T), Methionine (M), and Serine (S), additional to the original amino acid Isoleucine (I) on four superposed chains, all on the same site 320. Also, Figures 61B and 61C represents molecular surface created on two different amino acids.

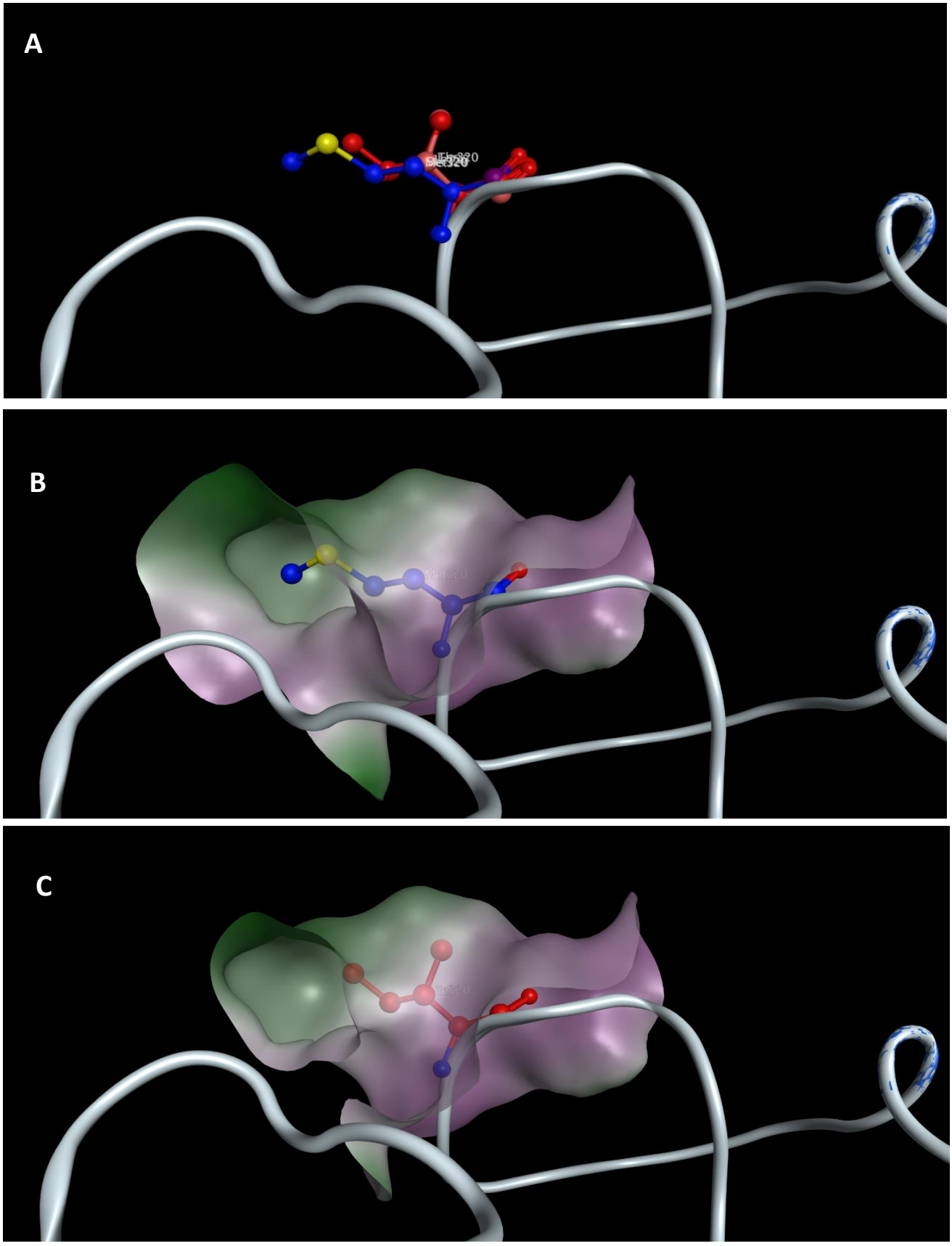

Figure 61: A. Isoleucine (I) residue on 1K3v solved structure, with three energetic minimized mutants Threonine (T), Methionine (M), and Serine (S). B. Molecular surface representation of the (M) residue created using MOE software. The surface is coloured based on lipophilicity, highlighting hydrophilic (purple), neutral (white), and lipophilic (green) regions. C. Molecular surface created on residue (I) using same features and same angel used to capture Figure 59 B.

b. YP_009513265.1 showed positive selection for Human metapneumovirus peptide sequence. 5FVD is a tetramer solved structure for Human metapneumovirus, solved by using by X-ray crystallography with 1.8 Å resolution (Renner et al., 2016). In SLR analysis one positive selected site was found in position 220, with two polymorphic amino acids; Tyrosine (Y) and Histidine (H). The residue found to be an internal residue, Figure 62 show the two amino acids on position 220 on aligned/superposed chains, no molecular surface study was performed.

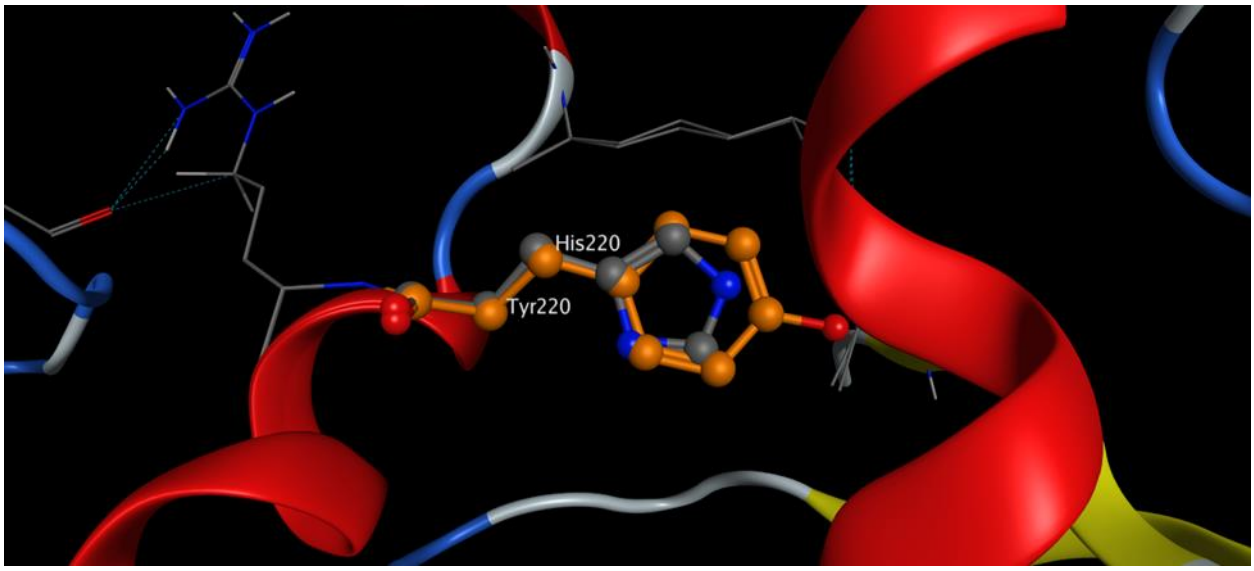

Figure 62: Shows residues Tyrosine (Y) and Histidine (H) on superposed chains on the same position 220 on 5FVD protein 3D structure which has a 100% match of Human metapneumovirus peptide sequence.

c. NP_040937.1 showed positive selection for Bovine viral diarrhea virus peptide sequence in multiple selected sites from SLR run. Two sites have been analysed in MOE for their high polymorphism number, both sites showed 9 polymorphic amino acids. According to this high number only four amino acids mutants were studied on superposed chains to avoid overlapping. According to the wide range of sequences years as collection dates where between 1990 to 2020, mutants were picked considering their diversity in sequence age, residue shape and occurrence within 70 sequences analysed. The solved structure PDB ID is 4JNT which is a dimer protein with identical chains, its structure was solved using X-ray crystallography with 4 Å resolution (Li et al., 2013). Figure 63A illustrates the selected

mutants Threonine (T), Serine (S), Histidine (H) additional to the original residue Glutamine (Q) on site number 944. Also, in Figures 63B and 63C molecular surface was created on two residues Glutamine (Q) and Histidine (H). Finally, the last surface was created on two selected sited 944 and 946 with mutant residues appearing on the earliest sequence collected on 1990 Serine (S), and the latest one collected on 2020 Glycine (G) as shown in Figure 63D.

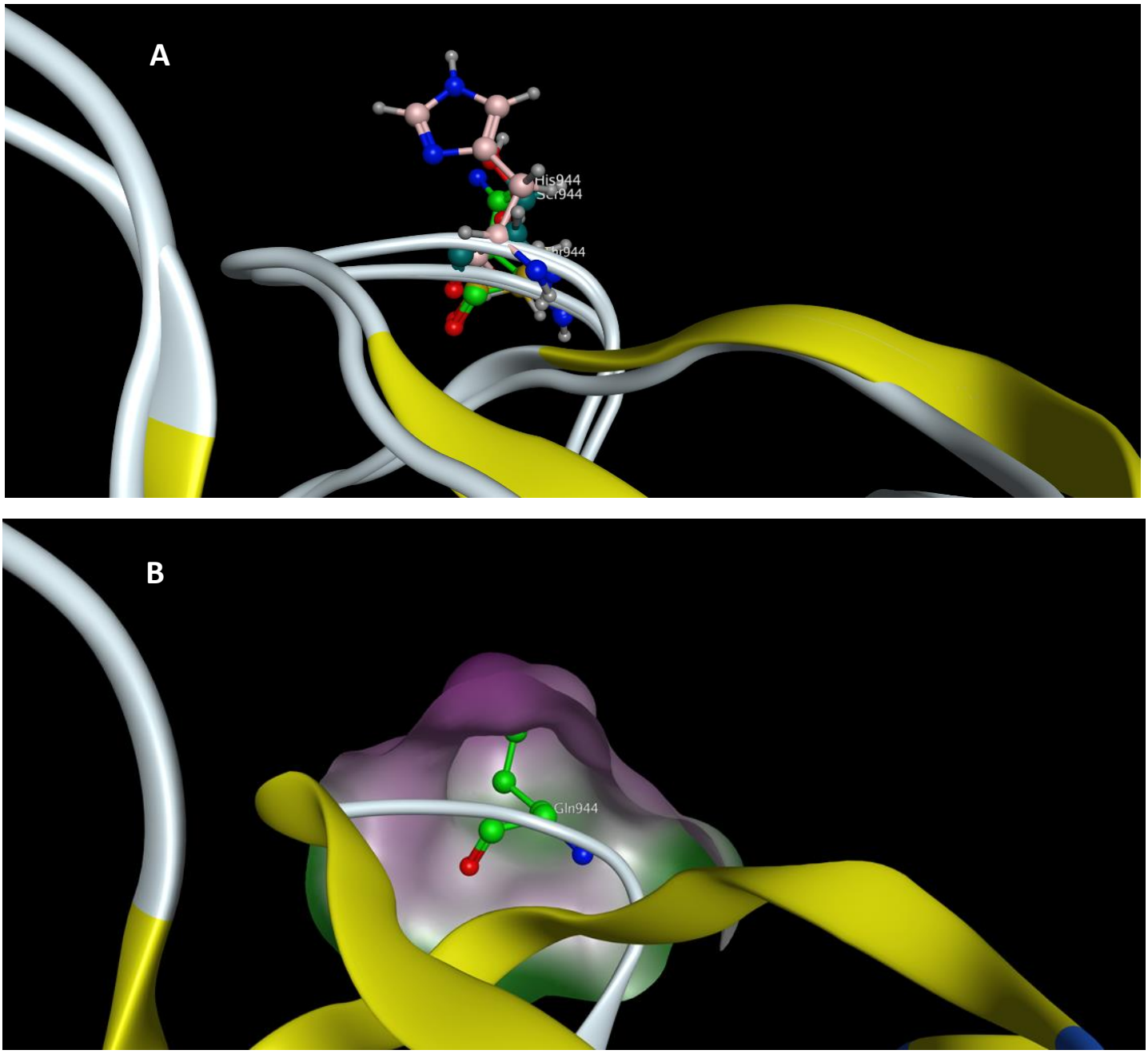

Figure 63: A. Glutamine (Q) on 4JNT solved structure, the 3D structure studied for Bovine viral diarrhea virus peptide sequence visualizing positive selected site with three energetic minimized mutants Threonine (T), Histidine (H), and Serine (S). B. Molecular surface created on the Gln (Q) residue only.

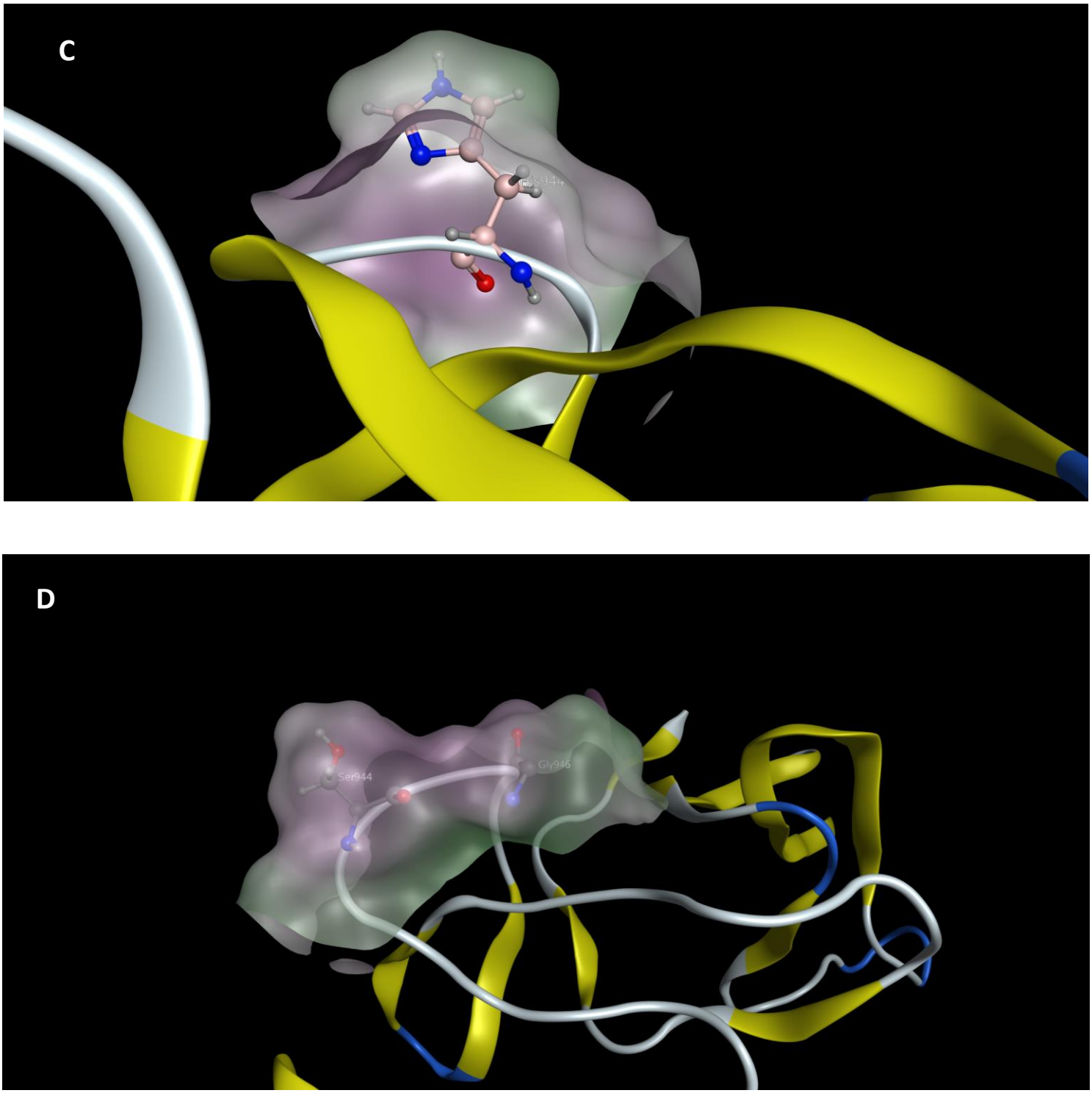

Figure 63: C. Molecular surface created on His (H) residue using same position, magnification, transparency, and other features as Figure 63B. D. Molecular surface created on more than one selected site with mutants from earliest and latest sequences.

d. NP_ 899212.1 showed positive selection for Machupo mammarenavirus peptide sequence. SLR showed two positive selected sites, Figure 64A represents site 170 with four different amino acids present in polymorphism analysis, which are Methionine (M), Lysine (K), Valine (V) and Arginine (R) highlighted on 6S9J, the solved structure of Machupo mammarenavirus contains 8 chains A to H solved using X-ray crystallography with resolution of 2.6 Å (Cohen-Dvashi et al., 2020). Figures 64B & 64C show molecular surfaces created on Methionine (M) and Valine (V) amino acids.

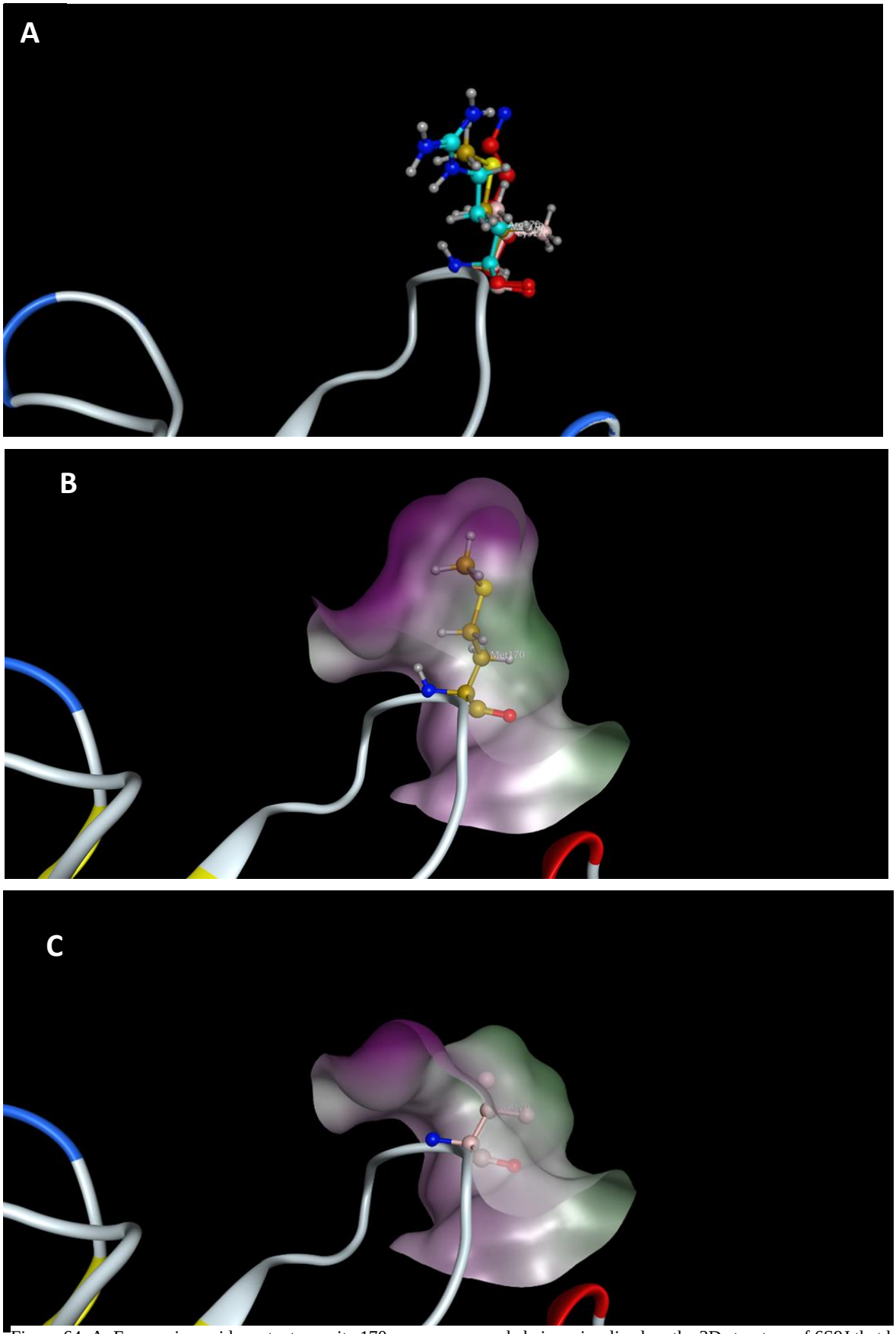

Figure 64: A. Four amino acids mutants on site 170 over superposed chains, visualized on the 3D structure of 6S9J that has 100% match of Machupo mammarenavirus peptide sequence . B. Molecular surface created on the Methionine (M) residue. C. Molecular surface created on Valine (V) residue using same position, magnification, transparency, and other features Figure 64B.

Furthermore, two antigenic epitope prediction platforms were used to search if the selected sites appeared within an epitope, the first selected site did not appear at any online prediction tool and the second one was found in “Antigenic” a component application within EMBOSS (Rice et al., 2000), but no evidence of being an antigenic site at System Biology laboratory of Chi Zhang (Yao et al., 2012).

Additionally, antigenic sites were highlighted in the structure and visualized by MOE with the two selected sites. Figure 65 illustrates two chains of the solved structure; chain B where both selected sites present and the adjacent chain A. also, both selected sites are highlighted with all mutants present, and pointed with arrows on the two surface residues.

Figure 65: 6S9j solved structure chains A and B only pointed. Two selected sites with different mutants’ amino acids on superposed chains, arrows showing selected site appeared within antigenic epitope.

Moreover, molecular surfaces were created using chain colour instead of lipophilic colour on both chains A and b with single amino acid on each selected site (not superposed chains), see Figure 66A**.** Also, residues found to be antigenic sites by EMBOSS antigenic online prediction

tool were highlighted through chain B sequence and a lipophilic molecular structure was created over them, see Figure 66B.

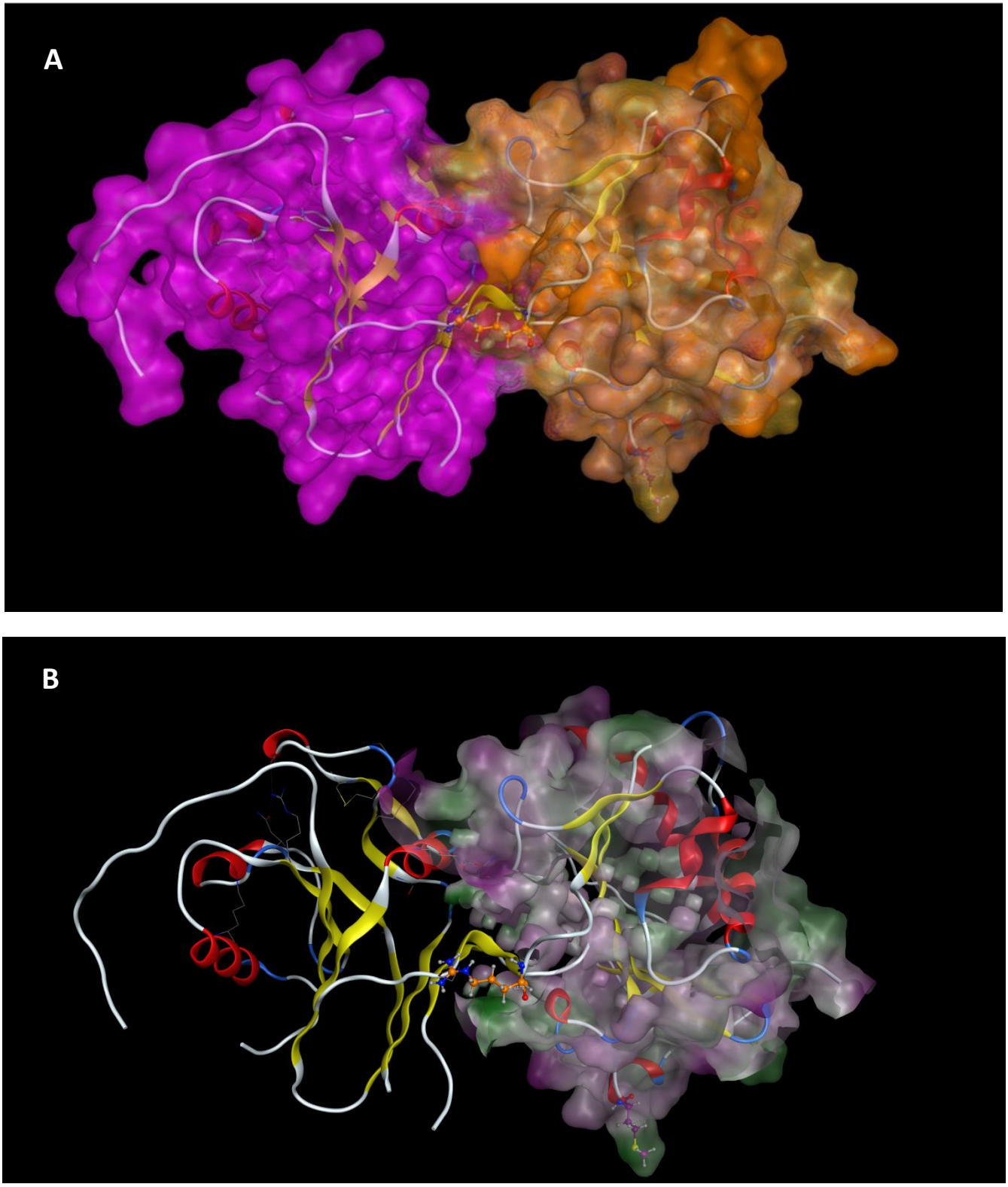

Figure 66: A. Molecular surface on whole chains using chain colour, showing both selected residues on chain B. B. lipophilic molecular surface created on only antigenic sites from EMBOSS antigenic prediction tool.

3.5.8 Overview of Positive selection results

Positive selection analysis was performed on a wider range of alignments, as R value filter was not considered, a total of 448 alignments were analysed for positive selection by SLR. 60% of alignments showed at least one positive selected site, as positive sites ranged from 1 to 203 site per alignment. Kappa and Omega values resulted from SLR run were plotted against length of alignments showing no linear correlation. Taxonomic distribution for alignments with positive selection showed Geminiviridae dominating with 41% among other families, while Sedoreoviridae and Spinareoviridae were prevalent among alignments with no positive sites. Polymorphism analysis showed 2 to 13 different amino acids for sequences in dataset per selected sites. Directional and diversifying selection were inferred through diversity scores and Omega values on phylogenetic trees for some alignments examples. Following, proteins found to be under positive selection were studied for domains. From 353 protein IDs, InterProscan identified 79 GO terms with structural molecular activity as the top hit annotation, then RNA binding and RNA dependent RNA polymerase scored as the second hit. Viral capsid was on the fourth hit for viral proteins under positive selection. Pfam domains showed strong taxonomic signals, particularly among Geminiviridae-related domains, while 91 domains lacked clans' associations. GO and Pfam data were linked to Omega values, to check if the highest GO and Pfam occurrence are more frequently targeted by positive selection, but it shows that selection does not always align with function frequency as the highest sitewise Omega for DNA replication and endodeoxyribonuclease activity GO terms. Furthermore, GO terms host linkage analyses revealed plant viruses were dominant in structural and replication functions, while mammalian viruses were more represented in enzymatic activities like helicases and endopeptidase. Finally, only 38 of positively selected proteins had 100% identity in BLAST search for PDB. Number of selected cases for different protein encoded functional annotations were visualised over MOE with mutants to assess structural and functional impact of positive selection.

Chapter 4: Discussion

The use of bioinformatics techniques is growing in the means of analysing large biological datasets, many of which are generated by various high-throughput technologies. Sequence alignments are widely used and provide a wealth of information. One of the vital methods used in research is sequence alignment as various data can be obtained e.g., regions, motifs, domains, and experimental results. Additionally evolutionary studies and homology modelling needs sequence alignments methods (Vihinen, 2020).

High quality alignments creation was the following goal after parsing GenBank records. The accuracy of all subsequent analysis depends on the quality of alignments dataset created, and their production required a combination of subject-specific expertise and computational methods. This procedure not only guaranteed the integrity of created datasets but also examine evolutionary parameters and patterns within alignments in more details.

