Selection Pressure on Surface Exposed Virus Proteins

Sareh Bagherichimeh, The University of Western Ontario

Supervisor: Poon, Art, The University of Western Ontario
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Pathology and Laboratory Medicine
© Sareh Bagherichimeh 2022

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Bioinformatics Commons, Computational Biology Commons, Genetics Commons, Genomics Commons, and the Other Immunology and Infectious Disease Commons

Recommended Citation
https://ir.lib.uwo.ca/etd/9070

This Dissertation/Thesis is brought to you for free and open access by Scholarship@Western. It has been accepted for inclusion in Electronic Thesis and Dissertation Repository by an authorized administrator of Scholarship@Western. For more information, please contact wlswadmin@uwo.ca.
Abstract

Viral infection requires the interaction between virus surface-exposed (SE) proteins and host cell receptors. This can result in an “arms race” that is assumed to drive accelerated rates of evolution, and some well known examples of diversifying selection involve surface proteins (HIV-1 env, influenza hemagglutinin). We conducted a systematic analysis to determine whether this is truly a distinctive feature of SE virus proteins, in comparison to non-SE proteins encoded by the same genomes.

We obtained reference and all neighbour genomes of 52 human viruses from the NCBI Viral Genomes database. The coding sequences (CDS) of each genome extracted by pairwise alignment against the reference CDSs, and labeled as SE or non-SE using the Gene Ontology database and the transmembrane predictor TMbed. After generating a codon-aware multiple sequence alignments, we used FUBAR to estimate the joint probability distribution over 20 non-synonymous and synonymous rates for each alignment (the evolutionary fingerprint). We calculated the cosine distance between every pair of fingerprints and visualized the results using PCA.

In total, we analyzed 670 sets of homologous genes (125 of which were SE) from 21 virus families. We found no clear separation of SE from non-SE labels by PCA. Additionally, there were no significant differences between SE and non-SE genes in the codon site-specific mean dN/dS ratios, dN–dS differences, dN or dS independently, or the percentage of positive and/or negatively selected sites (Wilcoxon rank sum test, \( p < 0.05 \)).

In closing, we did not find evidence that human virus genes encoding surface-exposed virus proteins undergo higher rates of adaptation than other protein-coding regions in the viral genome.
Lay Summary

Rapid evolution of viruses makes controlling virus infections extremely challenging. Therefore, it is important to further our understanding of how viruses evolve, and how this is shaped by the interaction between a virus and its host. For instance, the host immune system often targets the surface exposed (SE) components of a virus. As a result, there is an ongoing host-virus arms race where the SE components of virus proteins are constantly changing to escape detection by the host immune system, which in turn is constantly adapting to recognise and bind the virus proteins. A recent systematic study by Wang et al. (2020) found evidence of elevated rates of evolution in the host receptor proteins used by viruses, and other proteins expressed on the surface of host cells that directly interact with viruses, in primate genomes.

In this study, we examined whether the SE proteins of viruses are also under elevated rates of evolution. My aim is to conduct a systematic analysis of the evolutionary patterns in human virus genomes, by measuring the selective pressures acting on protein coding genes, and comparing these estimates between SE and non-SE virus proteins. I hypothesize that the genes encoding SE virus proteins are under higher rates of evolution than other protein-coding regions in the viral genome.

After examining genome sequences of 52 human viruses from the National Center for Biotechnology Information (NCBI) with a total of 670 genes (125 SE) belonging to 21 different virus families, we found no difference in the evolution rate of SE and non-SE proteins.
Acknowledgements

Above all I want to thank my supervisor Dr. Art Poon, first for giving me such an incredible learning opportunity. I have so much respect for the depth of your knowledge and skills and so grateful for the time you take to share it.

I’d also like to thank my committee members Dr. Prodger and Dr. Troyer for their insight and guidance. And to the members of the Poon Lab for all their support. Especially Mr. Emmanuel Wong and Ms. Laura Muñoz-Baena for their patience as they helped me understand or debug code.

I would also like to thank Ruth and Bob Jones for their continued encouragement and support and the care they always take editing my writing.

I’m very grateful to my friends Mona, Asa, Cristina and Arsalan, for keeping me sane during COVID lockdowns and over 2 years of working from home. And to my eucalyptus tree who acted as my ‘rubber duck’ for rubberducking. Finlay, my sister for sharing her playlists that I worked to and my parents for their endless support.

Dedicated to the memory of my late grandfather, Gholamreza Bagherichimeh. You installed a love and respect for nature in me. My favorite childhood memories are running around your backyard, building my tree house, fires, and gardening.

Keywords: virus evolution, comparative genomics, natural selection
# Contents

Abstract i

Abstract ii

Acknowledgements iii

List of Figures ix

List of Tables xiv

List of Appendices xv

1 Introduction 1

1.1 Viruses ........................................ 1

1.1.1 Polyproteins .................................. 2

1.1.2 Overlapping reading frames .................. 2

1.1.3 Virus Classifications ......................... 3

Dna and RNA viruses ................................ 4

Enveloped and non-enveloped viruses ................. 6

1.1.4 Surface exposed virus proteins ................ 7

1.2 Natural Selection ................................ 9

1.2.1 Host-Virus Arms Race ........................ 9

The Host in the Evolutionary Arms Race ............. 11

The Virus in the Evolutionary Arms Race ........... 12

1.3 Selection Analysis ............................... 13

1.3.1 Measuring selection from genetic sequences 13

1.3.2 Modeling variation in dN/dS ................... 15

1.3.3 FUBAR ........................................ 16

1.3.4 Gene selection analysis ...................... 17

Codon site selection analysis ....................... 17
<table>
<thead>
<tr>
<th>Evolutionary Fingerprinting of Genes</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.3.5 Assumptions in selection analysis</td>
<td>19</td>
</tr>
<tr>
<td>1. Population evolution</td>
<td>19</td>
</tr>
<tr>
<td>2. Recombination</td>
<td>19</td>
</tr>
<tr>
<td>3. Codon frequencies</td>
<td>19</td>
</tr>
<tr>
<td>4. Base composition</td>
<td>20</td>
</tr>
<tr>
<td>5. Alignment quality</td>
<td>20</td>
</tr>
<tr>
<td>6. Substitution models of codon evolution</td>
<td>20</td>
</tr>
<tr>
<td>7. Overlapping reading frames</td>
<td>21</td>
</tr>
<tr>
<td>1.4 Rationale and Hypothesis</td>
<td>21</td>
</tr>
</tbody>
</table>