Alignment datasets were created following specific criteria starting from the parsing step and ended by manually quality checking. At first, date collected sequences were the main criteria for GenBank record to be considered for studying "Tempo" in viral evolution. Also, for the purpose of "Mode" studying positive selection in coding regions of genomes. For both tempo and mode only coding sequences (CDS) of viral genomes were saved.

To maintain alignments quality, viral sequences with available reference genome were selected with high similarity matches to their reference genomes. Moreover, dataset has number of sequences in a controlled range to ensure phylogeny quality and to minimize thick sampling. With time gap considering within selected sequences, a minimal date range was set, accommodating both isochronous and heterochronous sequences to allow studying evolutionary changes over time.

Additionally, removal of sequences showing evidence of recombination reduces recombination effects on evolutionary studying which might lead to false positive or false negative results in some parameters. Finally, exclusion of some viruses; as well studied viruses e.g., Influenza, HIV, and SARS were removed from the datasets to avoid repetition and focus on viruses with less evolutionary studies present in literature, also viruses with larger genomes were removed from datasets to control time consumption computationally. Produced datasets from these

criteria in data collection and alignment, are expected to provide better understanding into the viral population evolution.

One of the major cons of the production of high-quality alignments is the significant reduction of dataset size, which affect the study of evolutionary parameters as taxonomic markers and decreasing spread of genetic diversity that might lead to bias in some taxonomical levels and segmented viruses' evolution.

In the discussion section, multiple analytical approaches used to explore viral evolution computationally will be discussed with examples from outcomes matched with provided in literature:

4.1 Recombination

It is important to know that this is not a dedicated project to study recombination on viruses, recombination primarily used as filter prior proceeding to molecular clock, substitution rate and selection pressure analysis.

Recombination was studied with Simplot Bootscan, to identify possible presence of recombinant sequences within viral datasets, showed that approximately 60% of alignments exhibited evidence of recombinant sequences.

Patino-Galindo et al. (2021) studied recombination in viral alignment datasets in a similar way done in this project but with some differences in methodology. Patino-Galindo perform recombination comparative study on 30 datasets contained genome sequences collected publicly in October 2017 for human viruses only. Recombination was analysed with more than one approach; the first was using conventional methods including RDP, Bootscan, Chimaera and 3seq, also a newly developed PH based, and LD method was used to detect recombination events. Authors discussed how their results agreed with the present literature in both cases with high and low presence of recombination with few exceptions.

Below, there are examples for recombination events findings in our data set and compared to what is found in existing literature.

4.1.1 Segmented viruses

- Concordant recombinant segmented viruses

  Recombination patterns were previously mentioned in Results section 3.2.4. Number of 6 segmented viruses showed concordant pattern where all species alignments exhibit

recombinant sequences. Following are three examples discussing this output and what was found in literature.

Pepper huasteco yellow vein virus (PHYVV)

The first example from the alignments date sets is Pepper huasteco yellow vein virus. Our recombination analysis detects recombinant sequences in PHYVV.

A single-stranded DNA virus primarily infects chiltepin plants in Mexico, a member of Geminiviridae family with two segmented genomes (segments A and B), revealed compelling evidence of recombination within populations of PHYVV. 6 recombinant sequences in DNA-A and 9 in DNA-B were found for PHYVV, and a recombination signal was found in at least one DNA component in 31% of the isolates (Rodelo-Urrego et al., 2015).

This example for Rodelo-Urrego supported project's findings for recombination events presence in PHYVV analysis.

Peanut stunt virus (PSV)

Peanut stunt virus is a positive stranded RNA virus belongs to Cucumovirus genus, under the family Bromoviridae. Similar to other Cucumovirus members, the genome is made up of three positive-polarity genomic RNAs called RNA1, RNA2, and RNA3.

PSV also showed a concordant pattern for recombination in two segments which undergo recombination analysis in the dataset RNA1 and RNA2 only. While in the literature RNA3 which encodes for two proteins has evidence of recombinant sequences presence, according to Kiss et al. (2008) the diverse phylogenetic origins and sequence comparisons of the two proteins encoded by RNA3 indicate the chance of recombination events presence during the evolutionary development of RNA3. My findings are partially supportive of the observation of Kiss et al, as we found recombinant sequences in RNA1 & RNA2 while RNA3 was not analysed as it did not pass the filtration threshold for number of sequences and collection date range.

Thottapalayam virus (TPMV)

Another example of concordant recombinant segmented virus in our dataset was Thottapalayam virus, a negative-sense, single-stranded RNA genome with three segments, is a member of the genus Hantavirus in the family Bunyaviridae showed no literature on recombination for TPMV while the studied dataset includes two segments

S and M where both showed recombination signals. Therefore, project's finding is the first illustration of recombination in TPMV.

- Concordant non-recombinant segmented viruses

  African horse sickness virus (AHSV)

  African horse sickness virus is a species of Oribivirus genus and Reoviridae family, AHSV can lead to fatal morbidity and mortality rates among horses. The non-enveloped double stranded RNA (dsRNA) virus contains 10 segments, the segmented genome codes for structural and non-structural proteins (Bremer et al., 1990; Zwart et al., 2015).

  AHSV appeared in the concordant pattern for non-recombinant viruses with five segments, where four of them showed recombinant sequences that can be removed and modify the alignment dataset. Our analysis was performed on five alignment data sets each for specific segment with number of sequences ranges between 21 to 30 for each alignment, Bootscan showed recombinant sequences in segments 1, 3, 4 and 6, while segment 10 had no recombination event.

  Ngoveni et al. (2019) studied recombination in complete genomes of 100 AHSVs. Recombination presence was studied in segments 1, 6,7 and 10. Genomic recombination events has been identified in segments 1 and 6, while segment 7 and 10 has only single cross over event.

  Our findings agreed with Ngoveni for recombination presence in segments 1 and 6 but disagreed with segment 10 which can be due timing of sequence collection and length of alignment scanned.

4.1.2 Taxonomical hierarchy level

- Concordant non-recombinant genus.

  Mastrevirus genus contributed in recombination analysis with 4 different species where all recorded no evidence of recombinant sequences presence, while the literature was contrary to the project's findings.

  Mastreviruses recombination studies in the literature showed that some recombination events were found. Mastreviruses show inter/intra-species recombination (Kanakala and

Kuria, 2018). Mastrevirus recombination analysis clearly shows the existence of many breaking points in the Rep and CP genes (Kraberger et al., 2013).

Different software used can explain to some extent the disagreement between literature and project's findings, also some species studied in the literature are different to the ones in our Mastrevirus dataset.

- Family concordance

  Polyomaviridae is an example of non-recombinant pattern family in the recombination analysis with 3 genera and 4 species which all didn't show any evidence of recombination.

  Polyomaviruses are double stranded DNA viruses; their genomes encode for regulatory proteins and structural proteins (VP1 and VP2). Species of Polyomaviridae were divided in number of genera according to their viral protein large tumour antigen (LTAg) connection. Carr discovered little to no evidence of recombination within LTag and discovered recombination hotspots at the margins of VP1 (Carr et al., 2017). Recombination events involving polyomavirus fragments and viruses from other viral families were also found (Moens et al., 2017).

  Project's findings in Polyomaviridae family recombination agreed to one study and disagreed with the second literature (Hughes and Friedman, 2000).

  Iflaviridae is a second example for family recombination, where all three species present showed recombination in a concordant recombinant pattern. The first virus is Deformed wing virus (DWV), which is a member of Iflavirus genus, an RNA virus which cause emerging infectious diseases (EIDs) in honeybees. There are now two primary genotypes of DWV: the previously common DWV-A and the recently identified and quickly spreading DWV-B. DWV-A and DWV -B distributed equally from 2008 to 2021, DWV-B likely displaced DWV-A due to its rapid global expansion since its first description in 2004. Recombination considered one of the causes leads to DWV-B dominancy over DWV-A in honeybees' populations (Paxton et al., 2022).

  Li et al. (2016) proved presence of recombination in Sacbrood virus, the second species in Iflaviridae family, a member of Iflavirus genus. The study revealed presence of recombinant sequences in Vietnam (Viet1) and Korean (Kor19) isolates.

The last species in Iflaviridae family with no rank in genus was La Jolla virus, which showed no literature on recombination study while in our dataset multiple recombinant sequences were found in the viral alignment.

As above examples on recombination showed agreeing and disagreeing with findings presented in the literature, this can be explained due to using different software with different algorithms, difference in alignment lengths for the analysed sequences, number of sequences in each dataset, sampling time and quality of alignments.

4.1.3 Recombination and hosts

Referring to Tables 10A and 10B, concordant genera and families were listed with their corresponding species for both concordant non-recombinant and concordant recombinant patterns with corresponding hosts for all species. Number of host types included mammalian, plant, fish, and insect, with a majority in plant hosts around 70% of all species. Primary, this can be for the reason of high number of plant viral alignments in the recombination study dataset before attempts of expanding the filtering restrictions. Also, this may contribute to an understanding that plant viruses are prone to recombination. Chare and Holmes (2006) performed a phylogenetic survey on a number of RNA plant viruses to study frequency of recombination. The study revealed that recombination is common in positive-sense RNA plant viruses, with more than one third of genome alignments showing evidence of recombination. They concluded that plant RNA viruses exhibit higher recombination rates compared to other RNA viruses, which may contribute to their rapid evolutionary potential. Factors as frequent co-infections and large populations in plants can aid in early recombination occurrence during viral infections. Recombinant genomes accumulation across different time points was detected in a study conducted both natural vector transmission and artificial inoculation on two Begomoviruses (Urbino et al., 2013).

4.1.4 Summary of recombination

Although recombination performed primarily on viral alignment dataset as a filtration process to ensure alignments quality, it revealed interesting taxonomic patterns to be concluded. The majority of recombinant alignments were associated with Potyviridae and Geminiviridae at family level, while the non-recombinant alignments with Reoviridae and

Geminiviridae. Which can indicate that certain families can be prone to recombination more than others. Followed by analysing segmented viruses where around 57% showed recombinant concordant pattern across viral segments and the remain 43% showed discordant pattern. This can suggest that although number of segmented viruses maintain constant recombination across segments still a large portion does not. Furthermore, across taxonomical hierarchy levels, only 6 out of 29 families and 14 out of 36 genera showed full concordance with majority showed discordance, this suggests that within the datasets analysed recombination cannot be considered as a taxonomical marker.

4.2 Molecular clock

4.2.1 Temporal signal of viral datasets

Measuring viruses temporal signal or clocklikeness provides understandings of how regularly underlying molecular evolutionary changes occur. The term clocklikeness reflects how much substitution rates in genome sequences follows a molecular clock model, has been found in number of articles. Clocklikeness is measured by measuring correlation between genetic divergence and temporal divergence.

According to Rambaut et al. (2016), It is advisable to verify that sequences under examination have enough "temporal signal" for accurate estimation before employing a molecular clock model to construct a time-scaled tree from heterochronous sequences. This is important before proceeding to any Bayesian analysis, since the reliable evolutionary rate estimation greater than zero and molecular models used are processed statistically according to this estimation. The Bayesian software permits inference to continue even in cases where the studied alignments have low or no temporal signal, so the RIRO conditions (rubbish in, rubbish out) error is easy to commit if Bayesian software is used without a prior check for temporal signal.

In this project TempEst software was used to study temporal signal within viral datasets, showed that 53% of alignments has high sufficient temporal signal with R value >= 0.5.

Molecular clock has a patchy and scattered literature compared to recombination where large number of studies performed on virus genomes, PubMed advanced search for viruses

and temporal signal ends up with 29 hits, below are examples of temporal signals found in the literature matches viruses from our dataset.

Furthermore, the small number of viruses which appears in our dataset with a study of clocklikeness is the first point on which this thesis makes a novel contribution in viral evolution, due to the limited number of published studies including temporal signal studying in viruses. Additionally, removal of viral alignments with low temporal signal improves molecular clock estimates quality but may cause a loss of some potentially important data, especially when dealing with understudied viruses. These low R values alignment datasets may hold information for future pandemics. Duchene et al. (2020a) studied evolutionary rates for several SARS-CoV-2 genomes at pandemics early stage, the authors demonstrated the essential steps to understand the importance and limitations of early data collected in outbreaks. Duchene mentioned that early SARS-CoV-2 genome data, despite having initially low temporal signal, helped in understanding the virus evolutionary rate and time of origin once phylodynamic threshold was reached.

Note: Temporal signal measured using correlation coefficient R values while most of literature used R^2, and some did not specify values.

- Infectious hematopoietic necrosis virus (IHNV)

  Infectious hematopoietic necrosis virus, a member of Novirhabdovirus genus and Rhabdoviridae family, IHNV appeared within a concordant pattern for Rhabdoviridae family with high correlation coefficient value of R^2= 0.53. According to Abbadi et al. (2021) correlation between sample dates and genetic distance over time was studied on Italian IHNV isolates using TempEst to detect temporal signal. Therefore, R^2= 0.82 which indicates ability of the virus to evolve due to pertinent temporal signal.

  This example agreed with our findings for IHNV having strong temporal signal.

- Potato virus A (PVA)

  Second example studied and found in the literature was Potato virus A, PVA showed a discordant pattern twice when both genera and family levels were analysed with Potyvirus genus and Potyviridae family levels, correlation coefficient recorded low value of R^2= 0.1. According to Fuentes et al. (2021); 47 PVA isolates were collected and by using TempEst, no temporal signal was found in dated non-recombinant PVA

alignments. Also, in this example findings agreed of PVA having low temporal signal as in alignment datasets.

- Rice yellow mottle virus (RYMV)

  Rice yellow mottle virus has been considered as a main threat to rice cultivation in Africa, since it is found in sub-Saharan Africa rice producing countries. RYMV is a single-stranded positive-sense RNA virus, a member of the Sobemovirus genus and the Solemoviridae family which has a concordant pattern with two genera and species scoring correlation coefficient value of R^2= 0.28 in our data sets. Issaka et al. (2021) studied RYMV spread among Niger valley with some countries from West and Central Africa, temporal signal found to be weak on sequences of retrieved isolates and root to tip analysis show R values <0.2. Similar to the two previous examples here our findings agreed with literature.

- Louping ill virus (LIV)

  Louping ill virus is a flavivirus from Flaviviridae family which is similar to tick-borne flavivirus in the British Isles (Jeffries et al., 2014). This positive-strand RNA vector-borne virus appeared in our dataset with the Flaviviridae family discordant pattern with R^2= 0.29. Clark et al. (2020) studied clock rate on a dataset of 26 LIV genomes and the correlation coefficient R^2= 0.35 indicated weak temporal signal. Project's findings match Clark's for LIV having low temporal signal in studied alignments.

- Beak and feather disease virus (BFDV)

  Beak and feather disease virus is a circular single-stranded DNA virus, a member of Circovirus genus and Circoviridae family. At least two proteins are encoded by the bidirectional transcription of the BFDV genome; capsid protein (CP) that is expressed from the complementary strand and a replication related protein (Rep) that is expressed from the virion strand. BFDV appeared in our dataset in the discordant pattern of Circovirus genus with high correlation coefficient of 0.8.

  Harkins et al. (2014) collected the BFDV genome sequenced data publicly available. A total number of 184 BFDV genomes studied, three data sets were aligned; a. (RF) recombinant free set which has the 184 genomes after removal of recombinant

sequences, b. (Rep) set which contains the RF dataset with the Rep gene out from the 184 genomes, c. (CP) set includes the CP alignments that encodes for CP out from the 184 genomes. Low clock rates recorded on BFDV three data sets using root to tip divergence, correlation coefficient ranged between 0.18 for the RF dataset, and 0.17 for the Rep dataset.

Unlike previous examples BFDV clocklikeness showed different findings in the project's analysis than Harkins, although both datasets are free from recombinant sequences, this could be from sample collection time difference and parts of genome examined.

- Omsk hemorrhagic fever virus (OHFV)

  Omsk hemorrhagic fever virus is one of Flavivirus genus members within Flaviviridae family, this tick-borne virus has a single-stranded, positive-sense RNA genome.

  Bondaryuk et al. (2023) studied OHFV evolution starting by studying molecular clock in 43 ORF date collected sequences using TempEst, clocklike behaviour was estimated as $R^2 = 0.42$ and 0.62. While in our dataset $R^2 = 0.25$ which considered low value for temporal signal, this difference could be due to number of sequences and time of sampling, the two values mentioned in the article are for ORFhet and ORFhet+iso that implies to time of sample collection, additionally we studied sequences of complete genome which includes ORFs and Envelope genes.

- Salivirus A (SalVA)

  Salivirus A, a member of Salivirus genus in Picornaviridae family. This non-enveloped positive sense single stranded RNA genome virus related to acute gastroenteritis known to infect human and chimpanzee causing acute diarrhea.

  In our dataset Salivirus A appeared with high R value of 0.6, and an example from literature agreed to this finding, Angeletti et al. (2021) studied evolution in 81 sequences of SalVA specifically in the VP1 region, starting by testing clocklikeness using TempEst, R showed the value of 0.56 in the root to tip divergence plot.

4.2.2 Family taxonomy distribution

According to Figures 30 & 31 in Results section, two pie charts represented family level taxonomic distribution for low and high "R" values, the highest occurrence families in both datasets were similar: Geminiviridae and Reoviridae with high and low correlation coefficient values. Geminiviridae, a family of viruses with plant hosts, and Reoviridae, has varied range of hosts as plants, animals, and humans. Reoviridae family is highly diverse among other dsRNA viruses, species members of this family have wide range of hosts including plants, mammals, fish, birds, insects and fungi (Quito-Avila et al., 2011).

The consistent presence of these two families across datasets with different temporal signals, also with a discordant pattern in concordance analysis previously studied can suggest evolutionary rates differences, variation in genome structure or even sampling density across more than one species within these families.

There are almost no publications discussing temporal signal or clocklikeness in these two families and other families with high and low temporal signals as: Picornaviridae and Nonoviridae in the present literature. Additionally, no literature reviewed temporal signal in families or genera, which missed the chance to understand concordance or discordance in different taxonomical levels. This lack of published articles can affect temporal and evolutionary understanding in top hit families and their viruses, according to this the project can be considered as the first attempt for a comprehensive study for temporal signal across viral alignments.

As previously mentioned only 53% of viral alignments had high R values in TempEst, which means that only half alignments with a strong temporal signal passed the filter and included for Bayesian analysis in BEAST. This exclusion of datasets with weaker temporal signals, chop large number of viral data, introduces possible biases but guarantees accurate evolutionary rate and divergence time estimates.

4.2.3 Temporal signal and hosts

Referring to Tables 13A and 13B, all concordant genera and families with their corresponding species were listed for both high and low R values, showed that the majority of these species are related to mammalian hosts especially in genus level. This can contribute to disease outbreaks understanding, with the suggestion of diverse evolution for mammalian

viruses. Harvey and Holmes (2022) addressed a question in their review: can viral evolution be shaped by host evolution? And reply on this by saying, although researchers studied how new viruses' lineages emerge, there is still less known about their rates and how virus lifecycle behave. They added that the evolutionary change in animals had an impact on viruses they harbour. Similar studies done as for bacteria in mammals linked to mammalian evolution.

As previously mentioned, lack of studies on temporal signals for specific taxonomy levels and viral hosts generally can impact future outbreaks predicting in genera and families with low or high temporal signals.

4.2.4 Summary of molecular clock

Concluding temporal signal analysis on viral datasets; the analysis revealed that 59% (percentage increase after easing attempts) of total alignments showed sufficient temporal signal and proceeded to further analysis for tempo and mode studying, with both Reoviridae and Geminiviridae being the most represented families across alignments regardless of R value. Among segmented viruses 47 out of 79 alignments showed concordant temporal signal patterns across segments, while 32 were discordant. For taxonomical hierarchy levels, only 5 out of 29 families and 16 out of 49 genera exhibited concordance, with the majority showing discordance. This wide variation suggests that temporal signal does not align with taxonomic classification and cannot be considered as a taxonomical marker within studied datasets.

4.3 Substitution Rate

Substitution rate search is similar to recombination in the number of articles, PubMed search provided with 395 number of hits. For the large number of substitution rate studies present, a search was done for viruses according to evolution speed category and how previous studies agreed or disagreed with our findings.

4.3.1 Evolution speed categories

For speed categories ranges please refer to Table 14 in Results section. Substitution rate is represented in an order of a magnitude with the unit substitution/ site/ year (s/s/y). This rate reflects the average number of nucleotide changes that occur at each site in the genome annually. For example, a rate of 1x 10^{-3} s/s/y means approximately 1 substitution per 1000 sites per year. Table 14 represented the five evolution categories showing that the fastest is evolving 10,000 times faster than the slowest since each category in 10 times faster in order of a magnitude. Below are examples for viruses covering the speed categories studied from our alignments.

- Very slow evolution

  Puumala orthohantavirus (PUUV)

  One species of the genus Orthohantaviruses in the family Hantaviridae is Puumala orthohantavirus. Hantaviruses are a negative sense enveloped single stranded RNA viruses, the genome of hantaviruses is divided into three segments: S for small, M for medium, and L for large (Reil et al., 2017).

  Puumala orthohantavirus recorded as the slowest virus in the means of lowest meanRate value in substitution rate analysis studied on M segment, with R value = 0.73. The meanRate was 9.96 x 10^{-6} s/s/y, however the 95% HPD is wide (6.62 x 10^{-10} to 2.9 x 10^{-5} s/s/y).

  Ramsden et al. (2008) calculated the nucleotide substitution rate for 3 hantaviruses species datasets, Puumala recorded a meanRate value of 6.22 x 10^{-4} s/s/y on a Bayesian analysis run using relaxed normal molecular clock with (95% HPD = 1.5 X 10^{-4} to 1.06 x 10^{-3} s/s/y). Anyway, on the same review it was mentioned that previously hantaviruses evolutionary rate has been estimated between 2 x 10^{-6} and 3 x 10^{-7} s/s/y. (Hughes and Friedman, 2000; Sironen et al., 2001) studied evolution on S segment only from PUUV, with nucleotide substitution rate 2.2 x 10^{-6} to 0.7 x 10^{-7} s/s/y.

  Literature showed that project's output agreed to two of them and disagreed on the latest one, this discrepancy could be due to different genes studied and time of sampling included.

  It is important to add that PUUV is a zoonotic virus primarily maintained in rodent reservoirs, particularly the bank vole (Myodes glarelous), and is only sporadically transmitted to humans. This may contribute to lower substitution rates compared to

viruses replicate and transmit within human population. Asymptomatic infections in rodent hosts can lead to reduced viral replication frequency and limit mutation opportunities. Moreover, the lack of strong immune driven selection pressure in theses natural reservoir hosts may limit viral diversity, which can lead to slower evolution than observed in human adapted viruses (Strandin et al., 2020; Weber de Melo et al., 2015).

- Slow evolution

  Coxsackievirus B3

  Coxsackievirus B3 is a single stranded RNA virus, a strain of Enterovirus B species, a member of Enterovirus genus and Picornaviridae family. Coxsackievirus B3 considered a public health concern being an active pathogen in pancreatitis, myocarditis, aseptic meningitis, and hand, foot, and mouth disease (HFMD) (Fairweather et al., 2012; Han et al., 2019).

  In our substitution rate analysis Coxsackievirus B3 dataset categorized in the slow evolving viruses with a meanRate value of 4.59×10^{-5} s/s/y and (95% HPD interval of 5.87×10^{-8} to 1.32×10^{-4} s/s/y), R value = 0.56.

  Yang et al. (2022) studied evolutionary history of Coxsackievirus B3 analysing sequences for P1 region using exponential molecular clock and GTR model, the substitution of CVB3 nucleotides was 4.82×10^{-3} s/s/y and (95% HPD= 3.51×10^{-3} to 6.05×10^{-3} s/s/y).

  Substitution rate from this projects' work is different than literature and this can be for the different genome region analysed and clock type, lognormal relaxed clock was used, and this also explain the wide 95% HPD interval than Bayesian analysis using lognormal relaxed clocks.

- Moderate evolution

  Human respirovirus 1 (HRV1)

  An RNA virus, a causative agent for common cold is Human respirovirus 1, a member of Respirovirus genus and Paramyxoviridae family. R value = 0.94. Once Bayesian phylogenetic analysis was applied on the alignment data set, the meanRate recorded 4.52 $\times 10^{-4}$ s/s/y and (95% HPD interval ranged from 3.2×10^{-4} to 5.87×10^{-4} s/s/y). According to the evolution speed categories, HRV1 has a moderate speed.

Takahashi et al. (2023) collected 66 HRV1 strains from several countries to study the evolution of fusion protein F gene, starting with estimating temporal signal for the sequences in the dataset, R^2= 0.87 which indicates that the 66 strains are suitable to analyse molecular clock. Using BEAST, rate of evolution was calculated for all HRV1 strains to be 8.5×10^{-4} s/s/y and (95% HPD is from 7.01×10^{-4} to 1.0×10^{-3} s/s/y). Findings of Takahashi's study agreed to the project's finding.

- Fast evolution

  Rotavirus C (RVC)

  A member of the Reoviridae family, Rotavirus is the primary cause of acute gastroenteritis in both humans and animals.

  Rotavirus has nine species known as A to D and F to J. Rotavirus A, B, and C are the most prevalent types that infect both humans and animals (Ferrari et al., 2022).

  In our analysed dataset RVC appeared in the concordant pattern of segmented viruses and concordant pattern for Rotavirus genus, where six segments fall in the fast evolution speed, the meanRate started from 1.01×10^{-3} for the (VP1) gene, 1.4×10^{-3} for the (VP6) gene, 2.4×10^{-3} for (VP3) gene, 2.5×10^{-3} for the (VP2) gene, 4.01×10^{-3} for (NSP4) gene and 5.89×10^{-3} for the (NSP2) gene, and (95% HPD intervals from 1.01×10^{-3} to 6.14×10^{-3} s/s/y).

  Joshi et al. (2023) studied RVC eleven segments, different genes nucleotide sequences included information on host, the country of origin, and collection date were collected from GenBank. Starting by measuring temporal signal for each gene dataset, root to tip regression showed no temporal signal for two genes and where excluded, for the nine remaining genes which showed weak to moderate temporal signal proceeded to Bayesian phylogenetic analysis. Substitution rate for RVC nine genes ranged from 6.5×10^{-4} to 1.19×10^{-3} s/s/y.

- Very fast evolution

A. GB virus C (GBVC)

GB virus C, also referred to as hepatitis G virus (HGV), belongs to the Flaviviridae family and Pegivirus genus (Leary et al., 1996). The genome of this single-strand RNA virus has a positive sense strand. The fact that GBVC is broadly spread in the healthy

human population and causes a sustained, asymptomatic infection in hosts makes it remarkable.

Romano et al. (2008) collected all GBVC publicly available dated sequences and separated them in four aligned sets according to viral proteins: E1, E2, 5′-UTR and NS5b. Nucleotide substitutions measured throughout Bayesian analysis recorded the highest in E2 region and the lowest in the 5′-UTR, as meanRate ranged between 2.2×10^{-2} and 4×10^{-3} s/s/y.

In our analysed data set for GBVC, the nucleotide substitution showed the highest value among other viral dataset with meanRate 3.76 x 10^{-2} s/s/y, with R value = 0.92 and categorized in the very fast evolving viruses. Here, our analysis agreed with present literature.

B. Alternanthera yellow vein virus (AYVV)

Alternanthera yellow vein virus, a member of Begomovirus genus and Geminiviridae family. This single stranded DNA genome virus, known to infect wide range of plants and fields in a number of South Asian countries as Pakistan and India (Shafiq et al., 2023).

Nawaz-Ul-Rehman et al. (2022) performed Bayesian evolutionary analysis on datasets for AYVV isolates within coat protein CP gene, nucleotide substitution rate was 4.75 x 10^{-3} s/s/y, according to the author this value is higher than other Begomoviruses substitution rates in CP genes.

According to our data sets, findings was different from literature as AYVV was classified in the very fast evolving viruses with R value = 0.9, and meanRate value was 1.23 x 10^{-2} when complete genome was studied by BEAST.

4.3.2 Substitution rate and taxonomy hierarchy

- Family taxonomy level

Secoviridae family is the plant infecting member of Picornavirales order. The Comovirinae subfamily, which also includes the genera Fabavirus, Nepovirus, and Comovirus, includes the majority of Secovirid species. Viral species of Secoviridae family have positive-sense single stranded RNA genomes (Sanfacon et al., 2009).