## 2 Methods

### 2.1 Method Overview

### 2.2 Data processing

#### 2.2.1 Data Collection

#### 2.2.2 Downsampling neighbour genomes

#### 2.2.3 Obtaining Coding Sequences

#### 2.2.4 Remove Overlapping Regions

### 2.3 Identifying Surface Exposed (SE) Genes

#### 2.3.1 Using Gene Ontology IDs

#### 2.3.2 Using a Transmembrane Protein Predictor

### 2.4 Selection Analysis

#### 2.4.1 Average \(dN/dS\) ratio

#### 2.4.2 Percentage of Positively & Negatively selected sites

#### 2.4.3 Evolutionary Fingerprint

### 2.5 Data visualization

## 3 Results

### 3.1 Virus genome data collection

#### 3.1.1 Obtaining coding sequences from Genome

#### 3.1.2 Distinguishing surface exposed genes

### 3.2 Selection analysis

### 3.3 Evolutionary fingerprints

## 4 Discussion

### 4.1 Conclusions

### 4.2 Limitations
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.2.1 Codon Level Selection</td>
<td>60</td>
</tr>
<tr>
<td>4.2.2 Variable proportion of SE genes exposed to the outside of the virion</td>
<td>60</td>
</tr>
<tr>
<td>4.2.3 Functional Constraints</td>
<td>61</td>
</tr>
<tr>
<td>4.2.4 Other sources of diversifying selection</td>
<td>62</td>
</tr>
<tr>
<td>4.3 Future directions</td>
<td>63</td>
</tr>
<tr>
<td>Bibliography</td>
<td>63</td>
</tr>
<tr>
<td>A  Supplementary Figures and Tables</td>
<td>79</td>
</tr>
</tbody>
</table>
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Application program interface</td>
</tr>
<tr>
<td>APOBEC3G</td>
<td>Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G</td>
</tr>
<tr>
<td>BC</td>
<td>Baltimore classification</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding sequence</td>
</tr>
<tr>
<td>DENV</td>
<td>Dengue virus</td>
</tr>
<tr>
<td>dN</td>
<td>Non-synonymous substitution</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>ds</td>
<td>Double-stranded</td>
</tr>
<tr>
<td>FUBAR</td>
<td>Fast Unconstained Bayesian Approximation</td>
</tr>
<tr>
<td>GLM</td>
<td>Generalized linear model</td>
</tr>
<tr>
<td>GOID</td>
<td>Gene Ontology ID</td>
</tr>
<tr>
<td>GTR</td>
<td>General time reversible</td>
</tr>
<tr>
<td>HBV</td>
<td>Hepatitis B viruses</td>
</tr>
<tr>
<td>HIV</td>
<td>Human immunodeficiency virus</td>
</tr>
<tr>
<td>IQR</td>
<td>Interquartile range</td>
</tr>
<tr>
<td>JEV</td>
<td>Japanese encephalitis virus</td>
</tr>
<tr>
<td>JSON</td>
<td>JavaScript Object Notation</td>
</tr>
<tr>
<td>MAFFT</td>
<td>Multiple alignment using fast Fourier transform</td>
</tr>
<tr>
<td>MSA</td>
<td>Multiple sequence alignment</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>OvRFs</td>
<td>Overlapping reading frames</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal component analysis</td>
</tr>
<tr>
<td>RefSeq</td>
<td>Reference sequence</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>SE</td>
<td>Surface Exposed</td>
</tr>
<tr>
<td>SFTS</td>
<td>Severe fever with thrombocytopenia syndrome</td>
</tr>
<tr>
<td>ss</td>
<td>Single Stranded</td>
</tr>
<tr>
<td>t-SNE</td>
<td>t-distributed stochastic neighbor embedding</td>
</tr>
<tr>
<td>TSV</td>
<td>Tab-separated values</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Different typologies of overlapping genes [108]. Each arrow represents an open reading frame (ORF). The arrow head indicates the direction of the reading frame. Overlap interactions can either be partial, with only a portion of each sequence overlapping at their ends, or nested, where one ORF is completely contained within another. Overlaps can also have different topologies: between sequences on the same strand (unidirectional), divergent, or convergent (between sequences on opposite strands). ............................................. 3

1.2 A Rotavirus A virion with capsid proteins labelled on the left, and with the different layers of the capsid shown in the diagram on the right. This image was generated by ViralZone (Swiss Institute of Bioinformatics, SIB) [44] for its web resource (https://viralzone.expasy.org), and was released into the public domain under a Creative Commons Attribution Share-Alike 4.0 license (https://commons.wikimedia.org/wiki/File:Rotavirus_virion.jpg). ................................. 8

1.3 Schematic illustration of an evolutionary arms race between host and virus species. ................................. 10

2.1 Cutting out a coding sequence (CDS) from a neighbour (query) genome by pairwise alignment to a homologous CDS in the reference genome. An annotated CDS from the reference genome (‘Ref’, red) is aligned pairwise against a query genome (dark blue) and the aligned region is cut out of the query genome and identified as a query CDS homologous to the ref CDS used in the alignment 27

2.2 Removing overlapping regions in the genome while respecting the codon boundaries of the target coding sequence. RefSeq genes A and B are an example of partially overlapping ORFs. In this case, the overlapping region from either the beginning or the end of each gene was removed at the respective codon boundaries. RefSeq genes C and D represent a nested overlap. Here, gene D is discarded as all of its codons are overlapping sites. The non-overlapping sections of Gene C are truncated at the codon boundaries and concatenated together into a new sequence. ................................. 29
2.3 Schematic depiction of the workflow for identifying surface-exposed genes. Genes were separated by their association with enveloped or non-enveloped viruses, based on Gene Ontology terms in the UNIPROT database (top layer). For enveloped viruses, genes lacking UNIPROT entries were identified using a transmembrane protein predictor (TMbed). Genes lacking both UNIPROT entries and TMbed predictions were manually annotated. For non-enveloped viruses, genes lacking UNIPORT entries were also manually annotated. 30

2.4 Diagram summarizing the workflow for the selection analysis for each gene. The multiple sequence alignment (MSA) and phylogenetic tree of the sequences for each gene are fed into FUBAR, which calculates the dN and dS values for each codon site, as well as the evolutionary fingerprint. 32

3.1 Pie charts depicting the distribution of all 21 virus families. The inner pie chart is based on the relative frequencies of 52 viruses distributed among the virus families, while the outer pie chart shows the relative frequencies of 670 genes among virus families for all data in this study. 36