Secoviridae family was included in the substitution rate concordance analysis due to its discordant pattern in evolution speed. Within Fabavirus genus, two species recorded different meanRate values. Prunus virus F, virus with two segments; RNA2 polyprotein 2 gene recorded 1.05 x 10^{-3} s/s/y, and RNA1 polyprotein 1 gene has a meanRate value of 5.5 x 10^{-4} s/s/y. The second species was Broad bean wilt virus 2 of RNA1 gene recorded 9.68 x 10^{-5} s/s/y. Additionally, Grapevine fanleaf virus belongs to Secoviridae family in Nepovirus genus has a meanRate value of 0.014 s/s/y, and Strawberry mottle virus under Stramovirus genus showed a meanRate value of 0.0134 s/s/y. According to our substitution rates categories, species from Secoviridae family are evolving under slow, moderate, fast and very fast evolution speeds.

Thompson et al. (2014) studied evolutionary traits of these Secoviridae viruses by downloading 27 Secovirids nucleotide sequences publicly available in GenBank. BEAST run to the full conversion of parameters, substitution rate of CP "coat protein" nucleotides were calculated, and meanRate values ranged from 9.29 x 10^{-3} to 2.74 x 10^{-3} s/s/y.

Project's findings on substitution rate for species belongs to Secoviridae family disagreed with Thompsons' review, this can be for our studying of whole genome evolution, different species involved in the study and time of sampling.

- Genus taxonomy level

  Allexivirus genus belongs to Alphaflexiviridae family, have positive sense single stranded RNA genomes, including garlic viruses A, B, C, D, E, X as their common species.

  Garlic virus was included in viral alignments datasets with two different species; Garlic virus B and Garlic virus D, in substitution rate analysis both falls in the very fast evolution category with meanRates values 0.0221 and 0.0249 respectively. This fast evolution can suggest an adaptive mechanism within Allexivirus genus or even at a higher level with Alphaflexiviridae family, both two taxonomical levels might have features promotes evolution to escape host recognition and moreover faster spreading among garlic harvests.

4.3.3 Segmented viruses' substitution rate

Segmented viruses in substitution rate analysis had exhibits different behaviour among segments in number of viruses. This could be due to a different gene function where each gene is under different selective pressure, that can behave differently with host immune system. Also, mutation rate can be affected with genome segment size. So, when genome parts evolve at different speed, this can contribute to a fact that segmentation can be part of adaptation as virus ability to adapt is stronger when segments evolve in a different rate.
Some segmented viruses have the ability to exchange genome segments during co-infection, in a process called reassortment. This feature allows generation of genetic diversity, that can lead to evolutionary behaviours such as immune evasion and adaptation to new hosts. Genetic compatibility between segments is one of the keys to successful reassortment (McDonald et al., 2016).

- Segmented discordant pattern

  One example of a segmented virus from our dataset is Tilapia Lake virus, the single member of Tilapinevirus genus and Amnoonviridae family. TiLV is a negative sense, single stranded RNA virus with a segmented genome of segments 1 to 10.

  Thawornwattana et al. (2021) performed Bayesian analysis on genomic sequences collected from NCBI from 2011 to 2019, and from additional segments that were amplified by RT-PCR for analysis. Temporal signal was sufficient for all data sets. Bayesian analysis performed using BEAST with strict clock model. Results showed evolution speed between 1.81 to 3.47×10^{-3} s/s/y for all segments 1 to 10. According to the project's substitution rates categories, Thawornwattana's finding for TiLV segments fall in fast evolving viruses category.

  On contrary, our substitution analysis run on all 10 segments of TiLV resulting in a different evolution speed as follows; 1.20×10^{-3} to 3.21×10^{-3} s/s/y for segments 1-3, 5, 7-10, while segments 4 and 6 has speed of 9.09×10^{-4} and 4.21×10^{-4} s/s/y respectively. In our findings TiLV evolve in the fast-evolving category except two segments had a moderate evolving speed.

  This difference in output can be for more than one reason, the top one is type of molecular clock used in both analysis, type and time of sequences sampling too.

- Segmented concordant pattern

  As previously mentioned in section 4.3.1, the genus Rotavirus is a member of Reoviridae family, a double stranded RNA genome virus with 11 segments. Rotaviruses A, B, C and H were found to infect humans. RVB start to be enteric virus in the period between 1982 to 1987 when severe epidemics of diarrhoea detected in China (Lahon et al., 2013).

  Human rotavirus B appeared in a concordant pattern in our dataset with 10 segments encoding for both structural and non-structural proteins. MeanRate values were recorded as follows: 1.15 x 10^{-3} s/s/y for (NSP2) gene, 1.57 x 10^{-3} for (NSP5), 1.62 x 10^{-3} for (VP7), 1.74 x 10^{-3} for (VP3), 1.78 x 10^{-3} for (VP2), 1.87 x 10^{-3} for (VP6), 2.02 x 10^{-3} for (VP1), 2.32 x 10^{-3} for (NSP3), 2.6 x 10^{-3} for (VP4). And (95% HPD intervals from 6.9 x 10^{-4} to 3.38 x 10^{-3} s/s/y). R values for the 10 segments recorded between 0.85 to 0.98 indicating strong temporal signal.

  Lahon et al. (2012) studied molecular evolution on RVB segments. Bayesian analysis measured nucleotide substitution using BEAST software with relaxed molecular clock, 69 sequences collected both publicly from GenBank and isolated samples sequenced of gastroenteritis cases. Results showed meanRates for Human RVB segments from 2.28 x 10^{-3} s/s/y for (NSP4) as the fastest followed by 2.26 x 10^{-3} for (NSP3) and the slowest meanRate was 1.61 x 10^{-3} s/s/y for (VP6) gene, with (95% HPD intervals from 0.52 x 10^{-3} to 3.67 x 10^{-3} s/s/y). Here project's findings agreed with Lahon's findings in RVB evolution speeds.

4.3.4 Substitution rate and hosts

Referring to Tables 19A and 19B, all concordant genera and families with their corresponding species were listed, viral species hosts showing similar evolution speed were mammalian and plant hosts with almost similar percentages.

Mammalian RNA viruses known to have high substitution rates. Recent findings showed that cell tropism which is specific cell types targeted by a virus can be the cause of long-term substitution rates in mammalian RNA viruses. Viruses infecting epithelial cells as in respiratory and gastrointestinal tracts, tend to have higher substitution rates and more

genetic diversity than viruses infecting neurons. This evolution speed variation can be due to viral generation time differences (Hicks and Duffy, 2014).

4.3.5 Summary of substitution rate

To conclude substitution rate analysis findings in the project, I will start by looking at the diverse evolutionary speed across 350 alignments with majority fall in the moderate and fast evolution around 40% and 45% respectively, while very slow and very fast evolution categories were rare. Family level patterns showed that Spinareoviridae and Polyomaviridae were dominant among slow evolving alignments. Moreover, Geminiviridae were appearing at moderate, fast and very fast evolving alignments. Segmented viruses showed mixed substitution rate pattern, with 52% has concordant rates across segments and 48% were discordant. At multiple taxonomical hierarchy levels, only 2 out of 13 families and 8 out of 27 genera displayed full concordance indicating that substitution rate does not align with taxonomical classification. Even though substitution rate revealed not being a taxonomical marker within alignments studied, it remains fundamental in understanding viral evolution.

4.4 Selection pressure

4.4.1 Alignments with positive selected sites

Analysis of 448 alignments using sitewise likelihood ratio SLR program, 268 alignments had at least one positively selected site, while 180 alignments had no site under statistically significant selective pressure, which means that about 60% of the alignments has sites under positive selection. This percentage showed that positive selection is pervasive but not universal, still 40% of alignments had no positive sites.

60% of positive selection, represents that large number of viral proteins are having adaptive changes, for instance, this could be in responding to host immune system, replication, and binding functions. While the remaining 40% of alignments where there is no detected positively selected site, may be under neutral or purifying selection.

Kustin and Stern (2021) studied viral evolution characteristics focusing on RNA viruses, they mentioned that the diversity of RNA viruses is remarkable; they may infect a wide range of hosts and different morphologies, genomic structures, and genetic features. However, they are all characterized by similar evolutionary properties, such as host cells depending, high rates of mutation, purifying and positive selection even if being irregular.
Hughes and Hughes (2007) studied diversity in number of viral sequences datasets publicly available including 27 different families. The authors used statistical methods to compare nonsynonymous to synonymous substitutions ratio between RNA and DNA viruses. Purifying selection was found to be prevalent in 222 separate viral sequence datasets when the pattern of nucleotide diversity was analysed. In every dataset except for 11, synonymous substitutions (dS) were higher than nonsynonymous substitution (dN). Hughes findings indicate that purifying selection has been more successful in lowering the frequency of nonsynonymous variants in RNA viruses than in DNA viruses, therefore RNA viruses are shown to be under stronger purifying selection.

Number of positive selected sites for each alignment ranged from 1 to 203 sites, a literature search was conducted for the highest number of positive selected sites recorded to compare outputs.

- Atypical porcine pestivirus 1 (APPV)

  Within the Flaviviridae family, the genus Pestivirus comprises single-stranded, positive-sense RNA viruses of veterinary significance (Blome et al., 2017).

  In project's findings, 203 positive sites were identified in Atypical porcine pestivirus 1 alignment which is the highest among studied datasets.

  Folgueiras-Gonzalez et al. (2020) studied molecular evolution of APPV genomes using Datamonkey evolution server on isolated farm sequences additionally to Publically available sequences from 2013 to 2019, no positive selection was found.

  Folgueiras-González's evolutionary study findings on atypical porcine pestivirus, were contrary to this projects' finding. Results difference probably due to different time range for sequences analysed, additional to methods selected; as the previous study used Datamonkey webserver, applying Mixed Effects Model of Evolution (MEME) algorithm which is a likelihood ratio test detects sites subjected to episodic diversifying selection (Murrell et al., 2012), while the SLR indicates the presence of selection by performing a likelihood-ratio test for selection at each site in the alignment.

- Schmallenberg virus (SBV) and Akabane virus (AKV)

  Padhi and Ma (2015) examined M segment data for Schmallenberg virus and Akabane virus which are both negative-sense RNA viruses. Selection pressure was studied using method CODEML of PAML software and yielded to the conclusion that both viruses M segments are affected with different selection pressures; AKV found being under strong purifying selection, and intense positive selection in SBV.

  Schmallenberg virus and Akabane virus are two species belonging to genus Orthobunyavirus and Peribunyaviridae family.

  SBV is a second example of alignment dataset with high number of selected sites, SBV recorded 77 positive selected sites in SLR run. While AKV recorded only one selected site for M segment sequences.

  Findings of Padhi and Ma, agreed with our SLR output findings.

- Coxsackievirus A (CVA)

  Coxsackievirus A is a positive single-stranded RNA virus, a member of the Enterovirus A species and Picornaviridae family.

  Cheng et al. (2022) used two methods to study dN/dS within Datamonkey program in VP1 gene for CVA sequences, codons identified to be under negative selection in both methods, the only codon with positive selection evidence was VP1-145. Additionally, in the discussion part the author mentioned that selection was also studied in Coxsackievirus B VP1 region in a previous study by (Henquell et al., 2013) and came to conclusion that CVB evolve under purifying selection pressure.

  In our selection analysis Coxsackievirus was represented by two viral species; Coxsackievirus A2 and Coxsackievirus B3, neither of which showed any positive selected sites in SLR run, which is almost similar to the findings of Cheng and Henquell. On the other hand (Khan and Khan, 2021) identified numerous codons within both structural (VP1, VP3, VP4) and non-structural (2A, 2C, 3A, 3D) viral proteins that show episodic positive selection within 11 datasets containing 527 genomes of CVA using Datamonkey server with (MEME) algorithm.

4.4.2 Kappa and Omega

Previously mentioned in Results section 3.5.1 overall Kappa (κ) and Omega (ω) values were calculated for each alignment run, Figures 45 and 46 are scatter plots for overall Omega and Kappa values against length of alignment which both showed no correlation.

It has been noticed in many cases Kappa has a high value with no selected sites detected on the sequences of alignment. The lowest Kappa value recorded in SLR run was 0.68 in Cardamom bushy dwarf virus with 2 positive selected sites, while the highest recorded 49.8 for Cache Valley virus which has 0 positive selected site.

Omega values are correspondingly with number of positive selected sites present in the alignment dataset. Omega values greater than one indicates positive selection, when non-synonymous protein change is greater than synonymous change, positive selection occurs. This means Omega reflects positive selection, but Kappa transition/transversion rate ratio explains mutation bias, it does not measure selection directly. Globally, transitions rates (purine to purine or pyrimidine to pyrimidine) are mostly higher than transversion rate (purine to pyrimidine and vice versa) which can be related to selection (Knies et al., 2008).

4.4.3 Family level taxonomy

Multiple taxonomical levels could have different rates of positive, neutral, or negative selection. According to pie chart in Figure 47 representing Family level taxonomic distribution for alignments with positive selected sites in SLR, Geminiviridae took the top hit with 17%. Geminiviridae viruses have a single-stranded DNA genome, infecting variety of plant species, including agricultural crops like tomatoes, cassava, cotton, beans, and peppers. In SLR analysis 45 species alignments showed positive selected sites ranging from 1 to 25 site per alignment. Literature search on Geminiviridea showed more than one study agreed with our finding.

Deom et al. (2021) used CODEML maximum likelihood method to study dN/dS for Geminiviruses publicly available sequences on C4(AC4) and C1 (AC1) genes. The nucleotide substitution ratio showed that sequences for C1(AC1) are under purifying selection, and C4(AC4) genes are evolving under positive selection.

Review of Medina-Puche et al. (2021) focused on the C4 protein encoded by Geminiviruses, studying its properties and evolutionary behaviour. C4 exhibits a wide range of functions during viral infection, showing variability both within individual Geminivirus species and

across different Geminiviruses. Sequences analysed for dN/dS revealed the fact that C4 genes were under positive selection in more than half of datasets.

Figure 48 showed family taxonomy distribution for alignments had no positive selection, Sedoreoviridae and Spinareoviridae had the top hits with 21% and 14% respectively, no literature was found for any of these families discussing selection pressure.

4.4.4 Positive selection in hosts

Referring to Table 20 representing number of positive sites for all type of host species yielded from polymorphism study for selected sites of ++ and more, the Table showed that mammalian and plant hosts had the highest number of positive selected sites in that study.

- Viral species selection with mammalian hosts

  With a total of 513 selected sites found in viral alignments from 65 mammalian host species, mammals' host had the highest number of positively selected sites and average sitewise Omega values, which can indicate that viruses experience strong selection pressure from mammalian hosts.

  Wang and Han (2021) have shown in their study that adaptive evolution is prevalent in host-virus interactions across mammalian orders.

- Plants

  111 viral species had plants as a host, with a total 332 positive selected sites within alignments, a literature search for plants immune system showed that plants have number of genes that help them detect and defend against harmful microbial infections. But this constant defence also enhance pathogens selection pressure to escape detection by the plant's immune system (Wang et al., 2022b). One of microbial infections mentioned in the review was Tomato yellow leaf curl virus, the C4 protein disrupts the production of salicylic acid by interacting with the calcium-sensing receptor (CAS) in chloroplasts, and start affecting plant immune response, followed by symptoms development.

  This virus was studied within alignments in SLR, 6 positive selected sites were found with an average sitewise omega value of 3.7 for the 6 sites present, positive selected sites were present in region encoded for AC1, AC2 and AC3 proteins.

4.4.5 Gene Ontology

- GO terms and functions

  Referring to pie chart in Figure 54, study of Gene Ontology terms on selected proteins showed that majority are related to structural molecule activity, binding, and replication. Starting with GO term "structural molecular activity" which refers to the role of a molecule in maintaining the stability or shape of larger complex. Structural molecular activity scored 10% of total GO terms; positive selection in these proteins can indicate adaptation to enhance viral structure and stability. Also, "RNA-binding" and "RNA-dependent RNA polymerases" both has 8% of total GO hits; positive selection suggests adaptation to viral replication and perseverance within the host. Followed by "viral capsid", suggests adaptation to viral replication and pathogenicity when positive selection found in proteins related to them. "ATP Binding" followed, with 6%; Positive selection in their proteins promotes energy utilizing and efficiency through viral replication.

  Next in the top hits are "DNA replication" and "DNA-templated transcription", where both shows adaptation to viral replication, gene expression, and viral stability with positive selected sites on their proteins (Ahlquist et al., 2003; Venkataraman et al., 2018).

- Positive selection and immune evasion

  Among the top hits GO terms recorded, "viral capsid" which scored 8% of GO hits among selected proteins is the highest related to immune evasion mechanism. "Viral capsid" has a role in viral infection as it contributes with immune evasion by genome protecting against host immune system, also capsid can undergo some alterations to escape host immunity.

  Woelk and Holmes (2002) performed selection pressure analysis using CODEML package to study nature of viral selection generally and to discover evolving nature of vector-borne RNA viruses compared to viruses with different chain of transmission. The target was viruses with human host, sequences retrieved from GenBank for both surface and internal structure including outer capsid and core capsid. Only four viruses showed the evidence of significant positive selection which are: Measles virus, Oropouche virus, hepatitis C virus and HIV virus. In general, the findings indicate that

vector-borne RNA viruses are not as prone to positive selection compared to viruses transmitted through other ways. This is more noticeable in genes of the envelope glycoprotein; such genes often contain sites that undergo adaptive evolution in nonvector-borne viruses. The author added that it is highly probable that the selection pressure in this scenario is linked to immune evasion. This is because the envelope and outer capsid proteins often contain epitopes that can be targeted by neutralising antibodies and T-cell responses. It is not surprising that the genes responsible for creating the core structure of the virus are highly conserved in both types. Out of the four viruses mentioned in the review, two of them were found in the positive selection analysis done in this project. Oropouche virus appeared with two segments in the SLR study showed three positive selected sites, these sites encode two distinct viral proteins. Functional analysis showed that both proteins are associated with GO:0019013 (viral nucleocapsid) derived from viral capsid function in ancestor chart, which may perform some adaptive changes to evade host immune defence. The second virus was Measles virus which showed 10 positive selected sites and encodes for four different proteins, these proteins appeared in the InterProscan output under more than one domain, and functionally related to several GO terms which are structural molecular activity, RNA-dependent-RNA polymerase, ATP binding and viral nucleocapsid.

- Positive selection and polymerases

  Polymerases are enzymes responsible for genetic material replication. In terms of viral selection pressure, their ability to adapt and escape host immune response is controlled by polymerases.

  The top hits in GO terms had polymerases encoding for their selected proteins, for this below we are discussing role of polymerases in viral selection pressure from literature. Genetic diversity created when nucleotide substitution is introduced by RNA-dependent RNA polymerases. It has been observed that the selection pressure effect on hosts and vectors mark regions of the genome lead to host adaptation (Nigam et al., 2019).

  To understand viral evolution and contain infections, features of their population in transmission is highly important. A naturally sustained degree of genetic heterogeneity within the population is caused by the error rate of RNA viruses RNA-dependent RNA polymerases. The fitness of viral populations depends on this diversity because it

enables the virus to quickly adapt to a new genetic environment in response to various selective pressure (Varble et al., 2014).

An evolving virus can be defined by its ability to generate wide range of genome variants in new generations. This first observed in RNA viruses due to quasispecies phenomenon affecting the nature of RNA polymerases. This property has significant implications for viral evolution, impacting viral viability, virulence, immune evasion, vaccine resistance, host adaptation, and interactions between different variants. Natural selection plays a role in fixing such mutations by applying various filters that decrease mutant variations population size. The formation of quasispecies reflects the dynamic relationship between polymerase activity, genetic diversity, and natural selection, highlighting the central role of polymerases in viral evolution (Dupre and Volmer, 2023).

In the projects viral datasets, one of the viruses showed positive selection in encoded proteins related to RNA-dependent-RNA polymerase is Bean yellow mosaic virus (BYMV), it showed two positive selected sites coding for one polyprotein and two mature peptides regions, coat protein CP and P3 protein. Parrella and Lanave (2009) studied positive selection on (BYMV) isolates using Datamonkey software, the selection analysis revealed evidence of positive selection in the N-terminal region of the CP, this region known for its variability among potyvirus genome.

- GO terms and hosts

  Section 3.5.6 in Results listed each GO term with its percentage for viral host type. According to Table 23 GO functions has different percentages among hosts; as in 68 % of viral capsids present with plant hosts, followed by RNA binding that is 56% in plant hosts. While RNA-dependent-RNA polymerase activity showed the highest percentage in mammalian hosts with 44%.

  Understanding differences in GO terms proportion among host types can start by knowing host immunity and cellular structure and processes. Also, host environment can impact viral proteins interactions. GO annotations can be affected by both viral and host genes similar functions as in cellular processes and molecular functions. Some newly added plant pathogens GO terms found to be related to some animal pathogens, for e.g., a viral protein in Herpesvirus resembles host proteins functionally which both annotated for host transcription modulation and transcriptional regulation. This

highlights the applicability of GO annotations in understanding host-pathogen interactions across various taxonomic groups (McCarthy et al., 2009).

Polymerase proteins mutations are usually involved in avian viruses adaptation to mammalian hosts. Some residues within polymerases proteins have been identified as determinants of host range and replication efficiency (Moncorge et al., 2010).

Adaptation to mammalian hosts involve mutations which enhance polymerase activity and allow viral replication at low temperature, as the ones found in human upper respiratory tract (Hayashi et al., 2015).

4.4.6 Polymorphism amino acids

As previously mentioned, Results section 3.5.3 polymorphism analysis was performed in alignments with positive selected sites, the low values correlation coefficient for polymorphism number with both Omega and thickness of alignments showed that number of polymorphic amino acids is not affected with strength of selection.
In the same section, four examples of viral dataset alignments with different polymorphism amino acids and Omega values were listed. According to the location of diverse amino acids in phylogenetic trees, positive selection type can be directional or diversifying. In terms of defining the two types of selection, Kosakovsky Pond et al. (2008) used phylogenetic Maximum Likelihood method (ML) to study positive selection in Influenzae virus, they mentioned that directional selection refers to the process when certain residue is consistently selected, which leads to high presence of this specific amino acid or allele over time, and this can be due to accelerated substitution rates toward the specific residue. While diversifying selection can be caused by genetic variation to favour more than one amino acid at certain site, leads to multiple variants in a population.

The first example represented high sitewise Omega with low diversity which indicates directional selection, then both; high Omega with high diversity and low Omega with high diversity suggests diversifying selection showing multiple mutations within lineages or branches, finally low Omega with low diversity suggests directional selection as only one

amino acid appeared among nodes. The four examples polymorphism location suggested that type of positive selection can be indicated by level of diversity not strength of selection.

Furthermore, for a better understanding of highest variable selected sites functionally, the highest polymorphic datasets collected and GO terms were identified for their selected proteins irrespective for their Omega values. Table 24 shows the top five alignments with positive selected sites under diversifying selection.

Starting by Rotavirus C, the 13 polymorphic amino acids site on VP1 gene with sitewise Omega value 2.7, alignment length was 1090 and did not show any other positive selected site on the peptide chain. Literature search on Rotavirus C and positive selection showed that some genes found to be under positive selection but not VP1, moreover Joshi et al. (2019) performed selection pressure analysis on RVC 11 genes using two softwares, the study came with a result that two codons of VP3 and NSP5 single codon are under diversifying selection, for the remaining genes no positive selection was observed.
The three GO terms associated with Rotavirus C structural protein VP1 are RNA binding, RNA-dependent RNA polymerase, and DNA-templated transcription. Positive selection with this high polymorphism indicates adaptive evolution mechanism enhancing viral replication and binding.

The following example with 12 polymorphic diverse site is Bovine viral diarrhea 1, the positive site appeared in the mature peptide segment encodes for structural protein E2, the mat-peptide has no related GO terms on InterProscan, and the term related to the Polyprotein had 13 GO functions. Selection on Bovine viral diarrhea 1 is further discussed in section 4.4.7

The third example was Kibale red colobus virus 1 also known as Simian haemorrhagic fever virus (SHFV-krc1), which recoded one of the highest alignments in positive selected sites number, the positive selected site in Table 24 has 12 polymorphic amino acids and found in the region encodes for large glycoprotein "ORF7 protein". It was found to be associated with only one GO function which is viral envelope. The high polymorphism observed in ORF7 with the high number of positive selected site in viral alignment suggests adaptive evolution applied on the encoded envelope protein, which can be related to host immune responses evasion. Bailey et al. (2014) studied positive selection in Kibale red colobus viruses open reading frames (ORFs). In Kibale red colobus virus 1 positive selection was identified in ORF7 and ORF3. The author added that SHFV-krc1 high

genetic diversity and adaptive potential elevates its infection frequency and rapid evolving potential among RNA viruses.

The last two examples are for Grapevine fanleaf virus and Potato virus S, which both selected sites found to be within mature peptides and has no specific GO terms related. Protein functions are listed in Table 24.

Table 24: Illustrates alignments sets with the highest number of polymorphisms with corresponding protein encoded and related GO functions for the selected protein.

Protein ID	Viral Alignment	Poly-morphism Number	Protein Product	GO Terms	GO Function
YP_392464	Rotavirus C (Site number 193)	13	Structural protein VP1	GO:0003723 GO:0003968 GO:0006351	RNA binding RNA dependent RNA polymerase DNA-templated transcription
NP_040937.1 NP_776263.1	Bovine viral diarrhea 1 (Site number 876)	12	Polyprotein Envelope glycoprotein E2-Structural protein E2	GO:0004386 GO:0005524 GO:0003968 GO:0004197 GO:0004252 GO:0016817 GO:0017111 GO:0070008 GO:0016032 GO:0019082 GO:0003723 GO:0039694 GO:0004252	Helicase activity, ATP binding, RNA dependent RNA polymerase, cysteine-type endopeptidase activity, hydrolase activity, acting on acid anhydrides, serine-type exopeptidase activity, ribonucleoside triphosphate phosphatase activity, viral process, viral protein processing, RNA dependent RNA polymerase, viral genome replication, serine-type endopeptidase activity.
YP_009344816	Kibale red colobus virus 1 (Site number 5194)	12	Large glycoprotein	GO:0019031	viral envelope
NP_619689.1 NP_734039.3	Grapevine fanleaf virus (Site number 2184)	10	253K polyprotein Protease cofactor	GO:0003723 GO:0003968 GO:0006351 GO:0003724	RNA binding, RNA dependent RNA polymerase activity, DNA-templated transcription, RNA helicase activity
YP_277428.1	Potato virus S (Site number 660)	9	RNA-directed RNA polymerase	GO:0003968 GO:0016817 GO:0003723 GO:0006351 GO:0006396 GO:0008174 GO:0005524	RNA dependent RNA polymerase activity, RNA binding, ATP binding, DNA templated transcription, RNA processing, mRNA methyltransferase activity.

4.4.7 Protein structure

As previously mentioned in Results section, blastall command was applied on selected proteins to get protein matches and percentages. From selected proteins list 33% found to have matches on PDB database, and only 8% has 100% match.

This limited percentage of protein structure availability in the PDB database may be due to several factors. Some proteins may not have been widely studied before, so their structures were never submitted to PDB. Additionally, proteins chosen, or secondary and tertiary structure analysis may not have undergone positive selection analysis previously. Furthermore, the PDB database has its own limitations, as the methods used to determine protein structures may not be applicable to all proteins.

Although structural models and epitope regions were generated using prediction tools including online antigenic epitope prediction software and MOE for proteins structure visualization and molecular surface modelling, the interpretation of functional impact was not based only on these predictions. For four representative proteins, the presence of positively selected sites was assessed and related to literature describing their evolutionary behaviours and functional roles. While these findings remain predictive in origin, integrating computational outputs with biological evidence supports informed hypotheses about how specific amino acid substitutions may affect protein morphology and antigenic properties.

Starting with the first example displayed in section 3.5.7 studying cases from alignments dataset tertiary structure on MOE was Porcine parvovirus, positive selected site identified on its peptide chain and later highlighted on 1K3V protein, Figures 59B and 59C represented the molecular surfaces created on Isoleucine (I) residue and the mutated residue on the same site Methionine (M), differences observed can be due to variation in both amino acids side chain structures. Methionine contains sulfer atom with longer side chain, which can increase flexibility and affect surface. While, Isoleucine is an aliphatic amino acid, which is non-polar and more hydrophobic which can lead to much rigid surface (Al Mughram et al., 2023; Aledo, 2019).

Ohmura et al. (2001) agrees to previous finding, they observed differences between Methionine and Isoleucine substitutions, on lysosome gene mutation study measuring stability changes using X-ray crystallography and structure. They report that the substitution of Isoleucine with Methionine can lead to slight instability, and this mainly can be from hydrophobicity change between the two residues.