3.2 Overlapping histograms representing the distributions of the log-transformed lengths (number of codons) of the 670 gene sequences in this study. We use colour to distinguish between the distributions for gene lengths with overlapping regions (blue), and the lengths after removing the overlapping regions (red), as described in Methods. 38

3.3 Stacked barplots summarizing the distribution of surface-exposed (SE) and non-surface exposed (non-SE) genes with respect to enveloped and non-enveloped viruses (A) and nucleic acid (B). 39

3.4 Stripchart summarizing the average dN/dS ratios for all genes from non-enveloped viruses, grouped by virus family. Every point represents a gene, coloured with respect to being surface-exposed (SE, red) and non-surface-exposed (non-SE, blue), respectively. Each row (strip) corresponds to a specific virus as labelled on the left, and the respective virus family labels are on the right. 41

3.5 Stripchart summarizing the average dN/dS ratios for all genes from enveloped viruses, grouped by virus family. Every point represents a gene, coloured with respect to being surface-exposed (SE, red) and non-surface-exposed (non-SE, blue), respectively. Each row (strip) corresponds to a specific virus as labelled on the left, and the respective virus family labels are on the right. Vertical line segments are drawn at dN/dS=1 (neutral evolution) for reference. 42
3.6 Box-and-whisker plots summarizing the distributions of (A) dN/dS ratio, (B) the difference in dN and dS, and individual (C) dN and (D) dS rate estimates across all viruses. Each point represents a gene, which are stratified into non-surface exposed (non-SE) and surface-exposed (SE) proteins. Thick line segments indicate the median for each distribution, and the upper and lower limits of the box represent the inter-quartile ranges. The \( p \)-values obtained from two-sample Wilcoxon rank sum tests of differences in group means are displayed for each plot.

3.7 Distributions of mean dN/dS estimates for genes stratified by virus types. (A) Stacked barplot displaying the numbers of genes with respect to nucleic acid (RNA and DNA) and non-enveloped or enveloped virus. (B) Box-and-whisker plot comparing mean dN/dS for genes in non-enveloped (indigo) and enveloped (cyan) viruses. (C) Box-and-whisker plot comparing mean dN/dS for genes in RNA (red) and DNA (blue) viruses.

3.8 Box-and-whisker plots comparing dN/dS ratios for SE and non-SE genes across categories. Genes are separated into four groups based on whether they belong to an enveloped or non-enveloped virus, and whether the genome is DNA or RNA. We report the \( P \) value for Wilcoxon rank sum tests comparing SE and non-SE genes in each category above each set of plots. The \( n \) value indicates the total number of SE and non-SE genes in each category.

3.9 Box-and-whisker plots summarizing the percentage of (A) negative sites, (B) positive sites and (C) sum of positive and negative sites across all genes. For each codon site, if the probability that dN is greater than dS is > 0.8, it was identified as a positively selected site and if probability that dS is greater than dN is > 0.8 as a negatively selected site. Percentage across genes were obtained by averaging out over the total number of codon sites. Each point represents a gene, which are stratified into non-surface exposed (non-SE) and surface exposed (SE) proteins.

3.10 Box-and-whisker plots summarizing the distributions of (A) dN/dS ratio, (B) the difference in dN and dS, and individual (C) dN and (D) dS rate estimates across all viruses. Each point represents a gene, which are stratified into non-surface exposed (non-SE) and surface-exposed (SE) proteins. Thick line segments indicate the median for each distribution, and the upper and lower limits of the box represent the inter-quartile ranges. The \( p \)-values obtained from two-sample Wilcoxon rank sum tests of differences in group means are displayed for each plot.
3.11 Box-and-whisker plots summarizing the distributions of (A) dN/dS ratio, (B) the difference in dN and dS, and individual (C) dN and (D) dS rate estimates across all viruses. This supplements the plot of dN/dS distributions in Figure 3.7. Each point represents a gene, which are stratified into enveloped (cyan) and non-enveloped (indigo) viruses. Thick line segments indicate the median for each distribution, and the upper and lower limits of the box represent the inter-quartile ranges. The *p*-values obtained from two-sample Wilcoxon rank sum tests of differences in group means are displayed for each plot. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ..
4.1 Flaviviruses (Zika virus) transmembrane proteins [76] (A) Number of transmembrane Flavivirus proteins including the envelope glycoprotein (E) the membrane protein (M) and the nonstructural proteins NS1 and NS2. (b) The E protein and its transmembrane, internal and external regions. .......................... 62

A.1 Measles virus heatmap ......................................................... 84
A.2 Scatterplots of genes distributed with respect to the first two principal components, coloured by virus family. .................................................. 85
A.3 Scatterplots of genes distributed with respect to the first two principal components, coloured by virus family. .................................................. 86
1.1 Current definition of Baltimore classifications of viruses, with examples of virus families. ................................................................. 5

3.1 Summary of generalized linear model (GLM) analyses of the mean dN/dS across genes. The basic gamma GLM is reported as ‘fixed effects’, and a second model with a random intercept for genes grouped by virus is reported as ‘mixed effects’. To interpret the coefficient estimates (‘effect’), we note that the gamma regression model takes the form $E(y) = (\alpha + \beta_1 x_1 + \ldots + \beta_n x_n)^{-1}$, where $E(y)$ is the predicted dN/dS, $\alpha$ is the intercept, $x_i$ is the $i$-th predictor variable and $\beta_i$ is the $i$-th coefficient. ......................................................... 47

A.1 List of viruses being studied. ...................................................... 80
List of Appendices

Appendix A Supplementary Figures and Tables ............................ 79
Chapter 1

Introduction

1.1 Viruses

Viruses are non-living, obligate intracellular parasites. As they do not have their own metabolic systems, viruses infect living cells in order to use the energetic resources and components of the host cell to replicate. The hijacking of the host cell by a virus takes away resources from the host. For some viruses, it leads to the destruction of the infected cells, which can have a deleterious effect on the host.

A virion is the extracellular form of a virus that is spread from one cell to another, either within or between hosts. It consists of genetic material in the form of DNA or RNA, packaged inside a protein coat for protection (the nucleocapsid). The capsid is a structure (often symmetrical) composed of one or multiple proteins. Viruses are capable of using a small number of genes to encode large protective capsid layers due to the symmetry and repeated pattern in their structure [114].

Viruses are extremely diverse in size and structure. One of the smallest known viruses is the Porcine circovirus type 1, which has a genome of 1,759 nucleotides and a capsid diameter of 17 nanometres (nm) [30]. On the other hand, one of the largest reported viruses, Mimivirus, has a diameter of about 400 nm and a genome of over 1.2 million nucleotides [83]. In terms
of structure, viruses can be spherical like influenza virus, helical (e.g., tobacco mosaic virus), polyhedral (e.g., adenoviruses), or have more complex shapes, such as bacteriophages.

1.1.1 Polyproteins

Virus genomes often encode polyproteins in a single open reading frame. Polyproteins can be converted into multiple mature peptides by getting proteolytically cleaved by host- or virus-encoded proteases. The resulting mature peptides can be involved in different virus functions [57, 8]. For example, a large fraction of the proteome of SARS coronavirus is derived from two large polyproteins. These polyproteins are encoded in two open reading frames (ORF1a and ORF1ab), which are proteolytically cleaved and processed into 16 nonstructural and enzymatic (e.g., RNA dependent RNA polymerase) proteins [42, 112]. HIV-1 genome also consists of polyproteins; the three major genes gag, pol and env encode polyproteins that are cleaved into multiple peptides that become different proteins following maturation [32]. In GenBank, these polyproteins are often annotated as a single coding sequence (CDS) feature that is further annotated to identify mature peptides.

1.1.2 Overlapping reading frames

Overlapping reading frames (OvRFs) are sections of the virus genome that code for more than one protein. They are a ubiquitous component of virus genomes and have been documented in all Baltimore classifications of viruses [13]. In Prokaryotes and viruses, genes with OvRFs are defined as overlapping genes. In contrast, overlapping genes in Eukaryotes are defined on the basis of overlaps between the boundaries of the primary transcripts. OvRFs can overlap with three different shifts in reading frames, given the 5’ or 3’ directions of the respective sequences. Genes can partially overlap or can fully overlap in the case of nested overlapping genes (Figure 1.1), and more than two genes can overlap.

A recent study by Muñoz-Baena et al. [67] found that the majority of overlaps in viruses have a +2 frameshift and identified them predominantly in double-stranded DNA (dsDNA) viruses. The number of OvRFs increases with genome length; however, the overlaps also tend
to be shorter in longer genomes. One of many hypotheses for the existence and frequency of the OvRFs is that they provide a mechanism to overcome the constraints of small virus genome sizes. Overlapping genes makes it possible for the same length of sequence to code for multiple proteins [7]. Furthermore, the coding of multiple proteins from the same nucleotide sequence decreases amino acid redundancy and increases the sensitivity to mutations, potentially increasing the efficiency of purifying selection (see section 1.2) [17].

Figure 1.1: Different typologies of overlapping genes [108]. Each arrow represents an open reading frame (ORF). The arrow head indicates the direction of the reading frame. Overlap interactions can either be partial, with only a portion of each sequence overlapping at their ends, or nested, where one ORF is completely contained within another. Overlaps can also have different topologies: between sequences on the same strand (unidirectional), divergent, or convergent (between sequences on opposite strands).

1.1.3 Virus Classifications

The taxonomic classification of viruses — for instance, the recognition and naming of new virus species — is a subject of ongoing debate. Viruses do not neatly fit into established
biological classification systems, due in part to a lack of morphological characters and the pseudo-living nature of viruses. Recent phylogenomic analyses have helped clarify some of the evolutionary relationships among viruses [72]. For instance, there are five classes of RNA viruses and reverse-transcribing viruses that share a common origin, while both the single- and double-stranded DNA viruses appear to have evolved independently.

The Baltimore classification (BC) system differentiates viruses based on the different routes of information transfer as opposed to genome structure alone. It takes into account the type of nucleic acid (DNA or RNA) comprising the virus genome, its strandedness (single- or double-stranded; ss, ds), sense (positive or negative) and replication method to place the virus in one of seven groups [50]. However, it is not a fixed system and is continuously changing to adjust for new discoveries. The BC system started out with six groups, and was later expanded to seven groups (usually labelled with Roman numerals, I-VII; Table 1.1) to incorporate hepadnaviruses. The hepadnaviruses, including hepatitis B virus (HBV), have a dsDNA genome and use reverse transcriptase to replicate via an RNA intermediate. In addition, the BC system was further adjusted to accommodate the discovery of ambisense RNA viruses, such as members of the virus families Arenaviridae and Bunyaviridae. These viruses have genome segments that contain a mix of positive- and negative-sense open reading frames. Whether these viruses warrant the creation of a new group within the BC system, or should instead be assigned to a previously defined group, is an ongoing debate. Furthermore, it is possible for a virus to be assigned to more than one BC group. For example, viruses in the family Bacilladnaviridae have a ssDNA genome that also contains short dsDNA regions. As a result, these viruses can be classified into both groups I and II [50].

**DNA and RNA viruses**

The composition and structure of viral genomes are also very diverse. Both DNA and RNA viruses can be single- or double-stranded with linear, circular or even segmented configurations. DNA viruses with larger amounts of genetic material (> 10kb) tend to be double-stranded, while DNA viruses with less genetic material can have circular genomes, and be either single- or double-stranded. Some RNA viruses are classified as retroviruses because
1.1. Viruses

<table>
<thead>
<tr>
<th>Class</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>dsDNA</td>
<td>Adenoviruses, Herpesviruses, Poxviruses</td>
</tr>
<tr>
<td>II</td>
<td>ssDNA, (+)sense</td>
<td>Parvoviruses</td>
</tr>
<tr>
<td>III</td>
<td>dsRNA</td>
<td>Reoviruses</td>
</tr>
<tr>
<td>IV</td>
<td>(+)ssRNA</td>
<td>Picornaviruses, Togaviruses</td>
</tr>
<tr>
<td>V</td>
<td>(−)ssRNA</td>
<td>Orthomyxoviruses, Rhabdoviruses</td>
</tr>
<tr>
<td>VI</td>
<td>ssRNA-RT, (+)sense RNA with DNA intermediate</td>
<td>Retroviruses</td>
</tr>
<tr>
<td>VII</td>
<td>dsDNA-RT</td>
<td>Hepadnaviruses</td>
</tr>
</tbody>
</table>

Table 1.1: Current definition of Baltimore classifications of viruses, with examples of virus families.