Protein ID from where the positive selected site was identified has been checked to be related to "viral capsid" and "structural molecular activity" GO terms in InterProscan. Mutation in surface residues of capsid proteins can alter their functions either by interacting with host receptors or immune molecules. So, if hydrophobicity or flexibility properties has been affected as a result of residues mutation, the binding capability might be affected as seen in Hueffer et al. (2004) study, when modifications in feline transferrin receptor (TfR) affect parvovirus entry to host cell. When (TfR) variants were created, the study showed an impact on receptor-mediated endocytosis, localization within membrane domains and receptor's ability to bind to viral capsid.

The second example mentioned in Results for Human metapneumovirus positive selection on an internal residue, which located within a protein does not involve into surface interactions, so no molecular surface was created.

The following example was for positive selected site on Bovine viral diarrhea virus 1. Figures 61B and 61C for molecular surfaces created over Glutamine (Q) and one of the mutants Histidine (H) showed difference which can be due to the two amino acids difference in both side chains. The imidazole ring in Histidine can lead to the bulkier surface compared to Glutamine (Bhattacharyya et al., 2003).

Additionally, BVDV1 was previously mentioned in section 4.4.6 within polymorphism analysis performed as one of the examples of highest polymorphism amino acids appeared at one site. The positive selected site studied on the solved structure has 9 different amino acids within 70 sequences in alignment dataset with collection dates ranges from 1990 to 2020, this site appeared encoding for (polyprotein) envelope glycoprotein E2 and associated with number of GO terms, see table 24 for GO terms annotations. While reviewing the literature on positive selection in BVDV, Miroslaw and Polak (2020) identified signals of positive selection in the E2 glycoprotein region of BVDV. As envelope glycoproteins are proteins on enveloped viruses surfaces and plays a critical role in primary stage of viral infection. The observed polymorphism in E2 is likely caused of selection pressure to escape host immune system during viral attachment and entry.

Also, King et al. (2002) studied the association of RNA dependent RNA polymerase mutation with resistance to antiviral compounds, by examining how resistance to a thiazole urea compound, an antiviral drug targets BVDV, develops. Their findings showed the importance of this domain in viral replication and survival, showing that changes in RdRp

alter its nucleotide binding function and later enhance virus ability to replicate in the presence of the drug.

The last example was for positive selected sites in Machupo mammarenavirus, Figures 62B and 62C showed lipophilic molecular surfaces done on two different amino acids Methionine (M) and Valine (V) respectively. The difference of the two molecular surfaces can be from the longer Methionine side chain which ends in a sulfur group compared to the branched Valine structure with shorter side chain.

Methionine sulfur atom promotes some interactions as oxidization that stabilize structure and function of protein, Methionine substitution with Valine can reduce stability due to the lack of sulfur atoms and other binding efficiency (Lim et al., 2019). Also, Methionine role in stabilizing interactions with other proteins through sulfur aromatic reactions cannot be optimized by any other hydrophobic amino acid as Valine or Isoleucine (Valley et al., 2012).

Additionally, two more molecular surfaces were created for chains A and B in the PDB structure to highlight the location of the antigenic sites as mentioned in Results section, Figures 64A showed the colour chain for the two chains and 64B shows lipophilic molecular surface only for epitopes on the peptide chains, these two images identify that only one of the two selected sites is an antigenic site and is highlighted on an arrow at Figure 63.

The region where positive sites appeared encodes for "glycoprotein precursor" has been associated with viral envelope GO term. The first site which was found in the antigenic region is highly exposed to the surface of the viral envelope and recognised by host immune system, positive selection in this site can reflect immune evasion, where mutation allows the virus to escape immunity by performing changes on the epitope.

While the second one which is not an antigenic site can be more related to functional selection that interacts to viral structure, the mutation on this site suggests that selection has the ability to improve the two monomeric interactions.

Pathogenic arenaviruses, including highly pathogenic mammarenaviruses such as Machupo virus, have the ability to activate weaker interferon (IFN) responses compared to non-pathogenic members, using innate immune evasion as a key factor of their emergence and adaptation (Moreno and Kunz, 2021).

In Machupo virus, the S segment encodes two main protein glycoprotein precursor and nucleoprotein. The viral glycoprotein is responsible for the virus initial attachment of the virus surface to host cell receptors, making it a target for immune recognition and antibody generation. Due to its high diversity, immunity cannot last long because of these surface

glycoproteins, as mutation driven by positive selection facilitate viral evasion from immune response (Naveed et al., 2022).

4.4.8 Summary of selection pressure

Finally, conclusion of selection pressure findings will pass through subsequent stages. Starting by the main finding of SLR run that 60% of alignment datasets had at least one positive selected site explains the adaptive evolution in viral genomes, allowing viruses to change overtime helping them to survive and spread. At the family level positive selected sites were mostly appearing in Geminiviridae and Potyviridae, while alignments with no positive selected sites found were dominated with Sedoreoviridae and Spinareoviridae families. For the variability of positive selection among alignments, concordance analysis was not conducted across segmented viruses or for taxonomical classification. Instead, functionally relevant protein domains were analysed, with Pfam domain hits closer to viral taxonomy. Further understandings from GO terms annotations indicated that positive selected proteins are associated with structural molecular activity, RNA binding, RNA-dependent RNA polymerase activity, and viral capsid proteins. This can disagree with the common understanding that viral evolution is driven by immune evasion, as the project findings suggests a stronger selection pressure on structural, replication and binding functions. Additionally, polymorphism analysis can support this as represented in some examples phylogeny, that greater amino acid diversity correlated with diversifying selection. Lastly, tertiary structure and molecular surfaces created over positive selected sites and their mutants confirmed that selection can affect protein structure supporting functional impact of evolution.

4.5 General conclusion

The project was initially driven by the question of whether tempo and mode of viral evolution exhibit a consistent pattern across taxonomic classification, and whether evolutionary parameters should serve as taxonomic markers. Viral evolution was studied by analysing tempo (substitution rate) and mode (selection pressure) across taxonomic levels, along with recombination and temporal signal patterns. Key findings showed that while evolutionary parameters display both consistency and variability, they generally do not align directly with taxonomic classification.

Recombination analysis highlighted taxonomic patterns, with Potyviridae and Geminiviridae frequently showed evidence of recombinant sequences, while Reoviridae lacked it. Segmented viruses had mixed patterns (57% concordant, 43% discordant), and only a minority of families and genera demonstrated full concordance, indicating that recombination or its absence is not a taxonomic marker. Similarly, temporal signal analysis found that 59% of alignments were suitable for evolutionary rate estimation. However, discordance was common across segments (47 concordant vs. 32 discordant) and taxonomic levels (only 5 out of 29 families and 16 out of 49 genera were concordant), indicating that temporal signal is taxonomically inconsistent.

Substitution rates varied widely, with most viruses falling into moderate 45% or fast 40% evolutionary speed categories. Geminiviridae appeared in multiple rate categories, while Spinareoviridae and Polyomaviridae dominated slow evolving groups. Segmented viruses displayed almost similar concordance (52% vs 48%), and minimal taxonomic consistency (only 2 out of 13 families and 8 out of 27 genera are concordant) which also show that substitution rate is not a taxonomic marker.

Selection pressure analysis displayed evolutionary variability, as 60% of alignments showed positive selection, particularly in Geminiviridae and Potyviridae. Functional annotations linked the positive selected sites to structural, replication and binding domains (e.g., capsid proteins, RNA polymerases), challenging the understanding that immune evasion dominates viral evolution. Structural modelling confirmed that selected sites might often alter protein conformation, emphasizing their functional significance.

Collectively, these results demonstrate that while viral evolution is shaped by recombination, mutation, selection and temporal patterns, these forces do not consistently align with formal taxonomic grouping, indicating that tempo and mode are not constrained with taxonomic structure.

4.6 Limitations of current study

1. Reduction of data amount collected for evolutionary study: during filtration processes, many viral genomes were removed with species have no reference genome available. Also, large genome size with high number of sequences per species were removed to minimize time loss computationally, and low sequences number have been removed for being unsuitable to build reliable phylogenetic tree. Additionally, primary metadata was lacked from some sequences as collection date and CDS which are essential to study tempo and mode.
2. Moreover, the use of one software for each parameter used to study tempo and mode in viral alignments datasets, although these tools are well established for their respective analysis, relying on only one method per parameter may introduce bias or limitations in the results.
3. This study entirely used viral genome sequences available in public repositories, primarily GenBank. Due to some high presentation of some viruses after viral pandemics or epidemics work submitted, availability of sequences bias is present, and this effect results in multiple ways, including overrepresentation of certain hosts, specific populations, certain viral families or some genome segments and specific protein products in final alignments. Although some well-studied viruses were removed when datasets went through filtering stages, taxonomic distribution exhibited some limitations of what is available in the alignment datasets.

4.7 Future work suggestions

1. High-performance computing (HPC) systems to enable processing of larger datasets which can include larger genomes as many DNA viruses, with minimal time loss computationally.
2. Employ additional software to detect presence of recombinant sequences in viral alignments using different algorithm, such as 3SEQ, to handle larger datasets.

3. Secondary and tertiary structure for all selected proteins available structures to investigate the effect of positively selected sites on the peptide chain. Also, using tools like SWISS MODEL for homology modelling of protein 3D structures.
4. Apply new generation of analysis using real-time phylogenetic tracking as Nextstrain platform to deal with larger datasets as Herpesviridae and Poxviridae.
5. Analyse high annotated phylogenies by Archaeopteryx, which is a software for visualization, analysis and editing of large phylogenetic trees.
6. Incorporating evolutionary characteristics additional to virus taxonomy. As these parameters reflect how viruses evolve over time and can help clarify relationships between viruses that are not accurately captured by current taxonomy features.
7. Linking evolutionary traits to host range and cross-species transmission. This can evaluate weather specific taxonomy group share evolutionary signatures linked to host adaptation or zoonotic potential and can identify viruses with pandemic risk.
8. Machine learning approaches can be further used to find patterns linking evolutionary traits (e.g., dN/dS ratios, recombination breakpoints) with taxonomy improving classification and outbreak prediction.

Chapter 5: Bibliography

Abbadi, M., Gastaldelli, M., Pascoli, F., Zamperin, G., Buratin, A., Bedendo, G., Toffan, A. & Panzarin, V. (2021) Increased virulence of Italian infectious hematopoietic necrosis virus (IHNV) associated with the emergence of new strains. *Virus Evol,* 7(2), veab056. 10.1093/ve/veab056.

Adriano Raiano. internationalization-framework. Available at: https://flowchart.js.org/ [2023].

Ahlquist, P., Noueiry, A. O., Lee, W. M., Kushner, D. B. & Dye, B. T. (2003) Host factors in positive-strand RNA virus genome replication. *J Virol,* 77(15), 8181-6. 10.1128/jvi.77.15.8181-8186.2003.

Al Mughram, M. H., Catalano, C., Herrington, N. B., Safo, M. K. & Kellogg, G. E. (2023) 3D interaction homology: The hydrophobic residues alanine, isoleucine, leucine, proline and valine play different structural roles in soluble and membrane proteins. *Front Mol Biosci,* 10, 1116868. 10.3389/fmolb.2023.1116868.

Aledo, J. C. (2019) Methionine in proteins: The Cinderella of the proteinogenic amino acids. *Protein Sci,* 28(10), 1785-1796. 10.1002/pro.3698.

Angeletti, S., Benvenuto, D., Fogolari, M., De Flora, C., Ceccarelli, G., Maida, I., Bazzardi, R., Spoto, S., Pascarella, S., Mugosa, B. & Ciccozzi, M. (2021) The Bayesian reconstruction and the evolutionary history of Salivirus type 1 and type 2: the worldwide spreading. *J Infect Dev Ctries,* 15(2), 280-288. 10.3855/jidc.12141.

Anisimova, M. & Yang, Z. (2004) Molecular evolution of the hepatitis delta virus antigen gene: recombination or positive selection? *J Mol Evol,* 59(6), 815-26. 10.1007/s00239-004-0112-x.

Ayres, D. L., Darling, A., Zwickl, D. J., Beerli, P., Holder, M. T., Lewis, P. O., Huelsenbeck, J. P., Ronquist, F., Swofford, D. L., Cummings, M. P., Rambaut, A. & Suchard, M. A. (2012) BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics. *Syst Biol,* 61(1), 170-3. 10.1093/sysbio/syr100.

Baele, G., Ayres, D. L., Rambaut, A., Suchard, M. A. & Lemey, P. (2019) High-performance computing in Bayesian phylogenetics and phylodynamics using BEAGLE. *Evolutionary genomics: statistical and computational methods*, 691-722.

Bakhache, W., Symonds-Orr, W., Mccormick, L. & Dolan, P. T. (2025) Deep mutation, insertion and deletion scanning across the Enterovirus A proteome reveals

constraints shaping viral evolution. *Nat Microbiol,* 10(1), 158-168. 10.1038/s41564-024-01871-y.
Barba-Montoya, J., Tao, Q. & Kumar, S. (2020) Using a GTR+Gamma substitution model for dating sequence divergence when stationarity and time-reversibility assumptions are violated. *Bioinformatics,* 36(Suppl_2), i884-i894. 10.1093/bioinformatics/btaa820.
Baur, S. (2023) *X2Go - everywhere@home*. Available at: https://wiki.x2go.org.
Benson, D. A., Karsch-Mizrachi, I., Clark, K., Lipman, D. J., Ostell, J. & Sayers, E. W. (2012) GenBank. *Nucleic Acids Res,* 40(Database issue), D48-53. 10.1093/nar/gkr1202.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) The Protein Data Bank. *Nucleic Acids Res,* 28(1), 235-42. 10.1093/nar/28.1.235.
Berrio, A., Gartner, V. & Wray, G. A. (2020) Positive selection within the genomes of SARS-CoV-2 and other Coronaviruses independent of impact on protein function. *PeerJ,* 8, e10234. 10.7717/peerj.10234.
Bhattacharyya, R., Saha, R. P., Samanta, U. & Chakrabarti, P. (2003) Geometry of interaction of the histidine ring with other planar and basic residues. *J Proteome Res,* 2(3), 255-63. 10.1021/pr025584d.
Blome, S., Beer, M. & Wernike, K. (2017) New Leaves in the Growing Tree of Pestiviruses. *Adv Virus Res,* 99, 139-160. 10.1016/bs.aivir.2017.07.003.
Bondaryuk, A. N., Belykh, O. I., Andaev, E. I. & Bukin, Y. S. (2023) Inferring Evolutionary Timescale of Omsk Hemorrhagic Fever Virus. *Viruses,* 15(7). 10.3390/v15071576.
Booker, T. R., Jackson, B. C. & Keightley, P. D. (2017) Detecting positive selection in the genome. *BMC Biol,* 15(1), 98. 10.1186/s12915-017-0434-y.
Bouckaert, R., Heled, J., Kuhnert, D., Vaughan, T., Wu, C. H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol,* 10(4), e1003537. 10.1371/journal.pcbi.1003537.
Bowden, R., Davies, R. W., Heger, A., Pagnamenta, A. T., De Cesare, M., Oikkonen, L. E., Parkes, D., Freeman, C., Dhalla, F., Patel, S. Y., Popitsch, N., Ip, C. L. C., Roberts, H. E., Salatino, S., Lockstone, H., Lunter, G., Taylor, J. C., Buck, D., Simpson, M. A. & Donnelly, P. (2019) Sequencing of human genomes with nanopore technology. *Nat Commun,* 10(1), 1869. 10.1038/s41467-019-09637-5.
Bremer, C. W., Huismans, H. & Van Dijk, A. A. (1990) Characterization and cloning of the African horsesickness virus genome. *J Gen Virol,* 71 (Pt 4), 793-9. 10.1099/0022-1317-71-4-793.
Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. (2015) NCBI viral genomes resource. *Nucleic Acids Res,* 43(Database issue), D571-7 https://www.ncbi.nlm.nih.gov/genome/viruses/. 10.1093/nar/gku1207.

Bromham, L. (2020) Substitution rate analysis and molecular evolution. No commercial publisher| Authors open access book.
Brooksbank, C., Bergman, M. T., Apweiler, R., Birney, E. & Thornton, J. (2014) The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Res,* 42(Database issue), D18-25. 10.1093/nar/gkt1206.
Brown, R. P. & Yang, Z. (2011) Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evol Biol,* 11, 271. 10.1186/1471-2148-11-271.
Buchmann, J. P. & Holmes, E. C. (2020) Collecting and managing taxonomic data with NCBI-taxonomist. *Bioinformatics*. 10.1093/bioinformatics/btaa1027.
Bush, R. M., Bender, C. A., Subbarao, K., Cox, N. J. & Fitch, W. M. (1999) Predicting the evolution of human influenza A. *Science,* 286(5446), 1921-5. 10.1126/science.286.5446.1921.
Cardona, G., Rossello, F. & Valiente, G. (2008) Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics,* 9, 532. 10.1186/1471-2105-9-532.
Carr, M., Gonzalez, G., Sasaki, M., Ito, K., Ishii, A., Hang'ombe, B. M., Mweene, A. S., Orba, Y. & Sawa, H. (2017) Discovery of African bat polyomaviruses and infrequent recombination in the large T antigen in the Polyomaviridae. *J Gen Virol,* 98(4), 726-738. 10.1099/jgv.0.000737.
Chaitanya, K. (2019) Structure and organization of virus genomes. In: *Genome and Genomics.* Springer.
Chang, J. (2020) Biopython: Tutorial and Cookbook. Self-publishing.
Chare, E. R. & Holmes, E. C. (2006) A phylogenetic survey of recombination frequency in plant RNA viruses. *Arch Virol,* 151(5), 933-46. 10.1007/s00705-005-0675-x.
Cheng, D., Chiu, Y. W., Huang, S. W., Lien, Y. Y., Chang, C. L., Tsai, H. P., Wang, Y. F. & Wang, J. R. (2022) Genetic and Cross Neutralization Analyses of Coxsackievirus A16 Circulating in Taiwan from 1998 to 2021 Suggest Dominant Genotype B1 can Serve as Vaccine Candidate. *Viruses,* 14(10). 10.3390/v14102306.
Chi, P. B. & Liberles, D. A. (2016) Selection on protein structure, interaction, and sequence. *Protein Sci,* 25(7), 1168-78. 10.1002/pro.2886.
Choi, S. C., Hobolth, A., Robinson, D. M., Kishino, H. & Thorne, J. L. (2007) Quantifying the impact of protein tertiary structure on molecular evolution. *Mol Biol Evol,* 24(8), 1769-82. 10.1093/molbev/msm097.
Clark, J. J., Gilray, J., Orton, R. J., Baird, M., Wilkie, G., Filipe, A. D. S., Johnson, N., Mcinnes, C. J., Kohl, A. & Biek, R. (2020) Population genomics of louping ill virus provide new insights into the evolution of tick-borne flaviviruses. *PLoS Negl Trop Dis,* 14(9), e0008133. 10.1371/journal.pntd.0008133.
Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & De Hoon, M. J. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics,* 25(11), 1422-3. 10.1093/bioinformatics/btp163.
Cohen-Dvashi, H., Amon, R., Agans, K. N., Cross, R. W., Borenstein-Katz, A., Mateo, M., Baize, S., Padler-Karavani, V., Geisbert, T. W. & Diskin, R. (2020) Rational design

of universal immunotherapy for TfR1-tropic arenaviruses. *Nat Commun,* 11(1), 67. 10.1038/s41467-019-13924-6.
Collins, A., Lau, W. & De La Vega, F. M. (2004) Mapping genes for common diseases: the case for genetic (LD) maps. *Hum Hered,* 58(1), 2-9. 10.1159/000081451.
Comeau, D. C., Liu, H., Islamaj Dogan, R. & Wilbur, W. J. (2014) Natural language processing pipelines to annotate BioC collections with an application to the NCBI disease corpus. *Database (Oxford),* 2014. 10.1093/database/bau056.
D'addabbo, P., Lenzi, L., Facchin, F., Casadei, R., Canaider, S., Vitale, L., Frabetti, F., Carinci, P., Zannotti, M. & Strippoli, P. (2004) GeneRecords: a relational database for GenBank flat file parsing and data manipulation in personal computers. *Bioinformatics,* 20(16), 2883-5. 10.1093/bioinformatics/bth321.
Delong, J. P., Al-Sammak, M. A., Al-Ameeli, Z. T., Dunigan, D. D., Edwards, K. F., Fuhrmann, J. J., Gleghorn, J. P., Li, H., Haramoto, K., Harrison, A. O., Marston, M. F., Moore, R. M., Polson, S. W., Ferrell, B. D., Salsbery, M. E., Schvarcz, C. R., Shirazi, J., Steward, G. F., Van Etten, J. L. & Wommack, K. E. (2022) Towards an integrative view of virus phenotypes. *Nat Rev Microbiol,* 20(2), 83-94. 10.1038/s41579-021-00612-w.
Deom, C. M., Brewer, M. T. & Severns, P. M. (2021) Positive selection and intrinsic disorder are associated with multifunctional C4(AC4) proteins and geminivirus diversification. *Sci Rep,* 11(1), 11150. 10.1038/s41598-021-90557-0.
Domingo, E. (2001) RNA Virus Genomes. *e LS*.
Drummond, A. J. & Bouckaert, R. R. (2015) *Bayesian evolutionary analysis ; with BEAST.* Cambridge, United Kingdom: Cambridge University Press.
Drummond, A. J. & Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol,* 7, 214. 10.1186/1471-2148-7-214.
Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol,* 29(8), 1969-73. 10.1093/molbev/mss075.
Duchene, S., Featherstone, L., Haritopoulou-Sinanidou, M., Rambaut, A., Lemey, P. & Baele, G. (2020a) Temporal signal and the phylodynamic threshold of SARS-CoV-2. *Virus Evol,* 6(2), veaa061. 10.1093/ve/veaa061.
Duchene, S., Ho, S. Y. W., Carmichael, A. G., Holmes, E. C. & Poinar, H. (2020b) The Recovery, Interpretation and Use of Ancient Pathogen Genomes. *Curr Biol,* 30(19), R1215-R1231. 10.1016/j.cub.2020.08.081.
Duffy, S., Shackelton, L. A. & Holmes, E. C. (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet,* 9(4), 267-76. 10.1038/nrg2323.
Dupre, G. & Volmer, R. (2023) Influence of viral genome properties on polymerase fidelity. *Trends Genet,* 39(1), 9-14. 10.1016/j.tig.2022.10.008.
Fairweather, D., Stafford, K. A. & Sung, Y. K. (2012) Update on coxsackievirus B3 myocarditis. *Curr Opin Rheumatol,* 24(4), 401-7. 10.1097/BOR.0b013e328353372d.

Federhen, S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res,* 40(Database issue), D136-43. 10.1093/nar/gkr1178.
Ferrari, E., Salogni, C., Martella, V., Alborali, G. L., Scaburri, A. & Boniotti, M. B. (2022) Assessing the Epidemiology of Rotavirus A, B, C and H in Diarrheic Pigs of Different Ages in Northern Italy. *Pathogens,* 11(4). 10.3390/pathogens11040467.
Finn, R. D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. & Bateman, A. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res,* 34(Database issue), D247-51. 10.1093/nar/gkj149.
Fischer, W., Giorgi, E. E., Chakraborty, S., Nguyen, K., Bhattacharya, T., Theiler, J., Goloboff, P. A., Yoon, H., Abfalterer, W., Foley, B. T., Tegally, H., San, J. E., De Oliveira, T., Network for Genomic Surveillance in South, A., Gnanakaran, S. & Korber, B. (2021) HIV-1 and SARS-CoV-2: Patterns in the evolution of two pandemic pathogens. *Cell Host Microbe,* 29(7), 1093-1110. 10.1016/j.chom.2021.05.012.
Fitch, W. M. & Ayala, F. J. (1995) Tempo and mode in evolution: genetics and paleontology 50 years after Simpson.
Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M. & Et Al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* 269(5223), 496-512. 10.1126/science.7542800.
Flemming, R. (2023) Pandemics in the Ancient Mediterranean World. *Isis,* 114, S288-S312. 10.1086/726988.
Folgueiras-Gonzalez, A., Van Den Braak, R., Simmelink, B., Deijs, M., Van Der Hoek, L. & De Groof, A. (2020) Atypical Porcine Pestivirus Circulation and Molecular Evolution within an Affected Swine Herd. *Viruses,* 12(10). 10.3390/v12101080.
Fourdan, O. (2011) *Xfce Desktop Environment*. Available at: https://xfce.org/ [2021].
Frank, S. A. (2002) *Immunology and evolution of infectious disease.* Princeton University Press.
Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics,* 28(23), 3150-2. 10.1093/bioinformatics/bts565.
Fuentes, S., Gibbs, A. J., Adams, I. P., Wilson, C., Botermans, M., Fox, A., Kreuze, J., Boonham, N., Kehoe, M. A. & Jones, R. a. C. (2021) Potato Virus A Isolates from Three Continents: Their Biological Properties, Phylogenetics, and Prehistory. *Phytopathology,* 111(1), 217-226. 10.1094/PHYTO-08-20-0354-FI.
Furuno, M., Kasukawa, T., Saito, R., Adachi, J., Suzuki, H., Baldarelli, R., Hayashizaki, Y. & Okazaki, Y. (2003) CDS annotation in full-length cDNA sequence. *Genome Res,* 13(6B), 1478-87. 10.1101/gr.1060303.
Fusté, B. (2012) " Next-generation" Sequencing (NGS): The new genomic revolution. *Capítol del llibre: Handbook of instrumental techniques for materials, chemical*

and biosciences research, Centres Científics i Tecnològics. Universitat de Barcelona, Barcelona, 2012. Part III. Biosciences technologies (BT), BT. 7, 6 p.

Gascuel, O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol,* 14(7), 685-95. 10.1093/oxfordjournals.molbev.a025808.

Gene Ontology, C. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res,* 40(Database issue), D559-64. 10.1093/nar/gkr1028.

Gene Ontology, C., Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., Van Auken, K., Ramsey, J., Siegele, D. A., Chisholm, R. L., Fey, P., Aspromonte, M. C., Nugnes, M. V., Quaglia, F., Tosatto, S., Giglio, M., Nadendla, S., Antonazzo, G., Attrill, H., Dos Santos, G., Marygold, S., Strelets, V., Tabone, C. J., Thurmond, J., Zhou, P., Ahmed, S. H., Asanitthong, P., Luna Buitrago, D., Erdol, M. N., Gage, M. C., Ali Kadhum, M., Li, K. Y. C., Long, M., Michalak, A., Pesala, A., Pritazahra, A., Saverimuttu, S. C. C., Su, R., Thurlow, K. E., Lovering, R. C., Logie, C., Oliferenko, S., Blake, J., Christie, K., Corbani, L., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., Sitnikov, D., Smith, C., Cuzick, A., Seager, J., Cooper, L., Elser, J., Jaiswal, P., Gupta, P., Jaiswal, P., Naithani, S., Lera-Ramirez, M., Rutherford, K., Wood, V., De Pons, J. L., Dwinell, M. R., Hayman, G. T., Kaldunski, M. L., Kwitek, A. E., Lauledерkind, S. J. F., Tutaj, M. A., Vedi, M., Wang, S. J., D'eustachio, P., Aimo, L., Axelsen, K., Bridge, A., Hyka-Nouspikel, N., Morgat, A., Aleksander, S. A., Cherry, J. M., Engel, S. R., Karra, K., Miyasato, S. R., Nash, R. S., Skrzypek, M. S., Weng, S., Wong, E. D., Bakker, E., et al. (2023) The Gene Ontology knowledgebase in 2023. *Genetics,* 224(1). 10.1093/genetics/iyad031.

Goldwasser, M. H. & Letscher, D. (2007) Teaching Object-Oriented Programming in Python. *Iticse 2007: 12th Annual Conference on Innovation & Technology in Computer Science Education*, 365-366.

Goodwin, S., Mcpherson, J. D. & Mccombie, W. R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17(6), 333-51. 10.1038/nrg.2016.49.

Gorlov, I. P., Kimmel, M. & Amos, C. I. (2006) Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. *Hum Mol Genet,* 15(7), 1143-50. 10.1093/hmg/ddl029.

Guido (1990) *Python*. Python documentation. Available at: https://docs.python.org/3/ [2021].

Han, Z., Zhang, Y., Huang, K., Wang, J., Tian, H., Song, Y., Yang, Q., Yan, D., Zhu, S., Yao, M., Wang, X. & Xu, W. (2019) Two Coxsackievirus B3 outbreaks associated with hand, foot, and mouth disease in China and the evolutionary history worldwide. *BMC Infect Dis,* 19(1), 466. 10.1186/s12879-019-4107-z.