They undergo reverse transcription to convert the RNA genetic material into complementary DNA. Retroviruses have a unique ability to use an integrase enzyme to insert the resulting double stranded (ds) DNA into the host cell genome. [50] The insertion of the viral genome into the genomic DNA of the host cell can cause numerous diseases, most notably cancers, as the integrated DNA can potentially disrupt functional components of the host genome. All dsDNA virus genomes consist of a single nucleic acid molecule. On the other hand, RNA viral genomes are frequently multi-partite (consisting of multiple genome segments); most are single-stranded. ssRNA genomes can be positive- or negative-stand and even ambisense (a mixture of +ve and −ve sense) [91, 97].

DNA and RNA viruses have substantially different rates of spontaneous mutation. Specifically, DNA virus genomes tend to be less mutation-prone than RNA virus genomes [75, 90]. The mutation rates for DNA viruses are on the order of $10^{-8}$ to $10^{-6}$ substitutions per nucleotide site per cell infection (s/n/c), while RNA viruses have a higher mutation rates that range between $10^{-6}$ and $10^{-4}$ s/n/c [89]. (A substitution occurs when a new mutation appears, is passed on to progeny, and increases in frequency until it is found in nearly all members of the population.) Virus species with smaller genomes tend to have a negative correlation between the genomic mutation rate and genome size, even after normalizing to rates per-site. This tendency is hypothesized to keep the per-genome mutation rate at relatively consistent levels [89, 62]. DNA viruses use a DNA polymerase to replicate the viral genome. This DNA polymerase has proofreading mechanisms. In contrast, the RNA polymerases used by RNA
viruses for replication, as well as the reverse transcriptase unique to retroviruses, lack such proofreading capabilities. This functional difference between polymerases contributes to the higher per-site mutation rate of RNA viruses. There are also reported differences in mutation rates between single- and double-stranded viruses, with single-stranded viruses tending to have higher mutation rates [62].

The genomic mutation rate of a virus greatly affects its ability to quickly adapt to changing environments. This is because the mutation rate determines the amount of genetic variation in the population on which natural selection can act. The correlation of higher mutation rates with higher rates of evolution makes public health efforts to combat RNA viruses more challenging. For example, influenza A virus circulates as a genetically diverse population, due to the high mutation rate, reassortment of genome segments, and persistent non-human host reservoirs. This genetic variation has made it difficult to respond to virus evolution with new vaccines that are rendered less effective by the next seasonal variant.

**Enveloped and non-enveloped viruses**

Enveloped and non-enveloped viruses both have a protective layer, the viral capsid, that surrounds the genetic material of the virus. The viral capsid is composed of structural proteins that self-assemble with weak non-covalent bonds. At the optimal pH range, the decrease in the free energy of the capsid proteins drives self-assembly. The surface charge and polarity of each structural protein is minimized, while maximizing their contact region with other capsid proteins. Moreover, mechanical traction or changes in pH or temperature can initiate the dissociation of capsid proteins, which results in the viral uncoating and release of genetic material into a host cell [45]. This was famously observed with tobacco mosaic virus capsid proteins *in vitro*, which self-assembled into capsid structures at the optimal pH and dissociated when pH was increased [31].

In non-enveloped viruses, the capsid is the outermost layer. However, in addition to the capsid, enveloped viruses also have a biological membrane that surrounds the viral capsid. As it is composed mainly of lipid bilayers, this membrane is fragile and susceptible to physical and chemical factors in the environment. Maintaining the integrity of the virus envelope is
necessary for host cell infection by enveloped viruses. As a result, enveloped viruses are more sensitive to heat, changes in pH, humidity and other factors [93]. However, the envelope also confers an evolutionary advantage, as it is derived mainly from the host cell’s own endoplasmic reticulum, Golgi apparatus or plasma membrane, which become embedded with viral proteins. Since much of the membrane surface is recognized as self by the host immune system, the envelope makes the virus less susceptible to immune responses. In addition, non-enveloped viruses require a highly immunogenic process of cell lysis to exit the host cell, which induces a response from the host immune system. In contrast, enveloped viruses can exit the host cell while leaving the cell membrane mostly intact [85].

### 1.1.4 Surface exposed virus proteins

Surface exposed proteins in viruses are proteins that are fully or partially exposed to the outside of the virion. Enveloped viruses have a plasma membrane made up of lipid bilayers (with hydrophobic and hydrophilic regions) as their outermost layer. Like cell surface-exposed host proteins, surface exposed proteins of enveloped viruses are either anchored directly into the membrane, or indirectly by being attached to lipids or proteins that are anchored in to the membrane. Integral or transmembrane proteins are a subset of surface exposed proteins that contain amphipathic (having both hydrophilic and hydrophobic regions) helices or beta strands that cross the membrane.

In non-enveloped viruses, the capsid is the outermost layer of the virion. However, not all capsid subunits are exposed to the outside of the virion. For example, Rotavirus A from the family of Reoviridae has outer, intermediate and inner capsid proteins (Figure 1.2). The outer capsid proteins (VP8, VP5 and VP7) are surface exposed and are attached to the intermediate capsid proteins, which are partially exposed. The inner capsid proteins beneath a layer of intermediate capsid proteins are not surface exposed.
Figure 1.2: A Rotavirus A virion with capsid proteins labelled on the left, and with the different layers of the capsid shown in the diagram on the right. This image was generated by ViralZone (Swiss Institute of Bioinformatics, SIB) [44] for its web resource (https://viralzone.expasy.org), and was released into the public domain under a Creative Commons Attribution Share-Alike 4.0 license (https://commons.wikimedia.org/wiki/File:Rotavirus_virion.jpg).
1.2 Natural Selection

One of the central mechanisms of evolution is natural selection. Selection drives the deterministic evolution of adaptive features that influence the survival and reproduction/replication of the organism/entity (fitness) [37]. Mutation is the ultimate source of all genetic variation on which natural selection can act. The majority of random mutations have a deleterious effect on the survival of the organism/entity. Consequently, the frequency of these mutations would decrease on average from one generation to the next, as it would be under ‘negative’ or ‘purifying’ selection. A mutation can occasionally confer a selective advantage that increases the fitness of the individual, resulting in ‘positive’ selection where the mutation frequency increases on average. In some cases, mutations under positive selection can reach fixation, where the mutation has been transmitted to all surviving individuals in the population. It is also possible for mutation to have no measurable effect on fitness but become fixed in the population due to genetic drift, a neutral stochastic process [1]. In either case, the fixation of a new mutation results in a substitution event.

1.2.1 Host-Virus Arms Race

The estimated number of viruses on Earth (about $10^{31} - 10^{32}$ particles) exceeds the number of host cells by orders of magnitude [71]. With every living cell assumed to be under constant attack from viruses, these parasites potentially exert the greatest selective pressure on living organisms [114]. Organisms do not evolve in isolation — we are constantly co-evolving with other organisms in our shared environment. In addition to the enormous abundance of viruses, many pairs of hosts and viruses have been co-evolving for a tremendously long amount of time. Over time, the evolution of one species influences the evolution of the other. This co-evolution is the concept underlying the Red-Queen hypothesis proposed by Van Valen [88], which states that continuous adaptation by one species is required just to maintain a consistent level of fitness in response to adaptation by the other species [15, 21]. In a prey/predator co-evolutionary relationship, the beneficial evolutionary changes in one species will tend to have negative consequences for other species. The negatively-impacted species would then become
under stronger selection to respond with their own evolutionary changes. For example, a virus may evolve to avoid detection by the host’s immune system, placing the host species under stronger selection pressure to improve its ability to detect the adapted virus (Figure 1.3). We assume that constantly staving off the advantage of their counterparts drives an elevated and perpetual rate of evolution in both species.

Figure 1.3: Schematic illustration of an evolutionary arms race between host and virus species.

Primates have been exposed to pathogenic viruses throughout their evolutionary history, and have consequently evolved mechanisms for preventing and restricting virus infection. Simultaneously, viruses have also been evolving to replicate or transmit more efficiently within hosts, for instance by gaining new accessory proteins that counteract the host immune and restriction factors. These processes have culminated in an ongoing evolutionary arms race. For example, the replication of human immunodeficiency virus (HIV) in the human host cells is restricted by the host factor apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3G (APOBEC3G), which increases the viral mutation rate during the reverse transcription of viral RNA to complementary DNA. In response, HIV has evolved the vif gene encoding an accessory protein that neutralises the activity of APOBEC by targeting it for degradation.
1.2. Natural Selection

by the host cell [63]. In addition, the human restriction factor bone marrow stromal antigen (BST-2), also known as tetherin, prevents the virus particle from budding off from the infected cell. Tetherin is, in turn, restricted by the HIV protein Vpu to promote virus budding [73].

The Host in the Evolutionary Arms Race

Viruses and host cells first come into contact through their surface-exposed components. Viruses initiate infection by binding to the host cell surface before they can enter the cell or deliver their genetic material. As a result, viruses have evolved to recognize and co-opt exposed proteins on the host cell surface. The proteins specifically recognized by the virus have been referred to as ‘viral receptors’ [103]. Viral receptors generally have some other function for the host cell. Some viral receptors are complex proteins, while others are molecules that are covalently attached to other host proteins, such as sialic acids that are bound by the influenza A virus. Due to the dichotomy in the functions served by viral receptors, they are assumed to be under two opposing directional evolutionary forces. Most host genes, viral receptors included, are under negative selection (removing mutations) to conserve the ‘normal’ functions that they provide for the host cell. However, the viral receptor is also under positive selection to escape recognition and binding by viruses to reduce the risk of infection [103].

A number of viral receptors have been studied individually in the context of co-evolution (see the following section). However, a recent study by Wang et al. [103] systematically analyzed 96 viral receptor genes in primates, determining that viral receptors exhibit significantly elevated rates of adaptation in comparison to other host genes. They were also able to show that the host-virus arms race resulted in elevated rates of adaptive evolution in viral receptors by identifying many sites undergoing significant positive selection on viral receptors, which were concentrated at the site of viral-receptor interfaces. Thus, Wang et al. [103] provide a more comprehensive picture of how viral receptors experience an elevated rate of evolution driven by the host-virus conflict. Finally, a study by Murrell et al. [69] explored the evolutionary history of genes encoding a number of host restriction factors (e.g., APOBEC3F, SERINC3; see next section) that are responsible for suppressing virus replication. They found a distinctive signature of selection (evolutionary fingerprint, see section 1.3.4) in ‘canonical’ arms-race
The Virus in the Evolutionary Arms Race

Similarly, there are several examples of accelerated evolution in virus genes due to an evolutionary arms race against its host. For example, the influenza A virus proteins hemagglutinin and neuraminidase are well-characterized targets of positive selection. This is evident by the elevated rates of non-synonymous substitution (see section 1.3.1) in the genes encoding these two surface proteins, compared to the rest of the genome [106, 79]. Elevated rates of evolution in surface-exposed proteins have also been observed, for example, in HIV-1 envelope glycoprotein [33]. Consequently, there has been a generally held view that positive selection is concentrated in regions of a virus genome encoding surface-exposed proteins [46, 40].

However, there is also abundant evidence of positive selection due to host-virus coevolution affecting genes that encode proteins that are not exposed on the surface of the virus. Several examples can be found in HIV-1, which is targeted by multiple host restriction factors as it enters and infects a host cell. For example, APOBEC 3F and 3G are mRNA-editing enzymes targeting the reverse transcription of the viral RNA genome into DNA. The resulting hypermutation effectively restricts virus infectivity. However, the HIV-1 vif gene has evolved to neutralize APOBEC activity [107, 65].

The host factor SAMHD1 also targets the reverse transcription of viral RNA genome. This is accomplished by reducing the concentration of cytoplasmic deoxynucleoside triphosphates (dNTPs) that would be used by the virus for reverse transcription. HIV-1 counteracts SAMHD1 by promoting its degradation with the accessory protein Vpx [56, 35, 43]. As mentioned above, the host factor tetherin (BST) prevents the virus particles from budding off from the host cell. Tetherin is restricted by both HIV-1 accessory proteins Vpu and Nef, although evidence for the latter is still being determined [26, 21].

Uncoating of viral capsids makes the viral RNA genomes accessible to host cell nucleases like TREX1, which actually reduces innate immune recognition by the host cell, benefiting HIV replication. HIV-1 exploits this mechanism by activating the SLX4 endonuclease complex.

In addition, SERINC3 and SERINC5 are transmembrane proteins that are highly conserved in eukaryotes. Multiple studies have found that these transmembrane proteins can be included in the membrane of the virus as it buds off. Their presence on the virus membrane makes the virus particle less infective, because it is more difficult for the virus particle to bind and fuse to another cell. However, when HIV-1 has the accessory gene encoding the Nef protein (the current wildtype state of HIV-1) then these transmembrane proteins are incorporated into the virus envelope less often, increasing its infectivity [26, 99]. These interactions result in intense adaptive selection pressures that dominate the evolutionary histories of both the host and viral genes involved in the arms race.