Harkins, G. W., Martin, D. P., Christoffels, A. & Varsani, A. (2014) Towards inferring the global movement of beak and feather disease virus. *Virology,* 450-451**,** 24-33. 10.1016/j.virol.2013.11.033.
Harvey, E. & Holmes, E. C. (2022) Diversity and evolution of the animal virome. *Nat Rev Microbiol,* 20(6)**,** 321-334. 10.1038/s41579-021-00665-x.
Hayashi, T., Wills, S., Bussey, K. A. & Takimoto, T. (2015) Identification of Influenza A Virus PB2 Residues Involved in Enhanced Polymerase Activity and Virus Growth in Mammalian Cells at Low Temperatures. *J Virol,* 89(15)**,** 8042-9. 10.1128/JVI.00901-15.
Haydon, D. T., Bastos, A. D., Knowles, N. J. & Samuel, A. R. (2001) Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates. *Genetics,* 157(1)**,** 7-15. 10.1093/genetics/157.1.7.
Heather, J. M. & Chain, B. (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics,* 107(1)**,** 1-8. 10.1016/j.ygeno.2015.11.003.
Heiden, W., Moeckel, G. & Brickmann, J. (1993) A new approach to analysis and display of local lipophilicity/hydrophilicity mapped on molecular surfaces. *J Comput Aided Mol Des,* 7(5)**,** 503-14. 10.1007/BF00124359.
Henquell, C., Mirand, A., Richter, J., Schuffenecker, I., Bottiger, B., Diedrich, S., Terletskaia-Ladwig, E., Christodoulou, C., Peigue-Lafeuille, H. & Bailly, J. L. (2013) Phylogenetic patterns of human coxsackievirus B5 arise from population dynamics between two genogroups and reveal evolutionary factors of molecular adaptation and transmission. *J Virol,* 87(22)**,** 12249-59. 10.1128/JVI.02075-13.
Hicks, A. L. & Duffy, S. (2014) Cell tropism predicts long-term nucleotide substitution rates of mammalian RNA viruses. *PLoS Pathog,* 10(1)**,** e1003838. 10.1371/journal.ppat.1003838.
Hinsen, K. (2000) The molecular modeling toolkit: a new approach to molecular simulations. *Journal of Computational Chemistry,* 21(2)**,** 79-85.
Hirshfield, S. & Ege, R. K. (1996) Object-oriented programming. *Acm Computing Surveys,* 28(1)**,** 253-255. Doi 10.1145/234313.234415.
Ho, S. Y. & Duchene, S. (2014) Molecular-clock methods for estimating evolutionary rates and timescales. *Mol Ecol,* 23(24)**,** 5947-65. 10.1111/mec.12953.
Hoffmann, A. (2013) III. 8. Evolutionary Limits and Constraints. In: *The Princeton guide to evolution.* Princeton University Press.
Holmes, E. C. (2003) Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol,* 11(12)**,** 543-6. 10.1016/j.tim.2003.10.006.
Holmes, E. C., Dudas, G., Rambaut, A. & Andersen, K. G. (2016) The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature,* 538(7624)**,** 193-200. 10.1038/nature19790.
Hueffer, K., Palermo, L. M. & Parrish, C. R. (2004) Parvovirus infection of cells by using variants of the feline transferrin receptor altering clathrin-mediated endocytosis, membrane domain localization, and capsid-binding domains. *J Virol,* 78(11)**,** 5601-11. 10.1128/JVI.78.11.5601-5611.2004.

Huelsenbeck, J. P. & Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17(8), 754-5. 10.1093/bioinformatics/17.8.754.
Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 294(5550), 2310-4. 10.1126/science.1065889.
Hughes, A. L. & Friedman, R. (2000) Evolutionary diversification of protein-coding genes of hantaviruses. *Mol Biol Evol,* 17(10), 1558-68. 10.1093/oxfordjournals.molbev.a026254.
Hughes, A. L. & Hughes, M. A. (2007) More effective purifying selection on RNA viruses than in DNA viruses. *Gene,* 404(1-2), 117-25. 10.1016/j.gene.2007.09.013.
Huynen, M. A. & Bork, P. (1998) Measuring genome evolution. *Proc Natl Acad Sci U S A,* 95(11), 5849-56. 10.1073/pnas.95.11.5849.
International Committee on Taxonomy of Viruses Executive, C. (2020) The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat Microbiol,* 5(5), 668-674. 10.1038/s41564-020-0709-x.
Issaka, S., Traore, O., Longue, R. D. S., Pinel-Galzi, A., Gill, M. S., Dellicour, S., Bastide, P., Guindon, S., Hebrard, E., Dugue, M. J., Sere, Y., Semballa, S., Ake, S., Lemey, P. & Fargette, D. (2021) Rivers and landscape ecology of a plant virus, Rice yellow mottle virus along the Niger Valley. *Virus Evol,* 7(2), veab072. 10.1093/ve/veab072.
Jeffries, C. L., Mansfield, K. L., Phipps, L. P., Wakeley, P. R., Mearns, R., Schock, A., Bell, S., Breed, A. C., Fooks, A. R. & Johnson, N. (2014) Louping ill virus: an endemic tick-borne disease of Great Britain. *J Gen Virol,* 95(Pt 5), 1005-1014. 10.1099/vir.0.062356-0.
Jenkins, G. M., Rambaut, A., Pybus, O. G. & Holmes, E. C. (2002) Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *J Mol Evol,* 54(2), 156-65. 10.1007/s00239-001-0064-3.
Joe Felsenstein (2004) *fdnadist manual*. EMBOSS. Available at: http://bioinf.ibun.unal.edu.co/cgi-bin/emboss/help/fdnadist [2022].
Joshi, M. S., Walimbe, A. M., Arya, S. A. & Gopalkrishna, V. (2023) Evolutionary analysis of all eleven genes of species C rotaviruses circulating in humans and domestic animals. *Mol Phylogenet Evol,* 186, 107854. 10.1016/j.ympev.2023.107854.
Joshi, M. S., Walimbe, A. M., Dilpak, S. P., Cherian, S. S. & Gopalkrishna, V. (2019) Whole-genome-based characterization of three human Rotavirus C strains isolated from gastroenteritis outbreaks in Western India and a provisional intra-genotypic lineage classification system. *J Gen Virol,* 100(7), 1055-1072. 10.1099/jgv.0.001284.
Kanakala, S. & Kuria, P. (2018) Chickpea chlorotic dwarf virus: An Emerging Monopartite Dicot Infecting Mastrevirus. *Viruses,* 11(1). 10.3390/v11010005.
Karsch-Mizrachi, I., Nakamura, Y., Cochrane, G. & International Nucleotide Sequence Database, C. (2012) The International Nucleotide Sequence Database

Collaboration. *Nucleic Acids Res,* 40(Database issue), D33-7. 10.1093/nar/gkr1006.
Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res,* 30(14), 3059-66. 10.1093/nar/gkf436.
Katoh, K. & Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol,* 30(4), 772-80. 10.1093/molbev/mst010.
Kerns, J. A., Emerman, M. & Malik, H. S. (2008) Positive selection and increased antiviral activity associated with the PARP-containing isoform of human zinc-finger antiviral protein. *PLoS Genet,* 4(1), e21. 10.1371/journal.pgen.0040021.
Khan, H. & Khan, A. (2021) Genome-wide population structure inferences of human coxsackievirus-A; insights the genotypes diversity and evolution. *Infect Genet Evol,* 95, 105068. 10.1016/j.meegid.2021.105068.
Kimura, M. (1989) The neutral theory of molecular evolution and the world view of the neutralists. *Genome,* 31(1), 24-31. 10.1139/g89-009.
King, R. W., Scarnati, H. T., Priestley, E. S., De Lucca, I., Bansal, A. & Williams, J. K. (2002) Selection of a thiazole urea-resistant variant of bovine viral diarrhoea virus that maps to the RNA-dependent RNA polymerase. *Antivir Chem Chemother,* 13(5), 315-23. 10.1177/095632020201300507.
Kini, R. M. & Chan, Y. M. (1999) Accelerated evolution and molecular surface of venom phospholipase A2 enzymes. *J Mol Evol,* 48(2), 125-32. 10.1007/pl00006450.
Kishino, H., Thorne, J. L. & Bruno, W. J. (2001) Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol,* 18(3), 352-61. 10.1093/oxfordjournals.molbev.a003811.
Kiss, L., Sebestyen, E., Laszlo, E., Salamon, P., Balazs, E. & Salanki, K. (2008) Nucleotide sequence analysis of peanut stunt virus Rp strain suggests the role of homologous recombination in cucumovirus evolution. *Arch Virol,* 153(7), 1373-7. 10.1007/s00705-008-0120-z.
Knies, J. L., Dang, K. K., Vision, T. J., Hoffman, N. G., Swanstrom, R. & Burch, C. L. (2008) Compensatory evolution in RNA secondary structures increases substitution rate variation among sites. *Mol Biol Evol,* 25(8), 1778-87. 10.1093/molbev/msn130.
Kodama, Y., Mashima, J., Kosuge, T., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y. & Takagi, T. (2018) DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Res,* 46(D1), D30-D35. 10.1093/nar/gkx926.
Koehler, M., Delguste, M., Sieben, C., Gillet, L. & Alsteens, D. (2020) Initial Step of Virus Entry: Virion Binding to Cell-Surface Glycans. *Annu Rev Virol,* 7(1), 143-165. 10.1146/annurev-virology-122019-070025.
Kolakofsky, D. (2015) A short biased history of RNA viruses. *RNA,* 21(4), 667-9. 10.1261/rna.049916.115.

Kosakovsky Pond, S. L., Poon, A. F., Leigh Brown, A. J. & Frost, S. D. (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. *Mol Biol Evol,* 25(9), 1809-24. 10.1093/molbev/msn123.

Kosiol, C., Vinar, T., Da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R. & Siepel, A. (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet,* 4(8), e1000144. 10.1371/journal.pgen.1000144.

Kosuge, T., Mashima, J., Kodama, Y., Fujisawa, T., Kaminuma, E., Ogasawara, O., Okubo, K., Takagi, T. & Nakamura, Y. (2014) DDBJ progress report: a new submission system for leading to a correct annotation. *Nucleic Acids Research,* 42(D1), D44-D49. 10.1093/nar/gkt1066.

Kraberger, S., Harkins, G. W., Kumari, S. G., Thomas, J. E., Schwinghamer, M. W., Sharman, M., Collings, D. A., Briddon, R. W., Martin, D. P. & Varsani, A. (2013) Evidence that dicot-infecting mastreviruses are particularly prone to inter-species recombination and have likely been circulating in Australia for longer than in Africa and the Middle East. *Virology,* 444(1-2), 282-291. 10.1016/j.virol.2013.06.024.

Kumar, S., Tamura, K. & Nei, M. (1994) MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput Appl Biosci,* 10(2), 189-91. 10.1093/bioinformatics/10.2.189.

Kustin, T. & Stern, A. (2021) Biased Mutation and Selection in RNA Viruses. *Mol Biol Evol,* 38(2), 575-588. 10.1093/molbev/msaa247.

Labute, P., Nilar, S. & Williams, C. (2002) A probabilistic approach to high throughput drug discovery. *Comb Chem High Throughput Screen,* 5(2), 135-45. 10.2174/1386207024607329.

Laguette, N. & Benkirane, M. (2015) Shaping of the host cell by viral accessory proteins. *Front Microbiol,* 6, 142. 10.3389/fmicb.2015.00142.

Lahon, A., Maniya, N. H., Tambe, G. U., Chinchole, P. R., Purwar, S., Jacob, G. & Chitambar, S. D. (2013) Group B rotavirus infection in patients with acute gastroenteritis from India: 1994-1995 and 2004-2010. *Epidemiol Infect,* 141(5), 969-75. 10.1017/S0950268812001537.

Lahon, A., Walimbe, A. M. & Chitambar, S. D. (2012) Full genome analysis of group B rotaviruses from western India: genetic relatedness and evolution. *J Gen Virol,* 93(Pt 10), 2252-2266. 10.1099/vir.0.043497-0.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics,* 23(21), 2947-8. 10.1093/bioinformatics/btm404.

Leary, T. P., Muerhoff, A. S., Simons, J. N., Pilot-Matias, T. J., Erker, J. C., Chalmers, M. L., Schlauder, G. G., Dawson, G. J., Desai, S. M. & Mushahwar, I. K. (1996) Sequence and genomic organization of GBV-C: a novel member of the flaviviridae associated with human non-A-E hepatitis. *J Med Virol,* 48(1), 60-7. 10.1002/(SICI)1096-9071(199601)48:1<60::AID-JMV10>3.0.CO;2-A.

Lemey, P., Pybus, O. G., Van Dooren, S. & Vandamme, A. M. (2005) A Bayesian statistical analysis of human T-cell lymphotropic virus evolutionary rates. *Infect Genet Evol,* 5(3), 291-8. 10.1016/j.meegid.2004.04.005.

Li, W.-H. (1997) Molecular evolution. Sinauer Assoc. *Inc.: Sunderland, MA, USA*, 309-334.

Li, Y., Wang, J., Kanai, R. & Modis, Y. (2013) Crystal structure of glycoprotein E2 from bovine viral diarrhea virus. *Proc Natl Acad Sci U S A,* 110(17), 6805-10. 10.1073/pnas.1300524110.

Li, Y., Zeng, Z. J. & Wang, Z. L. (2016) Phylogenetic analysis of the honeybee Sacbrood virus. *Journal of apicultural science,* 60(1), 31-38.

Lim, J. M., Kim, G. & Levine, R. L. (2019) Methionine in Proteins: It's Not Just for Protein Initiation Anymore. *Neurochem Res,* 44(1), 247-257. 10.1007/s11064-017-2460-0.

Lin, Y., Zhang, Y., Li, Q. & Yang, J. Supporting efficient query processing on compressed XML files. *Proceedings of the 2005 ACM symposium on Applied computing*. 660-665.

Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W. & Ray, S. C. (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol,* 73(1), 152-60. 10.1128/JVI.73.1.152-160.1999.

Long, J. S., Howard, W. A., Nunez, A., Moncorge, O., Lycett, S., Banks, J. & Barclay, W. S. (2013) The effect of the PB2 mutation 627K on highly pathogenic H5N1 avian influenza virus is dependent on the virus lineage. *J Virol,* 87(18), 9983-96. 10.1128/JVI.01399-13.

Loverdo, C. & Lloyd-Smith, J. O. (2013) Intergenerational phenotypic mixing in viral evolution. *Evolution,* 67(6), 1815-22. 10.1111/evo.12048.

Lucas, W. (2010) Viral Capsids and Envelopes: Structure and Function. *eLS*.

Luo, R., Delaunay-Moisan, A., Timmis, K. & Danchin, A. (2021) SARS-CoV-2 biology and variants: anticipation of viral evolution and what needs to be done. *Environ Microbiol,* 23(5), 2339-2363. 10.1111/1462-2920.15487.

Luo, W., Roy, A., Guo, F., Irwin, D. M., Shen, X., Pan, J. & Shen, Y. (2020) Host Adaptation and Evolutionary Analysis of Zaire ebolavirus: Insights From Codon Usage Based Investigations. *Front Microbiol,* 11, 570131. 10.3389/fmicb.2020.570131.

Massingham, T. & Goldman, N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics,* 169(3), 1753-62. 10.1534/genetics.104.032144.

Mccarthy, F. M., Mahony, T. J., Parcells, M. S. & Burgess, S. C. (2009) Understanding animal viruses using the Gene Ontology. *Trends Microbiol,* 17(7), 328-35. 10.1016/j.tim.2009.04.006.

Mcdonald, S. M., Nelson, M. I., Turner, P. E. & Patton, J. T. (2016) Reassortment in segmented RNA viruses: mechanisms and outcomes. *Nat Rev Microbiol,* 14(7), 448-60. 10.1038/nrmicro.2016.46.

Medina-Puche, L., Orilio, A. F., Zerbini, F. M. & Lozano-Duran, R. (2021) Small but mighty: Functional landscape of the versatile geminivirus-encoded C4 protein. *PLoS Pathog,* 17(10), e1009915. 10.1371/journal.ppat.1009915.

Miroslaw, P. & Polak, M. P. (2020) Variability of E2 protein-coding sequences of bovine viral diarrhea virus in Polish cattle. *Virus Genes,* 56(4), 515-521. 10.1007/s11262-020-01756-2.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D. & Bateman, A. (2021) Pfam: The protein families database in 2021. *Nucleic Acids Res,* 49(D1), D412-D419. 10.1093/nar/gkaa913.

Moens, U., Krumbholz, A., Ehlers, B., Zell, R., Johne, R., Calvignac-Spencer, S. & Lauber, C. (2017) Biology, evolution, and medical importance of polyomaviruses: An update. *Infect Genet Evol,* 54, 18-38. 10.1016/j.meegid.2017.06.011.

Moncorge, O., Mura, M. & Barclay, W. S. (2010) Evidence for avian and human host cell factors that affect the activity of influenza virus polymerase. *J Virol,* 84(19), 9978-86. 10.1128/JVI.01134-10.

Moreno, H. & Kunz, S. (2021) The Protein Kinase Receptor Modulates the Innate Immune Response against Tacaribe Virus. *Viruses,* 13(7). 10.3390/v13071313.

Murae, M., Shimizu, Y., Yamamoto, Y., Kobayashi, A., Houri, M., Inoue, T., Irie, T., Gemba, R., Kondo, Y., Nakano, Y., Miyazaki, S., Yamada, D., Saitoh, A., Ishii, I., Onodera, T., Takahashi, Y., Wakita, T., Fukasawa, M. & Noguchi, K. (2022) The function of SARS-CoV-2 spike protein is impaired by disulfide-bond disruption with mutation at cysteine-488 and by thiol-reactive N-acetyl-cysteine and glutathione. *Biochem Biophys Res Commun,* 597, 30-36. 10.1016/j.bbrc.2022.01.106.

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K. & Kosakovsky Pond, S. L. (2012) Detecting individual sites subject to episodic diversifying selection. *PLoS Genet,* 8(7), e1002764. 10.1371/journal.pgen.1002764.

National Library of Medicine (1999) *Sample GenBank Record*. National Center of Biotechnology Information. Available at: https://https.ncbi.nlm.nih.gov/genbank/samplerecord/ [2021].

Naveed, M., Makhdoom, S. I., Ali, U., Jabeen, K., Aziz, T., Khan, A. A., Jamil, S., Shahzad, M., Alharbi, M. & Alshammari, A. (2022) Immunoinformatics Approach to Design Multi-Epitope-Based Vaccine against Machupo Virus Taking Viral Nucleocapsid as a Potential Candidate. *Vaccines (Basel),* 10(10). 10.3390/vaccines10101732.

Nawaz-Ul-Rehman, M. S., Liaqat, I., Nahid, N., Saleem, F., Alkahtani, S., Al Qahtani, A., Ye, J. & Mubin, M. (2022) Alternanthera yellow vein virus (AYVV); a betasatellite independent begomovirus infecting Sonchus palustris in Pakistan. *Braz J Biol,* 82, e262248. 10.1590/1519-6984.262248.

Needleman, S. B. & Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol,* 48(3), 443-53. 10.1016/0022-2836(70)90057-4.
Ngoveni, H. G., Van Schalkwyk, A. & Koekemoer, J. J. O. (2019) Evidence of Intragenic Recombination in African Horse Sickness Virus. *Viruses,* 11(7). 10.3390/v11070654.
Nielsen, R. (2005) Molecular signatures of natural selection. *Annu Rev Genet,* 39, 197-218. 10.1146/annurev.genet.39.073003.112420.
Nielsen, R., Bustamante, C., Clark, A. G., Glanowski, S., Sackton, T. B., Hubisz, M. J., Fledel-Alon, A., Tanenbaum, D. M., Civello, D., White, T. J., J, J. S., Adams, M. D. & Cargill, M. (2005) A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol,* 3(6), e170. 10.1371/journal.pbio.0030170.
Nielsen, R. & Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics,* 148(3), 929-36. 10.1093/genetics/148.3.929.
Nigam, D., Latourrette, K., Souza, P. F. N. & Garcia-Ruiz, H. (2019) Genome-Wide Variation in Potyviruses. *Front Plant Sci,* 10, 1439. 10.3389/fpls.2019.01439.
Nylander, J. A., Ronquist, F., Huelsenbeck, J. P. & Nieves-Aldrey, J. L. (2004) Bayesian phylogenetic analysis of combined data. *Syst Biol,* 53(1), 47-67. 10.1080/10635150490264699.
Ohmura, T., Ueda, T., Hashimoto, Y. & Imoto, T. (2001) Tolerance of point substitution of methionine for isoleucine in hen egg white lysozyme. *Protein Eng,* 14(6), 421-5. 10.1093/protein/14.6.421.
Oliphant, T. E. (2007) Python for scientific computing. *Computing in science & engineering,* 9(3), 10-20.
Padhi, A. & Ma, L. (2015) Time-dependent selection pressure on two arthropod-borne RNA viruses in the same serogroup. *Infect Genet Evol,* 32, 255-64. 10.1016/j.meegid.2015.03.019.
Parrella, G. & Lanave, C. (2009) Identification of a new pathotype of Bean yellow mosaic virus (BYMV) infecting blue passion flower and some evolutionary characteristics of BYMV. *Arch Virol,* 154(10), 1689-94. 10.1007/s00705-009-0485-7.
Patino-Galindo, J. A., Filip, I. & Rabadan, R. (2021) Global Patterns of Recombination across Human Viruses. *Mol Biol Evol,* 38(6), 2520-2531. 10.1093/molbev/msab046.
Paxton, R. J., Schafer, M. O., Nazzi, F., Zanni, V., Annoscia, D., Marroni, F., Bigot, D., Laws-Quinn, E. R., Panziera, D., Jenkins, C. & Shafiey, H. (2022) Epidemiology of a major honey bee pathogen, deformed wing virus: potential worldwide replacement of genotype A by genotype B. *Int J Parasitol Parasites Wildl,* 18, 157-171. 10.1016/j.ijppaw.2022.04.013.
Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunic, I.,

Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., Sigrist, C. J. A., Sillitoe, I., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H. & Bateman, A. (2023) InterPro in 2022. *Nucleic Acids Res,* 51(D1), D418-D427. 10.1093/nar/gkac993.

Peck, K. M. & Lauring, A. S. (2018) Complexities of Viral Mutation Rates. *J Virol,* 92(14). 10.1128/JVI.01031-17.

Perez-Losada, M., Arenas, M., Galan, J. C., Bracho, M. A., Hillung, J., Garcia-Gonzalez, N. & Gonzalez-Candelas, F. (2020) High-throughput sequencing (HTS) for the analysis of viral populations. *Infect Genet Evol,* 80, 104208. 10.1016/j.meegid.2020.104208.

Perez-Losada, M., Arenas, M., Galan, J. C., Palero, F. & Gonzalez-Candelas, F. (2015) Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infect Genet Evol,* 30, 296-307. 10.1016/j.meegid.2014.12.022.

Pond, S. L. & Frost, S. D. (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics,* 21(10), 2531-3. 10.1093/bioinformatics/bti320.

Poon, A. F., Frost, S. D. & Pond, S. L. (2009) Detecting signatures of selection from DNA sequences using Datamonkey. *Methods Mol Biol,* 537, 163-83. 10.1007/978-1-59745-251-9_8.

Prechelt, L. (2003) Are scripting languages any good? A validation of Perl, Python, Rexx, and Tcl against C, C++, and Java. *Advances in Computers, Vol 57,* 57, 205-270. Doi 10.1016/S0065-2458(03)57005-X.

Quito-Avila, D. F., Jelkmann, W., Tzanetakis, I. E., Keller, K. & Martin, R. R. (2011) Complete sequence and genetic characterization of Raspberry latent virus, a novel member of the family Reoviridae. *Virus Res,* 155(2), 397-405. 10.1016/j.virusres.2010.11.008.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. (2018) Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol,* 67(5), 901-904. 10.1093/sysbio/syy032.

Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. (2016) Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol,* 2(1), vew007. 10.1093/ve/vew007.

Ramsden, C., Melo, F. L., Figueiredo, L. M., Holmes, E. C., Zanotto, P. M. & Consortium, V. (2008) High rates of molecular evolution in hantaviruses. *Mol Biol Evol,* 25(7), 1488-92. 10.1093/molbev/msn093.

Redondo, N., Zaldivar-Lopez, S., Garrido, J. J. & Montoya, M. (2021) SARS-CoV-2 Accessory Proteins in Viral Pathogenesis: Knowns and Unknowns. *Front Immunol,* 12, 708264. 10.3389/fimmu.2021.708264.

Reil, D., Rosenfeld, U. M., Imholt, C., Schmidt, S., Ulrich, R. G., Eccard, J. A. & Jacob, J. (2017) Puumala hantavirus infections in bank vole populations: host and virus dynamics in Central Europe. *BMC Ecol,* 17(1), 9. 10.1186/s12898-017-0118-z.

Renner, M., Bertinelli, M., Leyrat, C., Paesen, G. C., Saraiva De Oliveira, L. F., Huiskonen, J. T. & Grimes, J. M. (2016) Nucleocapsid assembly in

pneumoviruses is regulated by conformational switching of the N protein. *Elife,* 5, e12627. 10.7554/eLife.12627.
Retief, J. D. (2000) Phylogenetic analysis using PHYLIP. In: *Bioinformatics methods and protocols.* Springer.
Rice, P., Longden, I. & Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet,* 16(6), 276-7. 10.1016/s0168-9525(00)02024-2.
Rodelo-Urrego, M., Garcia-Arenal, F. & Pagan, I. (2015) The effect of ecosystem biodiversity on virus genetic diversity depends on virus species: A study of chiltepin-infecting begomoviruses in Mexico. *Virus Evol,* 1(1), vev004. 10.1093/ve/vev004.
Romano, C. M., Zanotto, P. M. & Holmes, E. C. (2008) Bayesian coalescent analysis reveals a high rate of molecular evolution in GB virus C. *J Mol Evol,* 66(3), 292-7. 10.1007/s00239-008-9087-3.
Rubio, A. E., Abraha, A., Carpenter, C. A., Troyer, R. M., Reyes-Rodriguez, A. L., Salomon, H., Arts, E. J. & Tebit, D. M. (2014) Similar replicative fitness is shared by the subtype B and unique BF recombinant HIV-1 isolates that dominate the epidemic in Argentina. *PLoS One,* 9(4), e92084. 10.1371/journal.pone.0092084.
Sanderson, M. J. (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics,* 19(2), 301-2. 10.1093/bioinformatics/19.2.301.
Sanfacon, H., Wellink, J., Le Gall, O., Karasev, A., Van Der Vlugt, R. & Wetzel, T. (2009) Secoviridae: a proposed family of plant viruses within the order Picornavirales that combines the families Sequiviridae and Comoviridae, the unassigned genera Cheravirus and Sadwavirus, and the proposed genus Torradovirus. *Arch Virol,* 154(5), 899-907. 10.1007/s00705-009-0367-z.
Sanger, F. (2005) Frederick Sanger—Biographical. *Nobelprize. org*, 1-4.
Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M. & Smith, M. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature,* 265(5596), 687-95. 10.1038/265687a0.
Sauquet, H. (2013) A practical guide to molecular dating. *Comptes Rendus Palevol,* 12(6), 355-367. 10.1016/j.crpv.2013.07.003.
Sayers, E. W., Cavanaugh, M., Clark, K., Pruitt, K. D., Schoch, C. L., Sherry, S. T. & Karsch-Mizrachi, I. (2021) GenBank. *Nucleic Acids Res,* 49(D1), D92-D96. 10.1093/nar/gkaa1023.
Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S. & Karsch-Mizrachi, I. (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford),* 2020. 10.1093/database/baaa062.
Sepulveda, J. L. (2020) Using R and Bioconductor in Clinical Genomics and Transcriptomics. *J Mol Diagn,* 22(1), 3-20. 10.1016/j.jmoldx.2019.08.006.

Shafiq, M., Ondrasek, G., Al-Sadi, A. M. & Shahid, M. S. (2023) Molecular Signature of a Novel Alternanthera Yellow Vein Virus Variant Infecting the Ageratum conyzoides Weed in Oman. *Viruses,* 15(12). 10.3390/v15122381.

Shi, W.-F., Zhang, Z., Dun, A.-S., Zhang, Y.-Z., Yu, G.-F., Zhuang, D.-M. & Zhu, C.-D. (2009) Positive selection analysis of VP1 genes of worldwide human enterovirus 71 viruses. *Virologica Sinica,* 24**,** 59-64.

Simon-Loriere, E. & Holmes, E. C. (2011) Why do RNA viruses recombine? *Nat Rev Microbiol,* 9(8)**,** 617-26. 10.1038/nrmicro2614.

Simpson, A. A., Hebert, B., Sullivan, G. M., Parrish, C. R., Zadori, Z., Tijssen, P. & Rossmann, M. G. (2002) The structure of porcine parvovirus: comparison with related viruses. *J Mol Biol,* 315(5)**,** 1189-98. 10.1006/jmbi.2001.5319.