1.3 Selection Analysis

Genetic variation is produced by different evolutionary mechanisms, including mutation and recombination, upon which selection operates. To examine the amount of evolutionary forces acting on viruses we need to quantify selection pressure. This is often performed at the level of protein-coding sequences, where mutations that do not alter the protein provide a ‘neutral’ baseline.

1.3.1 Measuring selection from genetic sequences

There are several comparative methods for measuring selection from genetic sequences, including Tajima’s $D$, the McDonald-Kreitman test, and dN/dS methods.

Tajima’s $D$ [95] is a test statistic for the presence of selection based on comparing two estimators of $\theta$, which is a measure of genetic variation as a product of the mutation rate and the effective population size. The first estimator of $\theta$ is based on the number ($S$) of segregating (polymorphic) sites in a multiple sequence alignment. A second estimator is based on the number ($\pi$) of nucleotide differences between all pairs of sequences in the alignment. These two estimators are expected to be concordant under neutral evolution. Hence, the deviation of
one from another is used as evidence of selection.

The McDonald-Kreitman (MK) test [66] detects selection by comparing nucleotide diversity (polymorphisms) within species to the amount of nucleotide divergence (fixed differences) among species [12]. Most studies employing the MK test have focused on nonsynonymous mutations, using synonymous mutations as a neutral reference. If nonsynonymous mutations tend to be neutral, then the ratio of nonsynonymous to synonymous polymorphisms within species ($P_N/P_S$) is expected to be equal to the ratio of nonsynonymous to synonymous substitutions among species ($D_N/D_S$). Otherwise, $D_N/D_S$ would be greater than $P_N/P_S$ for substitutions conferring a selective advantage to the organism (more fixed differences than expected), and $D_N/D_S$ would be less than $P_N/P_S$ for deleterious substitutions (fewer fixed differences than expected).

Currently, the most widely used method for detecting the pattern of natural selection from nucleotide sequences is the $dN/dS$ ratio test (also known as $K_a/K_s$ or $\omega$). In a protein coding region of the genome, every three nucleotides (i.e., a codon) is translated into one of the 20 amino acids, which are the building blocks of proteins. As such, a nucleotide mutation in the codon can change the resulting amino acid — a non-synonymous mutation — which in turn may modify the structure or function of the protein. On the other hand, a nucleotide mutation can be synonymous due to redundancy in the genetic code, where there can be multiple codons that code for the same amino acid. The rate of adaptation by a population at the codon level is quantified by comparing the synonymous (dS) and non-synonymous (dN) substitution rates, which are each adjusted for the expected numbers of synonymous or non-synonymous substitutions in the codon (based on the genetic code). We assume that dS represents neutral evolution, such that it can be used as a baseline to compare dN against. If we are observing a higher dN substitution rate than dS, we assume it is because non-synonymous substitutions are under positive selection. However, if dN is lower than dS, then we assume that non-synonymous substitutions are deleterious to the virus and so are being removed from the population by negative selection. [34, 70]
1.3.2 Modeling variation in dN/dS

There are number of computational methods available for measuring the strength and direction of selection. Codon-based models of molecular evolution are useful for identifying sites evolving under selection in protein coding genes [2, 23]. Generally, models of codon substitution are derived from models of nucleotide substitution. For instance, the general time-reversible (GTR) model uses six rate parameters to model symmetric rates between all pairs of nucleotides (such that A to G has the same rate as G to A). The Muse-Gaut model of codon substitutions adds a single parameter that modifies the GTR rates if the codon substitution that results from the nucleotide substitution is non-synonymous [51].

These models can also be expanded to obtain more detailed estimates of the pattern of selection. In other words, the dN and dS rate parameters can be estimated independently for specific codon sites (variation across the genome) or specific branches in a phylogenetic tree (rate variation over time) [109]. Selection that affects only some branches in the tree is called ‘episodic’ selection — otherwise we assume that selection acts the same throughout the tree, i.e., ‘pervasive’ selection. There are substantial limitations to our power to detect variation in dN and dS across sites and branches. Introducing too many rate parameters to the model will tend to increase the number of false positive results, since the model is being applied to a finite amount of data. Therefore, a model selection test is needed to confirm that this variation in dN and dS is statistically significant and supported by the data. Also, current comparative methods for analyzing sequences cannot answer specific questions on the biological significance of changing an amino acid at a specific codon site along a branch in the phylogenetic tree. In the extreme case (specific rate parameters for every combination of a site and a branch), it is a statistically meaningless analysis, as there is a sample size of one. As with every statistical analysis, there exists a balance between keeping the model simple enough to have statistical power, and making the model complex enough to be biologically informative.
1.3.3 FUBAR

The Fast Unconstained Bayesian Approximation (FUBAR) method for inferring selection [68] was developed to address the problem of adding too many model parameters to model variation in rates across sites. This method is best suited to detect sites in a sequence alignment undergoing pervasive diversifying selection, \textit{i.e.}, subject to sustained positive selection across the entire phylogeny. It enables the analysis of very large data sets (>500 sequences) that other methods would struggle to run through in a reasonable amount of time. FUBAR achieves this by using an \textit{a priori} specified set of dN and dS values in a $20 \times 20$ grid that covers a wide range of selection pressures. Fixing the possible dN and dS rates speeds up the analysis by using precomputed probabilities that would otherwise have to be recalculated throughout the analysis [51, 74].

FUBAR uses a random walk algorithm (Markov chain Monte Carlo [111]) to sample different assignments of sites to different cells in the dN/dS grid. This is a Bayesian method that assumes a uniform prior probability distribution (our state of knowledge before seeing the data) over this grid of dN and dS values. When the random walk is run for a sufficiently large number of steps, then the history of this walk should converge to a random sample of the posterior probability distribution over the grid (our updated state of knowledge after seeing the data). The probability that a site is under positive selection can be estimated by counting the number of times it was assigned to a part of the dN/dS grid where dN exceeds dS. Dividing this count by the size of the sample gives a posterior probability that dN > dS. A posterior probability of 0.9 is recommended threshold for identifying sites under significant positive selection.