Simpson, G. G. (1945) Tempo and mode in evolution. *Trans N Y Acad Sci,* 8**,** 45-60. 10.1111/j.2164-0947.1945.tb00215.x.

Sironen, T., Vaheri, A. & Plyusnin, A. (2001) Molecular evolution of Puumala hantavirus. *J Virol,* 75(23)**,** 11803-10. 10.1128/JVI.75.23.11803-11810.2001.

Smith, J. M. (1983) The Neutral Theory of Molecular Evolution - Kimura,M. *Nature,* 306(5944)**,** 713-714. DOI 10.1038/306713a0.

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res,* 26(1)**,** 320-2. 10.1093/nar/26.1.320.

Spielman, S. J., Weaver, S., Shank, S. D., Magalis, B. R., Li, M. & Kosakovsky Pond, S. L. (2019) Evolution of Viral Genomes: Interplay Between Selection, Recombination, and Other Forces. *Methods Mol Biol,* 1910**,** 427-468. 10.1007/978-1-4939-9074-0_14.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D. & Birney, E. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res,* 12(10)**,** 1611-8. 10.1101/gr.361602.

Strandin, T., Smura, T., Ahola, P., Aaltonen, K., Sironen, T., Hepojoki, J., Eckerle, I., Ulrich, R. G., Vapalahti, O., Kipar, A. & Forbes, K. M. (2020) Orthohantavirus Isolated in Reservoir Host Cells Displays Minimal Genetic Changes and Retains Wild-Type Infection Properties. *Viruses,* 12(4). 10.3390/v12040457.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J. & Rambaut, A. (2018) Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol,* 4(1)**,** vey016. 10.1093/ve/vey016.

Switzer, W. M., Salemi, M., Shanmugam, V., Gao, F., Cong, M. E., Kuiken, C., Bhullar, V., Beer, B. E., Vallet, D., Gautier-Hion, A., Tooze, Z., Villinger, F., Holmes, E. C. & Heneine, W. (2005) Ancient co-speciation of simian foamy viruses and primates. *Nature,* 434(7031)**,** 376-80. 10.1038/nature03341.

Takahashi, T., Akagawa, M., Kimura, R., Sada, M., Shirai, T., Okayama, K., Hayashi, Y., Kondo, M., Takeda, M., Ryo, A. & Kimura, H. (2023) Molecular evolutionary

analyses of the fusion protein gene in human respirovirus 1. *Virus Res,* 333**,** 199142. 10.1016/j.virusres.2023.199142.
Tamura, K., Stecher, G. & Kumar, S. (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol Evol,* 38(7)**,** 3022-3027. 10.1093/molbev/msab120.
Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences.
Te Velthuis, A. J. (2014) Common and unique features of viral RNA-dependent polymerases. *Cell Mol Life Sci,* 71(22)**,** 4403-20. 10.1007/s00018-014-1695-z.
Thawornwattana, Y., Dong, H. T., Phiwsaiya, K., Sangsuriya, P., Senapin, S. & Aiewsakun, P. (2021) Tilapia lake virus (TiLV): Genomic epidemiology and its early origin. *Transbound Emerg Dis,* 68(2)**,** 435-444. 10.1111/tbed.13693.
Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22(22)**,** 4673-80. 10.1093/nar/22.22.4673.
Thompson, J. R., Kamath, N. & Perry, K. L. (2014) An evolutionary analysis of the Secoviridae family of viruses. *PLoS One,* 9(9)**,** e106305. 10.1371/journal.pone.0106305.
Thorne, L. G., Bouhaddou, M., Reuschl, A. K., Zuliani-Alvarez, L., Polacco, B., Pelin, A., Batra, J., Whelan, M. V. X., Hosmillo, M., Fossati, A., Ragazzini, R., Jungreis, I., Ummadi, M., Rojc, A., Turner, J., Bischof, M. L., Obernier, K., Braberg, H., Soucheray, M., Richards, A., Chen, K. H., Harjai, B., Memon, D., Hiatt, J., Rosales, R., Mcgovern, B. L., Jahun, A., Fabius, J. M., White, K., Goodfellow, I. G., Takeuchi, Y., Bonfanti, P., Shokat, K., Jura, N., Verba, K., Noursadeghi, M., Beltrao, P., Kellis, M., Swaney, D. L., Garcia-Sastre, A., Jolly, C., Towers, G. J. & Krogan, N. J. (2022) Evolution of enhanced innate immune evasion by SARS-CoV-2. *Nature,* 602(7897)**,** 487-495. 10.1038/s41586-021-04352-y.
Toribio, A. L., Alako, B., Amid, C., Cerdeno-Tarraga, A., Clarke, L., Cleland, I., Fairley, S., Gibson, R., Goodgame, N., Ten Hoopen, P., Jayathilaka, S., Kay, S., Leinonen, R., Liu, X., Martinez-Villacorta, J., Pakseresht, N., Rajan, J., Reddy, K., Rosello, M., Silvester, N., Smirnov, D., Vaughan, D., Zalunin, V. & Cochrane, G. (2017) European Nucleotide Archive in 2016. *Nucleic Acids Res,* 45(D1)**,** D32-D36. 10.1093/nar/gkw1106.
Urbino, C., Gutierrez, S., Antolik, A., Bouazza, N., Doumayrou, J., Granier, M., Martin, D. P. & Peterschmitt, M. (2013) Within-host dynamics of the emergence of Tomato yellow leaf curl virus recombinants. *PLoS One,* 8(3)**,** e58375. 10.1371/journal.pone.0058375.
Valley, C. C., Cembran, A., Perlmutter, J. D., Lewis, A. K., Labello, N. P., Gao, J. & Sachs, J. N. (2012) The methionine-aromatic motif plays a unique role in stabilizing protein structure. *J Biol Chem,* 287(42)**,** 34979-34991. 10.1074/jbc.M112.374504.

Varble, A., Albrecht, R. A., Backes, S., Crumiller, M., Bouvier, N. M., Sachs, D., Garcia-Sastre, A. & Tenoever, B. R. (2014) Influenza A virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host Microbe,* 16(5), 691-700. 10.1016/j.chom.2014.09.020.

Venkataraman, S., Prasad, B. & Selvarajan, R. (2018) RNA Dependent RNA Polymerases: Insights from Structure, Function and Evolution. *Viruses,* 10(2). 10.3390/v10020076.

Vihinen, M. (2020) Guidelines for systematic reporting of sequence alignments. *Biol Methods Protoc,* 5(1), bpaa001. 10.1093/biomethods/bpaa001.

Vilar, S., Cozza, G. & Moro, S. (2008) Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr Top Med Chem,* 8(18), 1555-72. 10.2174/156802608786786624.

Vossen, M. T., Westerhout, E. M., Soderberg-Naucler, C. & Wiertz, E. J. (2002) Viral immune evasion: a masterpiece of evolution. *Immunogenetics,* 54(8), 527-42. 10.1007/s00251-002-0493-1.

Wang, H., Cui, X., Cai, X. & An, T. (2022a) Recombination in Positive-Strand RNA Viruses. *Front Microbiol,* 13, 870759. 10.3389/fmicb.2022.870759.

Wang, J. D. (2015) A Study of Comparing the Ambiguity of Existing Virus Taxonomy Structures Using Protein's Region Names in the Vector Space Model. *2015 Ieee Conference on Computational Intelligence in Bioinformatics and Computational Biology (Cibcb)*, 314-321.

Wang, Q., Wang, X., Ding, J., Huang, L. & Wang, Z. (2024) Structural insight of cell surface sugars in viral infection and human milk glycans as natural antiviral substance. *Int J Biol Macromol,* 277(Pt 1), 133867. 10.1016/j.ijbiomac.2024.133867.

Wang, W. & Han, G. Z. (2021) Pervasive Positive Selection on Virus Receptors Driven by Host-Virus Conflicts in Mammals. *J Virol,* 95(20), e0102921. 10.1128/JVI.01029-21.

Wang, Y., Pruitt, R. N., Nurnberger, T. & Wang, Y. (2022b) Evasion of plant immunity by microbial pathogens. *Nat Rev Microbiol,* 20(8), 449-464. 10.1038/s41579-022-00710-3.

Warehouse Project (April 2018). Python Packaging Working Group. Available at: https://pypi.org/project/pyflowchart/ [July 2023].

Weber De Melo, V., Sheikh Ali, H., Freise, J., Kuhnert, D., Essbauer, S., Mertens, M., Wanka, K. M., Drewes, S., Ulrich, R. G. & Heckel, G. (2015) Spatiotemporal dynamics of Puumala hantavirus associated with its rodent host, Myodes glareolus. *Evol Appl,* 8(6), 545-59. 10.1111/eva.12263.

Woelk, C. H. & Holmes, E. C. (2002) Reduced positive selection in vector-borne RNA viruses. *Mol Biol Evol,* 19(12), 2333-6. 10.1093/oxfordjournals.molbev.a004059.

Yadav, R., Chaudhary, J. K., Jain, N., Chaudhary, P. K., Khanra, S., Dhamija, P., Sharma, A., Kumar, A. & Handu, S. (2021) Role of Structural and Non-Structural Proteins

and Therapeutic Targets of SARS-CoV-2 for COVID-19. *Cells,* 10(4). 10.3390/cells10040821.

Yamaguchi-Kabata, Y. & Gojobori, T. (2000) Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J Virol,* 74(9), 4335-50. 10.1128/jvi.74.9.4335-4350.2000.

Yang, Q., Yan, D., Song, Y., Zhu, S., He, Y., Han, Z., Wang, D., Ji, T., Zhang, Y. & Xu, W. (2022) Whole-genome analysis of coxsackievirus B3 reflects its genetic diversity in China and worldwide. *Virology journal,* 19(1), 69.

Yang, Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci,* 13(5), 555-6. 10.1093/bioinformatics/13.5.555.

Yang, Z. (2005) The power of phylogenetic comparison in revealing protein function. *Proc Natl Acad Sci U S A,* 102(9), 3179-80. 10.1073/pnas.0500371102.

Yao, B., Zhang, L., Liang, S. & Zhang, C. (2012) SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. *PLoS One,* 7(9), e45152. 10.1371/journal.pone.0045152.

Zardecki, C., Dutta, S., Goodsell, D. S., Lowe, R., Voigt, M. & Burley, S. K. (2022) PDB-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Sci,* 31(1), 129-140. 10.1002/pro.4200.

Zdobnov, E. M. & Apweiler, R. (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics,* 17(9), 847-8. 10.1093/bioinformatics/17.9.847.

Zwart, L., Potgieter, C. A., Clift, S. J. & Van Staden, V. (2015) Characterising Non-Structural Protein NS4 of African Horse Sickness Virus. *PLoS One,* 10(4), e0124281. 10.1371/journal.pone.0124281.

Chapter 6: Supplementary Materials

The supplementary materials associated with this thesis are available through publicly available repositories, each assigned a DOI, and can be viewed by the following links:

https://doi.org/10.17635/lancaster/researchdata/723

https://doi.org/10.17635/lancaster/researchdata/724

https://doi.org/10.17635/lancaster/researchdata/725

https://doi.org/10.17635/lancaster/researchdata/726

These datasets contain all input and output files used in the methodological analysis, including sequence alignments, configuration files, result logs, and spreadsheets. The contents are organized according to the evolutionary parameters discussed throughout the thesis.

Supplementary Tables:

Table S 1**:** number of hits in each parsed taxonomic Family having 150 hits and more, see Figure 21

Family level name	**Hits Number**	**Percentages %**
Orthomyxoviridae	119924	76.51
Caliciviridae	15196	9.69
Retroviridae	1538	0.98
Paramyxoviridae	1261	0.80
Reoviridae	1250	0.80
Coronaviridae	1115	0.71
Geminiviridae	1081	0.69
Picornaviridae	1050	0.67
Astroviridae	880	0.56
Adenoviridae	838	0.53
Rhabdoviridae	643	0.41
Partitiviridae	640	0.41
Parvoviridae	574	0.37
Circoviridae	496	0.32
Papillomaviridae	487	0.31
Picobirnaviridae	482	0.31
Genomoviridae	454	0.29
Flaviviridae	397	0.25
Potyviridae	364	0.23
Herpesviridae	359	0.23
Phenuiviridae	357	0.23
Botourmiaviridae	337	0.21
Mitoviridae	331	0.21
Peribunyaviridae	326	0.21
Polyomaviridae	311	0.20
Hantaviridae	295	0.19
Poxviridae	280	0.18
Narnaviridae	275	0.18
Totiviridae	263	0.17
Nodaviridae	247	0.16
Betaflexiviridae	212	0.14
Anelloviridae	198	0.13
Tolecusatellitidae	194	0.12
Virgaviridae	192	0.12
Solemoviridae	192	0.12
Iridoviridae	180	0.11
Alphasatellitidae	172	0.11
Tombusviridae	172	0.11
Phycodnaviridae	162	0.10
Arenaviridae	159	0.10

Dicistroviridae	157	0.10
Iflaviridae	155	0.10
others	2549	1.63

Table S 2: number of hits in each alignment dataset Family having more than one hit, see Figure 23.

Family level name	**Hits Number**	**Percentages %**
Reoviridae	63	16.11
Geminiviridae	56	14.32
Potyviridae	23	5.88
Nanoviridae	20	5.12
Peribunyaviridae	16	4.09
Bromoviridae	16	4.09
Phenuiviridae	15	3.84
Picornaviridae	11	2.81
Paramyxoviridae	11	2.81
Alphasatellitidae	11	2.81
Fimoviridae	10	2.56
Amnoonviridae	10	2.56
Flaviviridae	9	2.30
Rhabdoviridae	7	1.79
Tolecusatellitidae	7	1.79
Parvoviridae	7	1.79
Betaflexiviridae	7	1.79
Virgaviridae	7	1.79
Hantaviridae	6	1.53
Circoviridae	5	1.28
Secoviridae	5	1.28
Arenaviridae	5	1.28
Arteriviridae	4	1.02
Iflaviridae	4	1.02
Retroviridae	4	1.02
Polyomaviridae	4	1.02
Nodaviridae	4	1.02
Caliciviridae	3	0.77
Alphaflexiviridae	3	0.77
Closteroviridae	3	0.77
Tombusviridae	3	0.77
Dicistroviridae	3	0.77
Partitiviridae	3	0.77
Tospoviridae	2	0.51

Adenoviridae	2	0.51
Solemoviridae	2	0.51
Birnaviridae	2	0.51
Orthomyxoviridae	2	0.51
Aspiviridae	1	0.26
Matonaviridae	1	0.26
Hepeviridae	1	0.26
Caulimoviridae	1	0.26
others	12	3.07

Table S 3**:** family level number of hits in each alignment dataset with recombinant sequences, see Figure 27**.**

Family level name	**Hits Number**	**Percentages %**
Potyviridae	19	15.32
Geminiviridae	18	14.52
Picornaviridae	8	6.45
Bromoviridae	6	4.84
Betaflexiviridae	5	4.03
Flaviviridae	5	4.03
Peribunyaviridae	5	4.03
Sedoreovirinae	4	3.23
Phenuiviridae	4	3.23
Secoviridae	4	3.23
Parvoviridae	3	2.42
Alphaflexiviridae	3	2.42
Paramyxoviridae	3	2.42
Hantaviridae	3	2.42
Dicistroviridae	3	2.42
Iflaviridae	3	2.42
Nanoviridae	2	1.61
Spinareovirinae	2	1.61
Tombusviridae	2	1.61
Rhabdoviridae	2	1.61
Arteriviridae	2	1.61
Caliciviridae	2	1.61
Retroviridae	2	1.61
Anelloviridae	1	0.81
Circoviridae	1	0.81
Hepeviridae	1	0.81
Closteroviridae	1	0.81
Nodaviridae	1	0.81

Fimoviridae	1	0.81
Tospoviridae	1	0.81
Amnoonviridae	1	0.81
Partitiviridae	1	0.81
Solemoviridae	1	0.81
Astroviridae	1	0.81
Hepadnaviridae	1	0.81
Caulimoviridae	1	0.81
Adenoviridae	1	0.81

Table S 4: family level number of hits in alignments with No recombinant sequences, see Figure 28.

Family level name	**Hits Number**	**Percentages %**
Reoviridae	57	21.43
Geminiviridae	38	14.29
Nanoviridae	18	6.77
Alphasatellitidae	11	4.14
Phenuiviridae	11	4.14
Bromoviridae	10	3.76
Peribunyaviridae	10	3.76
Fimoviridae	9	3.38
Amnoonviridae	9	3.38
Paramyxoviridae	8	3.01
Virgaviridae	7	2.63
Tolecusatellitidae	7	2.63
Rhabdoviridae	5	1.88
Arenaviridae	5	1.88
Polyomaviridae	4	1.50
Circoviridae	4	1.50
Flaviviridae	4	1.50
Potyviridae	4	1.50
Parvoviridae	3	1.13
Hantaviridae	3	1.13
Picornaviridae	3	1.13
Picornaviridae	3	1.13
Birnaviridae	2	0.75
Closteroviridae	2	0.75
Betaflexiviridae	2	0.75
Orthomyxoviridae	2	0.75
Partitiviridae	2	0.75
Retroviridae	2	0.75
Alphatetraviridae	1	0.38

Benyviridae	1	0.38
Matonaviridae	1	0.38
Endornaviridae	1	0.38
Kitaviridae	1	0.38
Togaviridae	1	0.38
Nodaviridae	1	0.38
Tombusviridae	1	0.38
Aspiviridae	1	0.38
Bornaviridae	1	0.38
Filoviridae	1	0.38
Nairoviridae	1	0.38
Tospoviridae	1	0.38
Amalgaviridae	1	0.38
others	7	2.63

Table S 5: segmented viruses species with concordant pattern the first 6 lines for concordant recombinant and the remain concordant non-recombinant, see Table8.

Concordant species	**Number of segments**	**Genus**	**Family**
Pepper huasteco yellow vein virus	2	Begomovirus	Geminiviridae
Sida micrantha mosaic virus	2	Begomovirus	Geminiviridae
Thottopalayam virus	2	Thottimvirus	Hantaviridae
Peanut stunt virus	2	Cucumovirus	Bromoviridae
Broad bean wilt virus	2	Fabavirus	Secoviridae
Wheat yellow mosaic virus	2	Bymovirus	Potyviridae
Nanovirus-like particle	3	unclassified Begomovirus	Alphasatellitidae
Faba bean necrotic stunt virus	3	Nanovirus	Nanoviridae
Milk vetch dwarf virus	3	Nanovirus	Nanoviridae
Pea necrotic yellow dwarf virus	8	Nanovirus	Nanoviridae
Cotton leaf curl Multan virus	2	Begomovirus	Geminiviridae
East African cassava mosaic virus	2	Begomovirus	Geminiviridae
Tomato leaf curl Palampur virus	2	Begomovirus	Geminiviridae
Tomato yellow leaf curl Kanchanaburi virus	2	Begomovirus	Geminiviridae
Tasmanian aquabirnavirus	2	Aquabirnavirus	Birnaviridae
African horse sickness virus	5	Orbivirus	Reoviridae
Bluetongue virus	2	Orbivirus	Reoviridae
Equine encephalosis virus	4	Orbivirus	Reoviridae
Human rotavirus B	10	Rotavirus	Reoviridae

Southern rice black-streaked dwarf virus	3	Fijivirus	Reoviridae
Mammalian orthoreovirus	9	Orthoreovirus	Reoviridae
Piscine orthoreovirus	2	Orthoreovirus	Reoviridae
Apple necrotic mosaic virus	2	Ilarvirus	Bromoviridae
Japanese soil-borne wheat mosaic virus	2	Furovirus	Virgaviridae
Potato mop-top virus	2	pomovirus	Virgaviridae
Vesicular stomatitis Indiana virus	2	Vesiculovirus	Rhabdoviridae
Guanarito mammarena virus	2	Mammarenavirus	Arenaviridae
Lymphocytic choriomeningitis mammarenavirus	2	Mammarenavirus	Arenaviridae
European mountain ash ringspot-associated emaravirus	2	Emaravirus	Fimoviridae
Pigeonpea sterility mosaic emaravirus	2	Emaravirus	Fimoviridae
Rose rosette emaravirus	2	Emaravirus	Fimoviridae
Gamboa virus	2	Orthobunyavirus	Peribunyaviridae
Punta Toro virus	2	Phlebovirus	Phenuiviridae
Rice grassy stunt tenuivirus	5	Tenuivirus	Phenuiviridae
Salmon isavirus	2	Isavirus	Orthomyxoviridae

Table S 6**:** segmented viruses species with discordant pattern with corresponding genus and family, see Table8.

Discordant species	**Number of segments**	**Genus**	**Family**
Faba bean necrotic yellows virus	6	Nanovirus	Nanoviridae
Blainvillea yellow spot virus	2	Begomovirus	Geminiviridae
Melon chlorotic mosaic virus	2	Begomovirus	Geminiviridae
Pedilanthus leaf curl virus	2	Begomovirus	Geminiviridae
Tomato yellow vein streak virus	2	Begomovirus	Geminiviridae
Changuinola virus	7	Orbivirus	Reoviridae
Palyam virus	8	Orbivirus	Reoviridae
Rotavirus C	3	Rotavirus	Reoviridae
Rice black streaked dwarf virus	6	Fijivirus	Reoviridae
Alfalfa mosaic virus	3	Alfamovirus	Bromoviridae
Prunus necrotic ringspot virus	3	Ilarvirus	Bromoviridae
Tobacco streak virus	3	Ilarvirus	Bromoviridae
Tomato chlorosis virus	2	Crinivirus	Closteroviridae
Redspotted grouper nervous necrosi virus	2	Betanodavirus	Nodaviridae
Fig mosaic emaravirus	4	Emaravirus	Fimoviridae
Dobrava-Belgrade orthohantavirus	2	Orthohantavirus	Hantaviridae

Aino virus	5	Orthobunyavirus	Peribunyaviridae
Cache Valley virus	3	Orthobunyavirus	Peribunyaviridae
La Crosse virus	3	Orthobunyavirus	Peribunyaviridae
Phasi Charoen-like phasivirus	3	Phasivirus	Phenuiviridae
Rice stripe tenuivirus	3	Tenuivirus	Phenuiviridae
Tilapia lake virus	9	Tilapinevirus	Amnoonviridae
Norovirus GI	2	Norovirus	Caliciviridae
Strawberry latent ringspot virus	2	Stralarivirus	Secoviridae
Barley yellow mosaic virus	2	Bymovirus	Potyviridae
Hubei mosquito virus 2.	2	N/A	N/A

Table S 7: recombination discordant genera with their five species or less and respective host for each species, see Table10.

Genus	**Species**	**Host**
Circovirus	Porcine circovirus 4	mammalian
	Beak and feather disease virus	avian
	Muscovy duck circovirus	avian
	Duck circovirus	avian
	Porcine circovirus-like virus P1	mammalian
Nanovirus	Faba bean necrotic stunt virus	plant
	Pea necrotic yellow dwarf virus	plant
	Faba bean necrotic yellows virus	plant
	Faba bean necrotic stunt virus	plant
	Milk vetch dwarf virus	plant
Cucumovirus	Cucumber mosaic virus	plant
	Peanut stunt virus	plant
Pegivirus	Human pegivirus 2	mammalian
	Simian pegivirus	mammalian
	Human hepegivirus	mammalian
Morbillivirus	Measles morbillivirus	mammalian
	Peste des petits ruminants virus	mammalian
	Feline morbillivirus	mammalian
	Dolphin morbillivirus	mammalian
Orthorubulavirus	Mumps orthorubulavirus	mammalian
	Human orthorubulavirus 2	mammalian
	Mammalian orthorubulavirus 5	mammalian
Orthohantavirus	Dobrava-Belgrade orthohantavirus	mammalian
	Puumala orthohantavirus	mammalian
Orthotospovirus	Groundnut ringspot virus	Plant
	Capsicum chlorosis virus	plant
Deltapartitivirus	Pepper cryptic virus	plant
	Fig cryptic virus	plant
Enterovirus	Enterovirus A	mammalian
	Enterovirus C	mammalian

	Enterovirus E	mammalian
Kobuvirus	Aichi virus 1	mammalian
	Porcine kobuvirus	mammalian
	Canine kobuvirus	mammalian
Bymovirus	Barley yellow mosaic virus	plant
	Wheat yellow mosaic virus	plant
Tenuivirus	Rice_stripe_tenuivirus.ref_NC_003753.1	plant
	Rice_grassy_stunt_tenuivirus.ref_NC_002324.1	plant
Emaravirus	European mountain ash ringspot-associated emaravirus	plant
	Pigeonpea sterility mosaic emaravirus	plant
	Rose rosette emaravirus	plant
	Fig mosaic emaravirus	plant
Rotavirus	Human rotavirus B	mammalian
	Rotavirus C	mammalian
	Bat rotavirus	mammalian

Table S 8: viral species with highest correlation coefficient values with genus and family levels, see Figure30.

Species	**R value**	**Family**	**Genus**
African horse sickness virus	0.92	Reoviridae	Orbivirus
Ageratum enation alphasatellite	0.97	Alphasatellitidae	Colecusatellite
Ageratum enation virus	0.92	Geminiviridae	Begomovirus
Ageratum yellow leaf curl betasatellite	0.9774	Tolecusatellitidae	Betasatellite
Akabane virus	0.93	Peribunyaviridae	Orthobunyavirus
Alternanthera yellow vein virus	0.9	Geminiviridae	Begomovirus
Barley yellow mosaic virus	0.87	Potyviridae	Bymovirus
Chandipura virus	0.93	Rhabdoviridae	Vesiculovirus
Enterovirus A	0.87	Picornaviridae	Enterovirus
Enterovirus D	0.89	Picornaviridae	Enterovirus
European mountain ash ringspot-associated emaravirus	0.81	Fimoviridae	Emaravirus
Faba bean necrotic stunt virus	0.92	Nanoviridae	Nanovirus
Fowl aviadenovirus C	0.94	Adenoviridae	Aviadenovirus
Garlic virus B	0.91	Alphaflexiviridae	Allexivirus
Guanarito mammarenavirus	0.87	Arenaviridae	Mammarenavirus
Guaroa virus	0.97	Peribunyaviridae	Orthobunyavirus
Human pegivirus 2	0.93	Flaviviridae	Pegivirus
Human respirovirus 1	0.948	Paramyxoviridae	Respirovirus
Human rotavirus B	0.98	Reoviridae	Rotavirus
Kyasanur Forest disease virus	0.89	Flaviviridae	Orthoflavivirus
Mammalian orthoreovirus 3	0.93	Reoviridae	Orthoreovirus
Moroccan watermelon mosaic virus	0.88	Potyviridae	Potyvirus
Nova virus	0.99	Hantaviridae	Mobatvirus
Pea leaf distortion virus	0.99	Geminiviridae	Begomovirus
Pennisetum mosaic virus	0.87	Potyviridae	Potyvirus
Peste des petits ruminants virus	0.96	Paramyxoviridae	Morbillivirus
Porcine reproductive and respiratory syndrome virus	0.85	Arteriviridae	Betaarterivirus
Porcine respirovirus 1	0.85	Paramyxoviridae	Respirovirus

Potato virus M	0.99	Betaflexiviridae	Carlavirus
Rift Valley fever virus	0.89	Phenuiviridae	Phlebovirus
Rotavirus C	0.866	Reoviridae	Rotavirus
Sida micrantha mosaic virus	0.92	Geminiviridae	Begomovirus
Sida mottle Alagoas virus	0.85	Geminiviridae	Begomovirus
Simian pegivirus	0.96	Flaviviridae	Pegivirus
Spondweni virus	0.99	Flaviviridae	Flavivirus
Sudan ebolavirus	0.87	Filoviridae	Ebolavirus
Tasmanian aquabirnavirus	0.89	Birnaviridae	Aquabirnavirus
Thottopalayam virus	0.99	Hantaviridae	Thottimvirus
Tobacco streak virus	0.97	Bromoviridae	Ilarvirus
Tomato chlorotic mottle virus	0.93	Geminiviridae	Begomovirus
Tomato leaf curl Palampur virus	0.84	Geminiviridae	Begomovirus
Tomato mottle mosaic virus	0.94	Virgaviridae	Tobamovirus
Trichodysplasia spinulosa-associated polyomavirus	0.84	Polyomaviridae	Alphapolyomavirus
Wallerfield virus	0.87	N/A	Negevirus

Table S 9: viral species with lowest correlation coefficient values with genus and family levels, see Figure31.

Species	**R value**	**Family**	**Genus**
Alfalfa mosaic virus	-0.3	Bromoviridae	Alfamovirus
Aphid lethal paralysis virus	-0.3	Dicistroviridae	Cripavirus
East African cassava mosaic virus	-0.3	Geminiviridae	Begomovirus
Parrot hepatitis B virus	-0.3	Hepadnaviridae	Avihepadnavirus
Pea seed-borne mosaic virus	-0.3	Potyviridae	Potyvirus
Rice stripe tenuivirus	-0.3	Phenuiviridae	Tenuivirus
Strawberry latent ringspot virus	-0.3	Secoviridae	Stralarivirus
Wheat streak mosaic virus	-0.3	Potyviridae	Tritimovirus
Pepper cryptic virus	-0.26	Partitiviridae	Deltapartitivirus
Jamestown Canyon virus	-0.25	Peribunyaviridae	orthobunyavirus
Duck circovirus	-0.229	Circoviridae	Circovirus
Sweet potato virus	-0.22	Potyviridae	Potyvirus
Alfalfa leaf curl virus	-0.2	Geminiviridae	Capulavirus
Maize rough dwarf virus	-0.2	Reoviridae	Fijivirus
Norovirus GI	-0.2	Caliciviridae	Norovirus
Cassava brown streak virus	-0.16	Potyviridae	Ipomovirus
Tobacco streak virus	-0.16	Bromoviridae	Ilarvirus
Cherry virus A	-0.157	Betaflexiviridae	Capillovirus
Powassan virus	-0.15	Flaviviridae	Flavivirus
Cotton leaf curl Gezira virus	-0.14	Geminiviridae	Begomovirus
Porcine circovirus	-0.13	Circoviridae	Circovirus
Rose rosette emaravirus	-0.11	Fimoviridae	Emaravirus
Grapevine rupestris stem pitting-associated virus	-0.07	Betaflexiviridae	Foveavirus
Opuntia virus	-0.06	Geminiviridae	Opunvirus
Grapevine geminivirus A	-0.05	Geminiviridae	Maldovirus
Cucumber mosaic virus	-0.04	Bromoviridae	Cucumovirus
Pigeonpea sterility mosaic emaravirus	-0.03	Fimoviridae	Emaravirus
Barley yellow mosaic virus	0.02	Potyviridae	Bymovirus

Canine astrovirus	0.026	Astroviridae	Mamastrovirus
Feline foamy virus	0.05	Retroviridae	Felispumavirus
Pea necrotic yellow dwarf virus	0.05	Nanoviridae	Nanovirus
Piscine orthoreovirus	0.05	Spinareoviridae	Orthoreovirus
Peanut mottle virus	0.06	Potyviridae	Potyvirus
Porcine parvovirus	0.06	Parvoviridae	Protoparvovirus
Homalodisca vitripennis reovirus	0.065	Reoviridae	Phytoreovirus
Southern tomato virus	0.069	Amalgaviridae	Amalgavirus
Palyam virus	0.079	Reoviridae	Orbivirus
Grapevine red blotch virus	0.102	Geminiviridae	Grablovirus
African horse sickness virus	0.115	Reoviridae	Orbivirus
Culex flavivirus	0.12	Flaviviridae	Flavivirus
Liao ning virus	0.12	Reoviridae	Seadornavirus
WU Polyomavirus	0.12	Polyomaviridae	Betapolyomavirus
East African cassava mosaic Kenya virus	0.122	Geminiviridae	Begomovirus
Okra leaf curl alphasatellite	0.13	Alphasatellitidae	N/A

Table S 10**:** family level number of hits in alignment datasets with high R values (more than 0.5), see Figure 32.

Family level name	**Hits Number**	**Percentages %**
Reoviridae	40	15.04
Geminiviridae	34	12.78
Nanoviridae	15	5.64
Picornaviridae	13	4.89
Peribunyaviridae	12	4.51
Paramyxoviridae	11	4.14
Potyviridae	11	4.14
Bromoviridae	10	3.76
Flaviviridae	10	3.76
Amnoonviridae	10	3.76
Rhabdoviridae	8	3.01
Fimoviridae	7	2.63
Alphasatellitidae	6	2.26
Hantaviridae	6	2.26
Virgaviridae	5	1.88
Alphaflexiviridae	5	1.88
Phenuiviridae	5	1.88
Tolecusatellitidae	4	1.50
Birnaviridae	3	1.13
Spinareoviridae	3	1.13
Tombusviridae	3	1.13
Arenaviridae	3	1.13

Secoviridae	3	1.13
Retroviridae	3	1.13
Polyomaviridae	2	0.75
Parvoviridae	2	0.75
Circoviridae	2	0.75
Sedoreoviridae	2	0.75
Closteroviridae	2	0.75
Betaflexiviridae	2	0.75
Nodaviridae	2	0.75
Filoviridae	2	0.75
Partitiviridae	2	0.75
Arteriviridae	2	0.75
Solemoviridae	2	0.75
Alphatetraviridae	1	0.38
Benyviridae	1	0.38
Pneumoviridae	1	0.38
Caliciviridae	1	0.38
Iflaviridae	1	0.38
Caulimoviridae	1	0.38
Adenoviridae	1	0.38

Table S 11**:** family level number of hits in alignment datasets with low R values (less than 0.5), see Figure 33.

Family level name	**Hits Number**	**Percentages %**
Geminiviridae	32	13.22
Reoviridae	23	9.50
Potyviridae	20	8.26
Spinareoviridae	16	6.61
Phenuiviridae	13	5.37
Nanoviridae	10	4.13
Flaviviridae	10	4.13
Peribunyaviridae	10	4.13
Alphasatellitidae	9	3.72
Bromoviridae	8	3.31
Parvoviridae	7	2.89
Betaflexiviridae	7	2.89
Hantaviridae	7	2.89
Sedoreoviridae	6	2.48
Polyomaviridae	5	2.07
Circoviridae	4	1.65

Virgaviridae	4	1.65
Caliciviridae	4	1.65
Picornaviridae	4	1.65
Birnaviridae	3	1.24
Fimoviridae	3	1.24
Nairoviridae	3	1.24
Orthomyxoviridae	3	1.24
Dicistroviridae	3	1.24
Secoviridae	3	1.24
Tolecusatellitidae	3	1.24
Anelloviridae	2	0.83
Paramyxoviridae	2	0.83
Arenaviridae	2	0.83
Tospoviridae	2	0.83
Arteriviridae	2	0.83
Iflaviridae	2	0.83
Benyviridae	1	0.41
Hepeviridae	1	0.41
Closteroviridae	1	0.41
Nodaviridae	1	0.41
Amalgaviridae	1	0.41
Partitiviridae	1	0.41
Astroviridae	1	0.41
Hepadnaviridae	1	0.41
Retroviridae	1	0.41
Adenoviridae	1	0.41

Table S 12**:** segmented viruses species with concordant pattern where all segments record "R" value. >0.5, with corresponding genus and family levels, see Table11

Concordant species (R>0.5)	**Number of segments**	**Genus**	**Family**
Apple necrotic mosaic virus	2	Ilarvirus	Bromoviridae
Broad bean wilt virus	2	Fabavirus	Secoviridae
Cotton leaf curl Multan virus	2	Begomovirus	Geminiviridae
Akabane virus	3	Orthobunyavirus	Peribunyaviridae
Bluetongue virus	2	Orbivirus	Reoviridae
Cache Valley virus	3	Orthobunyavirus	Peribunyaviridae
Cotton leaf curl Multan virus	2	Begomovirus	Geminiviridae

European mountain ash ringspot-associated emaravirus	2	Emaravirus	Fimoviridae
Faba bean necrotic stunt virus	4	Nanovirus	Nanoviridae
Fig mosaic emaravirus	3	Emaravirus	Fimoviridae
Guanarito mammarenavirus	2	Mammarenavirus	Arenaviridae
Hubei mosquito virus 2	2	N/A	N/A
Human rotavirus B	10	Rotavirus	Reoviridae
Japanese soil-borne wheat mosaic virus	2	Furovirus	Virgaviridae
Mammalian orthoreovirus	9	Orthoreovirus	Reoviridae
Melon chlorotic mosaic virus	2	Begomovirus	Geminiviridae
Nanovirus-like particle	3	unclassified Begomovirus	Alphasatellitidae
Nova virus	2	Mobatvirus	Mammantavirinae
Peanut stunt virus	2	Cucumovirus	Bromoviridae
Prunus necrotic ringspot virus	3	Ilarvirus	Bromoviridae
Sida micrantha mosaic virus	2	Begomovirus	Geminiviridae
Tasmanian aquabirnavirus	2	Aquabirnavirus	Birnaviridae
Thottopalayam virus	2	Thottimvirus	Hantaviridae
Tilapia lake virus	10	Tilapinevirus	Amnoonviridae
Tomato chlorosis virus	2	Crinivirus	Closteroviridae
Tomato leaf curl Palampur virus	2	Begomovirus	Geminiviridae
Vesicular stomatitis Indiana virus	2	Vesiculovirus	Rhabdoviridae

Table S 13**:** segmented viruses species with concordant pattern where all segments record "R" value <0.5, with corresponding genus and family levels, see Table11.

Concordant species (R<0.5)	**Number of segments**	**Genus**	**Family**
Banana bunchy top virus	3	Babuvirus	Nanoviridae
Blainvillea yellow spot virus	2	Begomovirus	Geminiviridae
Cardamom bushy dwarf virus	2	Babuvirus	Nanoviridae
Crimean-Congo hemorrhagic fever orthonairovirus	2	Orthonairovirus	Nairoviridae
Cucumber mosaic virus	3	Cucumovirus	Bromoviridae
Dobrava-Belgrade orthohantavirus	3	Orthohantavirus	Hantaviridae
Hantaan orthohantavirus	2	Orthohantavirus	Hantaviridae
Infectious bursal disease virus	2	Avibirnavirus	Birnaviridae
Norovirus GI	2	Norovirus	Caliciviridae
Okra leaf curl alphasatellite	2	N/A	Alphasatellitidae
Oropouche virus	2	Orthobunyavirus	Peribunyaviridae

Pedilanthus leaf curl virus	2	Begomovirus	Geminiviridae
Piscine orthoreovirus	7	Orthoreovirus	Reoviridae
Punta Toro virus	3	Phlebovirus	Phenuiviridae
Rose rosette emaravirus	2	Emaravirus	Fimoviridae
Salmon isavirus	3	Isavirus	Orthomyxoviridae
Schmallenberg virus	5	Orthobunyavirus	Peribunyaviridae
Strawberry latent ringspot virus	2	Stralarivirus	Secoviridae
Tomato yellow leaf curl Kanchanaburi virus	2	Begomovirus	Geminiviridae
Tomato yellow vein streak virus	2	Begomovirus	Geminiviridae

Table S 14**:** segmented viruses species with discordant pattern where segments record "R" value <0.5 and >0.5, with corresponding genus and family levels, see Table11.

Discordant species	**Number of segments**	**Genus**	**Family**
African horse sickness virus	8	Orbivirus	Reoviridae
Aino virus	2	Orthobunyavirus	Peribunyaviridae
Alfalfa mosaic virus	3	Alfamovirus	Bromoviridae
Avian orthoreovirus	10	Orthoreovirus	Spinareoviridae
Barley yellow mosaic virus	2	Bymovirus	Potyviridae
Changuinola virus	7	Orbivirus	Reoviridae
East African cassava mosaic Kenya virus	4	Begomovirus	Geminiviridae
Equine encephalosis virus	4	Orbivirus	Reoviridae
Faba bean necrotic yellow virus	6	Nanovirus	Nanoviridae
Gamboa virus	2	Orthobunyavirus	Peribunyaviridae
Infectious pancreatic necrosis virus	2	Aquabirnavirus	Birnaviridae
Jamestown Canyon virus	2	Orthobunyavirus	Peribunyaviridae
Jingmen tick virus	4	Jingmenvirus	Flaviviridae
La Crosse virus	3	Orthobunyavirus	Peribunyaviridae
Lymphocytic choriomeningitis mammarenavirus	2	Mammarenavirus	Arenaviridae
Milk vetch dwarf virus	3	Nanovirus	Nanoviridae
Palyam virus	8	Orbivirus	Reoviridae
Pea necrotic yellow dwarf virus	8	Nanovirus	Nanoviridae
Pepper huasteco yellow vein virus	2	Begomovirus	Geminiviridae
Phasi Charoen-like phasivirus	3	Phasivirus	Phenuiviridae
Pigeonpea sterility mosaic emaravirus	2	Emaravirus	Fimoviridae

Potato mop-top virus	2	pomovirus	Virgaviridae
Redspotted grouper nervous necrosi virus	2	Betanodavirus	Nodaviridae
Rice black streaked dwarf virus	9	Fijivirus	Reoviridae
Rice grassy stunt tenuivirus	5	Tenuivirus	Phenuiviridae
Rice stripe tenuivirus	4	Tenuivirus	Phenuiviridae
Rift Valley fever virus	3	Phlebovirus	Phenuiviridae
Rotavirus C	8	Rotavirus	Reoviridae
Seoul orthohantavirus	4	Orthohantavirus	Hantaviridae
Southern rice black-streaked dwarf virus	4	Fijivirus	Reoviridae
Tobacco streak virus	3	Ilarvirus	Bromoviridae
Tomato severe rugose virus	2	Begomovirus	Geminiviridae

Table S 15: molecular clock discordant genera with their corresponding species and respective host for each species, see Table13.

Genus	**Species**	**Host**
Circovirus	Beak and feather disease virus	avian
	Muscovy duck circovirus	avian
	Porcine circovirus 4	mammalian
	Duck circovirus	avian
	Canine circovirus	mammalian
	Porcine circovirus-like virus P1	mammalian
Nanovirus	Faba bean necrotic stunt virus	plant
	Pea necrotic yellow dwarf virus	plant
	Faba bean necrotic yellows virus	plant
	Milk vetch dwarf virus	plant
Mastrevirus	Panicum streak virus	plant
	Chickpea chlorosis Australia virus	plant
	Paspalum striate mosaic virus	plant
	Sweet potato symptomless virus 1	plant
	Chickpea chlorotic dwarf virus	plant
Aquabirnavirus	Infectious pancreatic necrosis virus	Fish
	Tasmanian aquabirnavirus	Fish
Morbillivirus	Measles morbillivirus	mammalian
	Peste des petits ruminants virus	mammalian
	Feline morbillivirus	mammalian
	Dolphin morbillivirus	mammalian
Rotavirus	Human rotavirus B	mammalian
	Rotavirus C	mammalian
	Bat rotavirus	mammalian
Fijivirus	Rice black streaked dwarf virus	plant
	Southern rice black-streaked dwarf virus	plant
Orthoreovirus	Mammalian orthoreovirus 3	mammalian
	Piscine orthoreovirus	mammalian

Benyvirus	Rice stripe necrosis virus	plant
	Beet necrotic yellow vein virus	plant
Cucumovirus	Peanut stunt virus	plant
	Cucumber mosaic virus	plant
Ilarvirus	Prunus necrotic ringspot virus	plant
	Prune dwarf virus	plant
	Blackberry chlorotic ringspot virus	plant
	Tobacco streak virus	plant
Tobamovirus	Tomato mottle mosaic virus	plant
	Tomato mosaic virus	plant
	Pepper mild mottle virus	plant
	Tomato brown rugose fruit virus	plant
	Cucumber green mottle mosaic virus	plant
Potexvirus	Pepino mosaic virus	plant
	Bamboo mosaic virus	plant
	Citrus yellow vein clearing virus	plant
Carlavirus	Potato virus	plant
	Garlic common latent virus	plant

Table S 16**:** slowest evolving 40 alignments species, genera and family names, see Table14.

Species	**Genus**	**Family**
Puumala orthohantavirus	Orthohantavirus	Hantaviridae
Piscine orthoreovirus	Orthoreovirus	Spinareoviridae
WU Polyomavirus	Betapolyomavirus	Polyomaviridae
European bat 1 lyssavirus	Lyssavirus	Rhabdoviridae
Coxsackievirus B3	Enterovirus	Picornaviridae
Pea necrotic yellow dwarf virus	Nanovirus	Nanoviridae
La Crosse virus	Orthobunyavirus	Peribunyaviridae
Avian orthoreovirus	Orthoreovirus	Spinareoviridae
Grapevine red blotch virus	Grablovirus	Geminiviridae
Cauliflower mosaic virus	Caulimovirus	Caulimoviridae
Tomato chlorosis virus	Crinivirus	Closteroviridae
Dolphin morbillivirus	Morbillivirus	Paramyxoviridae
Peanut stunt virus	Cucumovirus	Bromoviridae
Broad bean wilt virus 2	Fabavirus	Secoviridae
Australian bat lyssavirus	Lyssavirus	Rhabdoviridae
Crimean-Congo hemorrhagic fever orthonairovirus	Orthonairovirus	Nairoviridae
Omsk hemorrhagic fever virus	Orthoflavivirus	Flaviviridae
East African cassava mosaic Kenya virus	Begomovirus	Geminiviridae
Cache Valley virus	Orthobunyavirus	Peribunyaviridae
Palyam virus	Orbivirus	Sedoreoviridae

Rhinovirus C	Enterovirus	Picornaviridae
Rice black streaked dwarf virus	Fijivirus	Spinareoviridae
Gamboa virus	Orthobunyavirus	Peribunyaviridae
Marburg marburgvirus	Orthomarburgvirus	Filoviridae
Mammalian orthorubulavirus 5	Orthorubulavirus	Paramyxoviridae
Cucumber green mottle mosaic virus	Tobamovirus	Virgaviridae
Rice grassy stunt tenuivirus	Tenuivirus	Phenuiviridae
Banana bunchy top virus	Babuvirus	Nanoviridae
African horse sickness virus	Orbivirus	Sedoreoviridae
Jingmen tick virus	Mogiana tick virus	Flaviviridae
Tobacco vein banding mosaic virus	Potyvirus	Potyviridae
Infectious hematopoietic necrosis virus	Novirhabdovirus	Rhabdoviridae
Potato mop-top virus	Pomovirus	Virgaviridae
Mumps orthorubulavirus	Orthorubulavirus	Paramyxoviridae
Bluetongue virus	Orbivirus	Sedoreoviridae
Bell pepper alphaendornavirus	Alphaendornavirus	Endornaviridae
Soybean mosaic virus	Potyvirus	Potyviridae
Spondweni virus	Flavivirus	Flaviviridae
Dobrava-Belgrade orthohantavirus	Orthohantavirus	Hantaviridae
Porcine respirovirus 1	Respirovirus	Paramyxoviridae

Table S 17: fastest evolving 40 alignments species, genera, and family names, see Table14.

Species	**Genus**	**Family**
Enterovirus A	Enterovirus	Picornaviridae
Dasheen mosaic virus	Potyvirus	Potyviridae
GB virus C	Pegivirus	Flaviviridae
Garlic virus D	Allexivirus	Alphaflexiviridae
Faba bean necrotic stunt virus	Nanovirus	Nanoviridae
Thottopalayam virus	Thottimvirus	Hantaviridae
Garlic virus B	Allexivirus	Alphaflexiviridae
Nova virus	Hantaviridae	Hantaviridae
Chicken megrivirus	Megrivirus	Picornaviridae
Enterovirus C	Enterovirus	Picornaviridae
Sida mottle Alagoas virus	Begomovirus	Geminiviridae
Grapevine fanleaf virus	Nepovirus	Secoviridae
Strawberry mottle virus	Sadwavirus	Secoviridae
Nora virus	Orthonoravirus	Noraviridae
Pea leaf distortion virus	Begomovirus	Geminiviridae
Pennisetum mosaic virus	Potyvirus	Potyviridae

Alternanthera yellow vein virus	Begomovirus	Geminiviridae
Tomato leaf curl Sudan virus	Begomovirus	Geminiviridae
Potato virus M	Carlavirus	Betaflexiviridae
Porcine kobuvirus	Kobuvirus	Picornaviridae
Changuinola virus	Orbivirus	Sedoreoviridae
Coxsackievirus A2	Enterovirus	Picornaviridae
Lettuce mosaic virus	Potyvirus	Potyviridae
Porcine reproductive & respiratory syndrome virus	Betaarterivirus	Arteriviridae
Mammalian orthoreovirus 3	Orthoreovirus	Spinareoviridae
Parechovirus A	Parechovirus	Picornaviridae
Nanovirus-like particle	Geminialphasatellitinae	Alphasatellitidae
Salivirus A	Salivirus	Picornaviridae
Banana bunchy top virus	Babuvirus	Nanoviridae
Ageratum enation virus	Begomovirus	Geminiviridae
Bean golden mosaic virus	Begomovirus	Geminiviridae
Fig mosaic emaravirus	Emaravirus	Fimoviridae
Enterovirus D	Enterovirus	Picornaviridae
East African cassava mosaic virus	Begomovirus	Geminiviridae
Pigeonpea sterility mosaic emaravirus 1	Emaravirus	Fimoviridae
Rotavirus C	Rotavirus	Sedoreoviridae
Pepper golden mosaic virus	Begomovirus	Geminiviridae
Tobacco streak virus	Ilarvirus	Bromoviridae
Sugarcane mosaic virus	Potyvirus	Potyviridae
Beak and feather disease virus	Circovirus	Circoviridae

Table S 18: species with Lowest Coefficient of Variation values with corresponding genera and families.

Species	**CofOfVar value**	**Family**	**Genus**
Porcine respirovirus 1	0.052	Paramyxoviridae	Respirovirus
Human pegivirus 2	0.0564	Flaviviridae	Pegivirus
Louping ill virus	0.0644	Flaviviridae	Orthoflavivirus
Australian bat lyssavirus	0.0713	Rhabdoviridae	Lyssavirus
Guanarito mammarenavirus	0.0731	Arenaviridae	Mammarenavirus
Atypical porcine pestivirus 1	0.0784	Flaviviridae	Pestivirus
Nova virus	0.0988	Hantaviridae	Mobatvirus
Prunus virus F	0.1058	Secoviridae	Fabavirus
Changuinola virus	0.1132	Sedoreoviridae	Orbivirus
Bluetongue virus	0.1134	Sedoreoviridae	Orbivirus

Tobacco vein banding mosaic virus	0.1222	Potyviridae	Potyvirus
Canine kobuvirus	0.1309	Picornaviridae	Kobuvirus
Garlic common latent virus	0.1342	Betaflexiviridae	Carlavirus
Bovine foamy virus	0.1397	Retroviridae	Bovispumavirus
Dobrava-Belgrade orthohantavirus	0.1493	Hantaviridae	Orthohantavirus
Moroccan watermelon mosaic virus	0.15	Potyviridae	Potyvirus
Guaroa virus	0.1595	Peribunyaviridae	Orthobunyavirus
Human rotavirus B	0.1668	Sedoreoviridae	Rotavirus
Lymphocytic choriomeningitis mammarenavirus	0.17	Arenaviridae	Mammarenavirus
Potato virus M	0.1718	Betaflexiviridae	Carlavirus
GB virus C	0.1775	Flaviviridae	Pegivirus
Fig mosaic emaravirus	0.18	Fimoviridae	Emaravirus
Mammalian orthoreovirus 3	0.1868	Spinareoviridae	Orthoreovirus
Melon chlorotic mosaic virus	0.1916	Geminiviridae	Begomovirus
Puumala orthohantavirus	0.1944	Hantaviridae	Orthohantavirus
Apple necrotic mosaic virus	0.218	Bromoviridae	Ilarvirus
La Crosse virus	0.24	Peribunyaviridae	Orthobunyavirus
Enterovirus D	0.24	Picornaviridae	Enterovirus
Avian orthoreovirus	0.2434	Spinareoviridae	Orthoreovirus
Alfalfa mosaic virus	0.244	Bromoviridae	Alfamovirus
Tilapia lake virus	0.2478	Amnoonviridae	Tilapinevirus
Potato virus S	0.2492	Betaflexiviridae	Carlavirus
Rice stripe tenuivirus	0.25	Phenuiviridae	Tenuivirus
Rotavirus C	0.2601	Sedoreoviridae	Rotavirus
Teschovirus A	0.27	Picornaviridae	Teschovirus
Mumps orthorubulavirus	0.2754	Paramyxoviridae	Orthorubulavirus
Avian orthoreovirus	0.2761	Spinareoviridae	Orthoreovirus
Tomato mottle mosaic virus	0.2883	Virgaviridae	Tobamovirus
Crimean-Congo hemorrhagic fever orthonairovirus	0.2891	Nairoviridae	Orthonairovirus
Rhinovirus C	0.2944	Picornaviridae	Enterovirus

Table S 19: species with Highest Coefficient of Variation values with corresponding genera and families, see Table 15.

Species	**CofOfVar value**	**Family**	**Genus**
Piscine orthoreovirus	7.3033	Spinareoviridae	Orthoreovirus
Vesicular stomatitis Indiana virus	4.46	Rhabdoviridae	Vesiculovirus
Porcine reproductive and respiratory syndrome virus	3.4	Arteriviridae	Ampobartevirus
Japanese soil-borne wheat mosaic virus	3.2652	Virgaviridae	Furovirus
Cache Valley virus	3.0718	Peribunyaviridae	Orthobunyavirus
Sudan ebolavirus	2.8	Filoviridae	Orthoebolavirus
Tomato leaf curl Sudan virus	2.7	Geminiviridae	Begomovirus
Seoul orthohantavirus	2.6	Hantaviridae	Orthohantavirus
Akabane virus	2.5825	Peribunyaviridae	Orthobunyavirus

Fowl aviadenovirus C	2.5725	Adenoviridae	Aviadenovirus
Equine arteritis virus	2.557	Arteriviridae	Alphaarterivirus
Beet necrotic yellow vein virus	2.4064	Benyviridae	Benyvirus
Onion yellow dwarf virus	2.4012	Potyviridae	Potyvirus
Schmallenberg virus	2.39	Peribunyaviridae	Orthobunyavirus
WU Polyomavirus	2.3871	Polyomaviridae	Betapolyomavirus
Equine encephalosis virus	2.36	Sedoreoviridae	Orbivirus
Garlic virus D	2.3593	Alphaflexiviridae	Allexivirus
Parrot bornavirus 4	2.3275	Bornaviridae	Orthobornavirus
Prunus necrotic ringspot virus	2.3	Bromoviridae	Ilarvirus
Pea necrotic yellow dwarf virus	2.2456	Nanoviridae	Nanovirus
Southern rice black-streaked dwarf virus	2.2429	Spinareoviridae	Fijivirus
Tomato brown rugose fruit virus	2.1752	Virgaviridae	Tobamovirus
Faba bean necrotic stunt virus	2.1398	Nanoviridae	Nanovirus
Canine morbillivirus	2.0235	Paramyxoviridae	Morbillivirus
Trichodysplasia spinulosa-associated polyomavirus	2.016	Polyomaviridae	Alphapolyomavirus
Prune dwarf virus	2.009	Bromoviridae	Ilarvirus
Leek yellow stripe virus	2.008	Potyviridae	Potyvirus
Aichi virus 1	1.9818	Picornaviridae	Kobuvirus
Cotton leaf curl Multan virus	1.9656	Geminiviridae	Begomovirus
Potato mop-top virus	1.9625	Virgaviridae	Pomovirus
Ageratum yellow leaf curl betasatellite	1.9545	Tolecusatellitidae	Betasatellite
Croton yellow vein mosaic betasatellite	1.95	Tolecusatellitidae	Betasatellite
Goose hemorrhagic polyomavirus	1.9398	Polyomaviridae	Gammapolyomavirus
Cotton leaf curl virus	1.9256	Geminiviridae	Begomovirus
Zucchini yellow mosaic virus	1.9	Potyviridae	Potyvirus
Tobacco streak virus	1.88	Bromoviridae	Ilarvirus
Enterovirus A	1.8508	Picornaviridae	Enterovirus
Hubei mosquito virus 2	1.844	N/A	N/A
Turnip curly top virus	1.838	Geminiviridae	Turncurtovirus
Tomato leaf curl Palampur virus	1.8178	Geminiviridae	Begomovirus

Table S 20: family level number of hits in alignments with CofOfVar values 0 to 1, see Figure 34.

Family level name	**Hits Number**	**Percentages %**
Sedoreoviridae	30	16
Spinareoviridae	25	13
Geminiviridae	16	8
Flaviviridae	11	6
Amnoonviridae	9	5
Paramyxoviridae	8	4

Picornaviridae	7	4
Nanoviridae	7	4
Bromoviridae	7	4
Potyviridae	6	3
Hantaviridae	5	3
Peribunyaviridae	5	3
Fimoviridae	5	3
Secoviridae	4	2
Alphasatellitidae	4	2
Arenaviridae	4	2
Rhabdoviridae	4	2
Betaflexiviridae	3	2
Alphaflexiviridae	3	2
Phenuiviridae	3	2
Nairoviridae	2	1
Virgaviridae	2	1
Retroviridae	2	1
Solemoviridae	2	1
Circoviridae	1	1
Birnaviridae	1	1
Closteroviridae	1	1
Tombusviridae	1	1
Filoviridae	1	1
Caulimoviridae	1	1
Anelloviridae	1	1
Polyomaviridae	1	1
Hepeviridae	1	1
Matonaviridae	1	1
Kitaviridae	1	1
Togaviridae	1	1
Nodaviridae	1	1
Aspiviridae	1	1
calciviridae	1	1
iflaviridae	1	1

Table S 21: family level number of hits in alignments with CofOfVar values > 1, see Figure 35.

Family level name	**Hits Number**	**Percentages %**
Geminiviridae	23	15
Nanoviridae	11	7
Potyviridae	10	6
Peribunyaviridae	10	6
Spinareoviridae	8	5
Flaviviridae	7	5
Picornaviridae	7	5
Hantaviridae	6	4
Virgaviridae	5	3
Alphasatellitidae	5	3
Bromoviridae	5	3
Paramyxoviridae	5	3
Parvoviridae	4	3
Polyomaviridae	4	3
Rhabdoviridae	4	3
Phenuiviridae	3	2
Tolecusatellitidae	3	2
Birnaviridae	3	2
Sedoreoviridae	3	2
Alphaflexiviridae	2	1
Tombusviridae	2	1
Secoviridae	2	1
Benyviridae	2	1
Fimoviridae	2	1
Partitiviridae	2	1
Arteriviridae	2	1
Caliciviridae	2	1
Closteroviridae	1	1
Endornaviridae	1	1
Nodaviridae	1	1
Orthomyxoviridae	1	1
Noraviridae	1	1
Circoviridae	1	1
Alphatetraviridae	1	1
bornaviridae	1	1
filoviridae	1	1
Ammnoniviridae	1	1
Solemoviridae	1	1
Adenoviridae	1	1

Table S 22: family level number hits for slow evolving alignments, see Figure 41.

Family level name	**Hits Number**	**Percentages %**
Spinareoviridae	7	23
Polyomaviridae	4	13
Nanoviridae	2	6
Rhabdoviridae	2	6
Peribunyaviridae	2	6
Geminiviridae	1	3
Sedoreoviridae	1	3
Bromoviridae	1	3
Closteroviridae	1	3
Kitaviridae	1	3
Alphaflexiviridae	1	3
Flaviviridae	1	3
Filoviridae	1	3
Paramyxoviridae	1	3
Picornaviridae	1	3
Secoviridae	1	3
Caulimoviridae	1	3
Retroviridae	1	3
Arenaviridae	1	3

Table S 23: family level number hits for moderate evolving alignments, see Figure 42.

Family level name	**Hits Number**	**Percentages %**
Spinareoviridae	17	12
Geminiviridae	16	11
Sedoreoviridae	12	8
Paramyxoviridae	11	8
Flaviviridae	10	7
Peribunyaviridae	10	7
Potyviridae	9	6
Phenuiviridae	6	4
Virgaviridae	6	4
Nanoviridae	5	3
Bromoviridae	4	3
Hantaviridae	3	2
Parvoviridae	3	2
Solemoviridae	3	2
Alphaflexiviridae	2	1
Amnoonviridae	2	1
Nairoviridae	2	1
Nodaviridae	2	1
Picornaviridae	2	1
Secoviridae	2	1
Closteroviridae	1	1
Endornaviridae	1	1
Filoviridae	1	1
Orthomyxoviridae	1	1
Retroviridae	1	1
Rhabdoviridae	1	1
Tombusviridae	1	1
Anelloviridae	1	1
Polyomaviridae	1	1
Alphatetraviridae	1	1
Hepeviridae	1	1
Matonaviridae	1	1
Togaviridae	1	1
betaflexviridae	1	1
Bornaviridae	1	1
Aspiviridae	1	1
Arenaviridae	1	1
Iflaviviridae	1	1
Adenoviridae	1	1

Table S 24: family level number hits for fast evolving alignments, see Figure 43.

Family level name	**Hits Number**	**Percentages %**
Geminiviridae	19	13
Sedoreoviridae	18	13
Nanoviridae	10	7
Spinareoviridae	9	6
Alphasatellitidae	9	6
Amnoonviridae	8	6
Bromoviridae	7	5
Fimoviridae	7	5
Flaviviridae	6	4
Picornaviridae	6	4
Hantaviridae	5	4
Potyviridae	5	4
Rhabdoviridae	5	4
Peribunyaviridae	3	2
Caliciviridae	3	2
Tolecusatellitidae	3	2
Birnaviridae	2	1
Tombusviridae	2	1
Circoviridae	2	1
Benyviridae	2	1
Partitiviridae	2	1
Arteriviridae	2	1
Arenaviridae	1	1
Paramyxoviridae	1	1
Parvoviridae	1	1
Secoviridae	1	1
Virgaviridae	1	1
Betaflexiviridae	1	1

Table S 25: family level number hits for fast evolving alignments, see Figure 44.

Family level name	Hits Number	Percentages %
Picornaviridae	5	23
Geminiviridae	4	18
Alphasatellitidae	2	9
Potyviridae	2	9
Hantaviridae	2	9
Secoviridae	2	9
Nanoviridae	1	5
Sedoreoviridae	1	5
Betaflexiviridae	1	5
Flaviviridae	1	5
Noraviridae	1	5

Table S 26: segmented viruses names with number of concordant and discordant patterns for evolution speed, see Table 17.

Segmented virus name	Number of segments	Concordant	Discordant
African horse sickness virus	6	*	
Apple necrotic mosaic virus	2	*	
Avian orthoreovirus	10		*
Bluetongue virus	2		*
Bovine viral diarrhea virus	2	*	
Broad bean wilt virus 2	2		*
Changuinola virus	5		*
Crimean-Congo hemorrhagic fever orthonairovirus	2	*	
Cucumber mosaic virus	2	*	
East African cassava mosaic Kenya virus	2	*	
European mountain ash ringspot-associated emaravirus	2	*	
Faba bean necrotic yellows virus	3		*
Fig mosaic emaravirus	3	*	
Guanarito mammarenavirus	2		*
Human rotavirus B	9	*	
La Crosse virus	2	*	
Mammalian orthoreovirus 3	9		*
Melon chlorotic mosaic virus	2	*	
Milk vetch dwarf virus	2	*	

Nanovirus-like particle	3	*	
Nova virus	2		*
Peanut stunt virus	2		*
Prunus virus F	2		*
Rice black streaked dwarf virus	5		*
Rotavirus C	6	*	
Tasmanian aquabirnavirus	2		*
Tilapia lake virus	9		*

Table S 27**:** lists discordant genera in evolution speed with their associated species, and the respective host for each species, see Table 19**.**

Genus	**Species**	**Host**
Nanovirus	Faba bean necrotic yellows virus	plant
	Milk vetch dwarf virus	plant
Begomovirus	Melon chlorotic mosaic virus	plant
	East African cassava mosaic Kenya virus	plant
	Mungbean yellow mosaic India virus	plant
	Euphorbia yellow mosaic virus	plant
	Sida micrantha mosaic virus	plant
	Tomato leaf curl Taiwan virus	plant
	Euphorbia leaf curl virus	plant
	Squash leaf curl China virus	plant
	Pepper yellow vein Mali virus	plant
	South African cassava mosaic virus	plant
	Sida mottle Alagoas virus	plant
Aquabirnavirus	Tasmanian aquabirnavirus	Fish
	Infectious pancreatic necrosis virus	Fish
Orbivirus	Bluetongue virus	Mammalian
	Changuinola virus	Mammalian
	African horse sickness virus	Mammalian
	Palyam virus	Mammalian
	Equine encephalosis virus	Mammalian
Orthoreovirus	Avian orthoreovirus	Avian
	Piscine orthoreovirus	Fish
	Mammalian orthoreovirus 3	Mammalian
Cucumovirus	Peanut stunt virus	plant
	Cucumber mosaic virus	plant
Potexvirus	Bamboo mosaic virus	plant
	Pepino mosaic virus	plant
	Citrus yellow vein clearing virus	plant
Carlavirus	Garlic common latent virus	plant
	Potato virus M	plant
	Potato virus S	plant
Orthoflavivirus	Kyasanur Forest disease virus	Mammalian
	Louping ill virus	Mammalian

	Omsk hemorrhagic fever virus	Mammalian
	Tembusu virus	Avian
Pegivirus	Human pegivirus 2	Mammalian
	GB virus C	Mammalian
	Simian pegivirus	Mammalian
Morbillivirus	Peste des petits ruminants virus	Mammalian
	Dolphin morbillivirus	Mammalian
	Feline morbillivirus	Mammalian
Orthorubulavirus	Human orthorubulavirus 2	Mammalian
	Mumps orthorubulavirus	Mammalian
Mammarenavirus	Guanarito mammarenavirus	Mammalian
	Lymphocytic choriomeningitis mammarenavirus	Mammalian
	Lassa mammarenavirus	Mammalian

Table S 28**:** lists top alignments with the highest number of selected sites in SLR run, with number of sites for each alignment and alignment length, see Figure 46.

Species	**Selected sites number**	**Alignment length**
Atypical porcine pestivirus 1	203	3361
Kibale red colobus virus 1	156	5634
Infectious pancreatic necrosis virus	82	1120
Schmallenberg virus	77	1403
Porcine parvovirus	49	3308
Turnip yellows virus	42	4017
Human pegivirus 2	35	3057
Human metapneumovirus	35	4182
Parrot hepatitis B virus	33	2469
Porcine parvovirus 7	31	1146
Mammalian orthorubulavirus 5	30	5311
Human respirovirus 1	27	5747
Human polyomavirus 6	26	2466
Grapevine red blotch virus	25	1087
WU Polyomavirus	25	2546
Tomato chlorosis virus	25	2693
Human hepegivirus	25	3057
Cauliflower mosaic virus	24	2410
Canna yellow streak virus	24	3041
Infectious hematopoietic necrosis virus	23	3470
Bovine viral diarrhea virus 2	21	4072
Feline immunodeficiency virus	19	3282
Equine arteritis virus	18	9303
Goose hemorrhagic polyomavirus	17	2666

Canine morbillivirus	17	4992
Porcine respirovirus 1	17	5728
Chickpea chlorosis Australia virus	16	1330
Alfalfa leaf curl virus	16	1611
Sweet potato leaf curl virus	15	1109
Paspalum striate mosaic virus	15	1381
Porcine parvovirus 6	15	1851
Carp sprivivirus	15	3554
Dolphin morbillivirus	15	4880
Human polyomavirus 1	14	2573
Cherry virus A	14	2806
Turnip mosaic virus	14	3226
Watermelon mosaic virus	14	3298
Opuntia virus 1	12	1069
Grapevine fanleaf virus	12	2284
Tomato brown rugose fruit virus	12	3155
Wheat yellow mosaic virus	11	903

Table S 29**:** alignments with highest Kappa values with their selected sites number and length of alignment, see Figure 48.

Species	**Selected sites number**	**Alignment length**	**Kappa value**
Cache Valley virus	0	1000	49.854083
Tilapia lake virus	0	1550	33.380386
Enterovirus D	1	6586	22.13513
Schmallenberg virus	7	972	21.151355
Akabane virus	0	972	20.546807
Piscine orthoreovirus	0	1071	19.350751
Guaroa virus	2	6600	18.611731
Nudaurelia capensis omega virus	3	1900	17.856207
La Crosse virus	0	950	16.452747
Bluetongue virus	7	2800	15.107787
Human rotavirus B	4	3480	13.980576
Rubella virus	5	9500	13.71153
Parrot bornavirus 4	2	14100	12.981975
Tasmanian aquabirnavirus	9	3300	12.270175
Tobacco vein banding mosaic virus	4	9200	11.899526
Equine encephalosis virus	0	1050	11.87419
Bovine foamy virus	8	17600	11.805897
African horse sickness virus	3	1900	11.76885
Wheat yellow mosaic virus	11	2700	11.703355

Peanut mottle virus	7	9200	11.255764
Jamestown Canyon virus	0	6700	11.184043
Aino virus	0	4200	10.96756
Sudan ebolavirus	10	19500	10.842078
Porcine respirovirus 1	17	17100	10.30115
Fig mosaic emaravirus	0	1050	10.269758
Rift Valley fever virus	0	1530	10.244485
Louping ill virus	6	10242	10.241165
Guanarito mammarenavirus	10	6800	10.239489
Omsk hemorrhagic fever virus	1	10200	10.041709
Rotavirus C	1	3270	10.02947
Crimean-Congo hemorrhagic fever orthonairovirus	7	5053	9.999788
Mammalian orthoreovirus 3	0	2100	9.97049
Tembusu virus	1	10275	9.90453
Measles morbillivirus	10	15600	9.83519
European mountain ash ringspot-associated emaravirus	0	900	9.823991
Barley yellow mosaic virus	0	2670	9.61264
Pea necrotic yellow dwarf virus	0	440	9.187401
Marburg marburgvirus	9	14619	9.089941
Blueberry mosaic associated virus	0	950	9.068464
Human metapneumovirus	35	12564	8.952811
Avian paramyxovirus 4	2	13000	8.731252

Table S 30**:** alignments with lowest Kappa values with their selected sites number and length of alignment.

Species	**Selected sites number**	**Alignment length**	**Kappa value**
Cardamom bushy dwarf virus	2	858	0.68489
South African cassava mosaic virus	1	1695	0.862234
East African cassava mosaic virus	2	3900	1.021513
Ageratum enation alphasatellite	0	900	1.022634
Pepper golden mosaic virus	0	1639	1.038241
Pepper huasteco yellow vein virus	0	1600	1.043518
Porcine hokovirus	4	6300	1.077987
Beet curly top Iran virus	2	2550	1.086134
Cotton leaf curl Gezira virus	11	3300	1.090997
Faba bean necrotic stunt virus	0	354	1.091787
Opuntia virus 1	12	3200	1.099199
Pepper yellow vein Mali virus	9	3246	1.100567

Ageratum conyzoides symptomless alphasatellite	0	900	1.107122
East African cassava mosaic Kenya virus	4	3325	1.140427
Melon chlorotic mosaic virus	2	1600	1.143055
Banana bunchy top virus	3	351	1.167815
Nanovirus-like particle	3	900	1.183554
Chickpea chlorotic dwarf virus	9	3894	1.20599
Tomato leaf curl alphasatellite	2	900	1.210222
Chickpea chlorosis Australia virus	16	3990	1.251175
Cotton leaf curl Multan virus	3	3300	1.258343
Mungbean yellow mosaic virus	0	1600	1.283267
Ageratum yellow vein virus	8	3300	1.328515
Soybean chlorotic blotch virus	2	1800	1.351046
Panicum streak virus	2	2200	1.355706
Milk vetch dwarf virus	1	850	1.356938
Tomato leaf curl Palampur virus	4	3700	1.38035
Goose hemorrhagic polyomavirus	17	7900	1.380719
Beak and feather disease virus	1	1600	1.380912
Turnip curly top virus	0	3400	1.396485
Faba bean necrotic yellows virus	0	850	1.407373
Pepino mosaic virus	3	6362	1.412726
Fig cryptic virus	4	1400	1.429241
Sida micrantha mosaic virus	1	1600	1.449858
Tomato mottle leaf curl virus	7	2800	1.473462
Pea necrotic yellow dwarf virus	1	500	1.475586
Bean golden mosaic virus	0	1640	1.488728
Sweet potato leaf curl virus	15	3300	1.523176
Tomato yellow spot virus	3	1600	1.52957
Cotton leaf curl virus betasatellite	3	350	1.529854
Canine circovirus	3	1719	1.52999

Table S 31**:** alignments with highest overall Omega values with their selected sites number and length of alignment, see Figure 49.

Species	**Selected sites number**	**Alignment length**	**Omega value**
Pea necrotic yellow dwarf virus	2	309	1.109028
Porcine circovirus-like virus P1	3	340	1.044579
Schmallenberg virus	77	4209	0.885172

Chilli leaf curl alphasatellite	7	900	0.719185
Cotton leaf curl betasatellite	2	350	0.604925
Cotton leaf curl virus betasatellite	3	350	0.545121
Tilapia lake virus	0	340	0.519293
Cotton leaf curl Burewala betasatellite	0	350	0.517361
Croton yellow vein mosaic betasatellite	2	340	0.513211
Citrus chlorotic dwarf associated virus	5	3300	0.477948
Fig cryptic virus	4	1400	0.376989
Rice grassy stunt tenuivirus	9	1400	0.349485
Porcine parvovirus 5	6	4700	0.322979
Human polyomavirus 6	26	7300	0.308219
Beet necrotic yellow vein virus	3	684	0.304211
Banana bunchy top virus	3	351	0.292199
Chickpea chlorosis Australia virus	16	3990	0.291264
Grapevine red blotch virus	25	3262	0.285353
Tomato severe rugose virus	9	2853	0.277608
Turnip yellows virus	42	12133	0.277598
Ageratum yellow leaf curl betasatellite	0	400	0.276631
Rose rosette emaravirus	5	1050	0.270273
Melon chlorotic mosaic alphasatellite	5	1350	0.267329
Southern rice black-streaked dwarf virus	3	1600	0.259718
Porcine circovirus 4	1	1500	0.25372
WU Polyomavirus	25	7638	0.250952
Pepper yellow vein Mali virus	9	3246	0.244283
Pepper cryptic virus 1	3	1200	0.236223
Human polyomavirus 7	11	7300	0.235699
Porcine hokovirus	4	6300	0.235564
Sweet potato symptomless virus 1	6	4200	0.228241
Bluetongue virus	7	2800	0.227096
Kibale red colobus virus 1	156	16900	0.223645
Tomato chlorosis virus	25	8000	0.221784
Tomato chlorotic mottle virus	8	2800	0.22177
Tomato leaf curl Sudan virus	7	3303	0.221018
Cotton leaf curl Burewala alphasatellite	0	946	0.220832
Mammalian orthorubulavirus 5	30	15000	0.220061
Paspalum striate mosaic virus	15	4100	0.218218
Tomato mottle leaf curl virus	7	2800	0.216352
Homalodisca vitripennis reovirus	0	1000	0.211773

Table S 32**:** alignments with lowest overall Omega values with their selected sites number and length of alignment.

Species	**Selected sites number**	**Alignment length**	**Kappa value**
Rhinovirus A	2	7743	0.011585
Piscine orthoreovirus	1	3858	0.013165
GB virus C	1	8526	0.016644
Faba bean necrotic stunt virus	1	500	0.021024
Canine kobuvirus	1	7300	0.022247
Punta Toro virus	1	6200	0.025777
Thottopalayam virus	1	3300	0.026871
African horse sickness virus	1	1047	0.027897
Pennisetum mosaic virus	4	9100	0.028997
Rotavirus C	1	3270	0.029663
Fig mosaic emaravirus	1	6800	0.030196
Jingmen tick virus	1	2793	0.031259
Rubella virus	5	9500	0.031687
Aichi virus 1	6	7200	0.031928
Chicken megrivirus	1	8200	0.033037
Wheat streak mosaic virus	1	9100	0.034353
Rice stripe tenuivirus	1	8700	0.034748
Deformed wing virus	5	8600	0.034903
Cache Valley virus	2	6700	0.034953
Akabane virus	1	6700	0.035722
Powassan virus	1	8610	0.037108
Human rotavirus B	1	1150	0.037286
Guaroa virus	2	6600	0.03798
Feline calicivirus	3	7600	0.0387
Oropouche virus	2	4260	0.039013
Palyam virus	1	950	0.042249
Pepino mosaic virus	3	6362	0.042536
Enterovirus D	1	6586	0.042711
Crimean-Congo hemorrhagic fever orthonairovirus	2	11837	0.044993
Omsk hemorrhagic fever virus	1	10200	0.046002
European bat 1 lyssavirus	1	10854	0.04629
Porcine kobuvirus	3	7300	0.047494
Posavirus 1	5	8800	0.048216
Piura virus	3	8800	0.048389
Pea necrotic yellow dwarf virus	1	500	0.049888
Broad bean wilt virus 2	1	3100	0.049999
Avian paramyxovirus 4	2	13000	0.050309
Teschovirus A	5	6700	0.051599
Vesicular stomatitis Indiana virus	2	13800	0.052296

Rice black streaked dwarf virus	3	4393	0.053538
Japanese soil-borne wheat mosaic virus	3	6500	0.054514

Table S 33: family level number hits for alignments with positive selected sites in SLR, see Figure 50.

Family level name	**Hits Number**	**Percentages %**
Geminiviridae	45	17.1
Potyviridae	18	6.8
Spinareoviridae	15	5.7
Flaviviridae	15	5.7
Sedoreoviridae	12	4.6
Phenuiviridae	11	4.2
Alphasatellitidae	10	3.8
Nanoviridae	10	3.8
Bromoviridae	9	3.4
Peribunyaviridae	9	3.4
Parvoviridae	8	3.0
Paramyxoviridae	8	3.0
Picornaviridae	8	3.0
Polyomaviridae	7	2.7
Circoviridae	6	2.3
Betaflexiviridae	5	1.9
Fimoviridae	5	1.9
Virgaviridae	4	1.5
Alphaflexiviridae	4	1.5
Rhabdoviridae	4	1.5
Birnaviridae	3	1.1
Secoviridae	3	1.1
Tolecusatellitidae	3	1.1
Closteroviridae	3	1.1
Nodaviridae	3	1.1
Arenaviridae	3	1.1
Amnoonviridae	3	1.1
Partitiviridae	3	1.1
Anelloviridae	2	0.8
Retroviridae	2	0.8
Benyviridae	2	0.8
Filoviridae	2	0.8

Nairoviridae	2	0.8
Arteriviridae	2	0.8
Solemoviridae	1	0.4
Hepadnaviridae	1	0.4
Caulimoviridae	1	0.4
Alphatetraviridae	1	0.4
Matonaviridae	1	0.4
Endornaviridae	1	0.4
Kitaviridae	1	0.4

Table S 34: family level number hits for alignments with no positive selected sites in SLR, see Figure 51.

Family level name	**Hits Number**	**Percentages %**
Sedoreoviridae	38	21.1
Spinareoviridae	25	13.9
Nanoviridae	14	7.8
Peribunyaviridae	13	7.2
Hantaviridae	11	6.1
Geminiviridae	7	3.9
Phenuiviridae	7	3.9
Amnoonviridae	7	3.9
Picornaviridae	7	3.9
Fimoviridae	5	2.8
Alphasatellitidae	4	2.2
Bromoviridae	4	2.2
Flaviviridae	4	2.2
Rhabdoviridae	4	2.2
Secoviridae	4	2.2
Tolecusatellitidae	4	2.2
Arenaviridae	3	1.7
Potyviridae	3	1.7
Birnaviridae	2	1.1
Aspiviridae	2	1.1
Orthomyxoviridae	2	1.1
Caliciviridae	2	1.1
Virgaviridae	1	0.6
Alphaflexiviridae	1	0.6
Betaflexiviridae	1	0.6
Nairoviridae	1	0.6
Tospoviridae	1	0.6
Dicistroviridae	1	0.6

Iflaviridae	1	0.6
Astroviridae	1	0.6

Table S 35: Gene Ontology terms hits in InterProscan with corresponding function, see Figure 57.

GO Id	**InterProscan hits**	**Function**
GO:0005198	99	structural molecule activity
GO:0003723	82	RNA binding
GO:0003968	78	RNA-dependent RNA polymerase activity
GO:0019028	75	viral capsid
GO:0005524	59	ATP binding
GO:0006260	46	DNA replication
GO:0006351	43	DNA-templated transcription
GO:0006508	38	proteolysis
GO:0016888	37	endodeoxyribonuclease activity
GO:0004197	28	cysteine-type endopeptidase activity
GO:0003724	21	RNA helicase activity
GO:0016032	18	viral process
GO:0019079	17	viral genome replication
GO:0008234	14	cysteine-type peptidase activity
GO:0006396	13	RNA processing
GO:0008174	13	mRNA methyltransferase activity
GO:0080009	13	mRNA methylation
GO:0044423	13	virion component
GO:0046740	13	transport of virus in host, cell to cell
GO:0016818	12	hydrolase activity, acting on acid anhydrides
GO:0018144	12	RNA-protein covalent cross-linking
GO:0019031	12	viral envelope
GO:0003676	11	nucleic acid binding
GO:0003677	10	DNA binding
GO:0004482	10	mRNA (guanine-N7-)-methyltransferase activity
GO:0006370	10	7-methylguanosine mRNA capping
GO:0004386	10	helicase activity
GO:0016779	9	nucleotidyltransferase activity
GO:0018142	9	protein-DNA covalent cross-linking
GO:0019013	9	viral nucleocapsid
GO:0019058	8	viral life cycle

GO:0019048	8	modulation by virus of host process
GO:0030430	8	host cell cytoplasm
GO:0008168	7	methyltransferase activity
GO:0032259	7	methylation
GO:0016020	6	membrane
GO:0060967	6	negative regulation of gene silencing by RNA
GO:0003688	4	DNA replication origin binding
GO:0019069	4	viral capsid assembly
GO:0016070	4	RNA metabolic process
GO:0046983	4	protein dimerization activity
GO:0003725	4	double-stranded RNA binding
GO:0019064	4	fusion of virus membrane with host plasma membrane

Table S 36: Pfam clan names associated with Pfam IDs with their number of hits, see Figure 60. Abbreviations: **P-loop***, refers to the "P-loop containing nucleoside triphosphate hydrolase superfamily."

Pfam domain name	**Hits**	**Pfam ID**	**Pfam Clan**
Geminivirus rep protein central domain	28	PF08283	N/A
Geminivirus Rep catalytic domain	28	PF00799	Rep-like domain
Viral RNA-dependent RNA polymerase	22	PF00680	RNA dependent RNA polymerase
Geminivirus coat protein/nuclear export factor BR1 family	22	PF00844	N/A
Viral (Superfamily 1) RNA helicase	17	PF01443	P-loop*
RNA helicase	16	PF00910	P-loop*
picornavirus capsid protein	15	PF00073	Nucleoplasmin-like/VP (viral coat and capsid proteins) superfamily
Viral methyltransferase	13	PF01660	Viral methyltranferase superfamily
Potyvirus coat protein	12	PF00767	N/A
Peptidase family C4	12	PF00863	Peptidase clan PA
Potyviridae polyprotein	12	PF08440	N/A
RNA dependent RNA polymerase	12	PF00978	RNA dependent RNA polymerase
Geminivirus AL2 protein	12	PF01440	N/A
Helicase conserved C-terminal domain	11	PF00271	P-loop*
Mononegavirales RNA dependent RNA polymerase	10	PF00946	N/A
Mononegavirales mRNA-capping region V	10	PF14318	N/A
Flavivirus DEAD domain	10	PF07652	P-loop*
Helper component proteinase	9	PF00851	Peptidase clan CA
Putative viral replication protein	9	PF02407	Rep-like domain
Protein P3 of Potyviral polyprotein	8	PF13608	N/A
Potyvirus P1 protease	8	PF01577	Peptidase clan PA

Parvovirus coat protein VP2	8	PF00740	ssDNA viruses Nucleoplasmin-like/VP coat superfamily
Geminivirus AL3 protein	8	PF01407	N/A
Bunyavirus RNA dependent RNA polymerase	7	PF04196	N/A
FtsJ-like methyltransferase	7	PF01728	FAD/NAD(P)-binding Rossmann fold Superfamily
DEAD/DEAH box helicase	6	PF00270	P-loop*
Polyomavirus coat protein	6	PF00718	Nucleoplasmin-like/VP (viral coat and capsid proteins) superfamily
mRNA (guanine-7-) methyltransferase (G-7-MTase)	6	PF12803	N/A
Parvovirus non-structural protein NS1	6	PF01057	P-loop*
L protein N-terminus	5	PF15518	PD-(D/E)XK nuclease superfamily
3C cysteine protease (picornain 3C)	5	PF00548	Peptidase clan PA
Peptidase S7 Flavivirus NS3 serine protease	5	PF00949	Peptidase clan PA
Flavivirus RNA-directed RNA polymerase	5	PF20483	RNA dependent RNA polymerase
Flavivirus non-structural Protein NS1	5	PF00948	N/A
Phospholipase A2-like domain	5	PF08398	Phospholipase A2 superfamily
WCCH motif	5	PF03716	N/A
Rep protein catalytic domain like	5	PF08724	Rep-like domain
Paramyxovirus structural protein V/P N-terminus	4	PF13825	N/A
Origin of replication binding protein	4	PF02217	Rep-like domain