FUBAR is available as one of several methods on the Datamonkey web application [104]. This web interface only reports sites under positive selection, although a full set of reports can be downloaded as a JSON (JavaScript Object Notation)-formatted file. It can also be run locally using the command-line program HyPhy [52].
1.3.4 Gene selection analysis

Codon site selection analysis

dN and dS can be calculated per codon site by fitting a model of evolution to the observed distribution of substitutions in the phylogeny. The ratio of dN/dS can be used as a test statistic, where if there is no difference between synonymous and non-synonymous substitution then the ratio would be 1 and positive selection and negative selection is detected when dN/dS > 1 and dN/dS < 1, respectively. Another test statistic would be the difference in dN and dS (dN − dS) where neutral selection implies that dN − dS = 0 and positive or negative selection can be detected if dN − dS > 0 and dN − dS < 0, respectively.

To examine the selection pressure acting on an entire gene, the codon site-specific estimates of dN and dS can be averaged out over the length of the gene. A drawback to this method is the loss of information that occurs by averaging out the variation in rates of evolution among codon sites in the gene. For example, positive selection may target an active site of a protein undergoing immune escape. However, the rest of the protein can be under strong negative (purifying) selection in order to preserve its function. Hence, when averaging selection pressures over the entire sequence, the strong positive selective pressure at a small number of sites can be masked [110].

We can also measure the number of codon sites in the gene undergoing positive or negative selection by applying a threshold to the dN and dS rate estimates for every codon site, or to the significance level of a test statistic applied to these numbers. Furthermore, to quantify selection pressure acting on a gene we can calculate the percentage of codon sites in a gene sequence under positive and/or negative selection. These approaches recover some of the information that would otherwise be lost by averaging dN/dS across the entire gene.

Evolutionary Fingerprinting of Genes

FUBAR enables a new computational framework for comparing the patterns of negative and positive selection experienced by different, unrelated genes. An evolutionary fingerprint is the unique ‘mosaic’ of rates across codon sites of a gene, which may be evolving rapidly
or resisting change in response to selection [53]. dN and dS rates can be measured across the protein coding regions of the entire genome, either at the level of individual codon sites or averaged across the gene. However, there is usually a loss of information associated with these dN/dS rate estimates, either because there is not enough information to accurately estimate site-specific dN and dS rates, or from averaging rates across the gene. Evolutionary fingerprinting minimizes this information loss across a gene as it encompasses the joint distribution of dN and dS rates across the entire gene.

Mapping different gene alignments to the same grid of dN and dS rates means that it is possible to compare patterns of selection between completely unrelated genes and organisms. This comparison used to be done with the earth mover’s distance, also known as the Wasserstein metric [86]. To understand this distance, supposed that the distribution of posterior probability over the $20 \times 20$ grid, representing different combinations of dN and dS values, can be moved around like a pile of dirt. The earth mover’s distance calculates the smallest amount of dirt that needs to be moved in order to make one pile (distribution) into the same shape as another [68]. Subsequently, the developers of FUBAR switched to using the cosine similarity (equivalent to Pearson’s correlation) because of its better statistical characteristics [69].

A study by Murrell et al. [69] adapted FUBAR to characterize the evolutionary fingerprints of genes. Restriction factor genes code for proteins that block viral infection and replication, and they have some of the highest documented rates of adaptive evolution in primates that can be attributed to their involvement in the host-virus arms race. Murrell et al. calculated the evolutionary fingerprints of a number of primate genes encoding restriction factors, and compared it to diverse sample of host genes with other functions. They were able to identify a characteristic evolutionary fingerprint common to restriction factor genes involved in evolutionary arms races (canonical arms race genes). The study is also a proof of concept for how evolutionary fingerprinting of genes can be used as a hypothesis generating tool. Evolutionary fingerprinting is a powerful tool that allows the retrieval of an information layer that is left untapped by straightforward test statistics of dN and dS [69].
1.3.5 Assumptions in selection analysis

There are a number of factors that can bias the selection inference derived from homologous gene sequences.

1. Population evolution

Selection methods were developed primarily in a phylogenetic context with the assumption that the observed nucleotides in a gene are fixed in the population [25]. However, when applied to gene sequences from the same organism at population level, intra-species comparison, this assumption is no longer valid. Unfixed (segregating) mutations that are misidentified as fixed substitutions can lead to overestimation of site-specific dN/dS ratios [54].

2. Recombination

Another complicating process is intragenic recombination, which occurs when multiple viral genomes co-infect the same host cell, such that their genetic material can recombine to create new ‘chimeric’ genome sequences [5]. As such, recombination can result in the merger of genetic fragments with different evolutionary histories. Most comparative methods for measuring selection do not account for recombination, as they are designed to compare gene sequences from different species, where recombination is assumed to occur less frequently between species [1]. As such, when applied to sequence comparison between virus strains (divergent populations) with frequent intragenic recombination, it can also lead to overestimation of site-specific dN/dS [3, 84, 4].

3. Codon frequencies

Some amino acids are located and are involved directly at the active sites of a protein, while others could be adding to the solubility and stability of the protein structure. Consequently, the frequencies of the 20 amino acids can vary among sites, from which we can extrapolate that the codon frequencies could also vary in frequency among sites. There also can be variation in codon frequencies that is independent of variation in amino acids, as codon variation can be induced by the differences in the frequencies of synonymous codons. Heterogeneous codon
frequencies among sites can lead to underestimation of dN/dS [6]. Despite this known effect of variation in codon frequencies, most dN/dS selection analysis methods assume for simplicity that codon frequencies are the same for all sites in the gene [22].

4. Base composition

Non-stationary base composition occurs when the proportions of different nucleotides in a sequence change over time. Heterogeneous base composition over time can lead to over or under estimation of dN/dS [4]. A study in 2018 [38] found that decreasing GC-content over time tended to result in overestimation of dN/dS, and conversely, increasing GC-content caused underestimation of dN/dS.

5. Alignment quality

A multiple sequence alignment (MSA) of protein coding sequences is codon-aware if all sequences share a common reading frame, and gaps are preferentially placed at codon boundaries, i.e., so that a run of three gaps does not span two codons (AA---A). This type of alignment is needed not only to identify codon substitutions among sequences, but also to identify whether it is a synonymous (dS) or a non-synonymous (dN) substitution. As such, the quality of the MSA influences the accuracy of selection analysis. The incorrect alignment of non-homologous positions leads to overestimation of dN/dS on average [64, 92].

6. Substitution models of codon evolution

The rate of synonymous substitutions (dS) is used as a baseline measure of the evolutionary rate of a sequence. Under some of the original models used for measuring selection by estimating dN and dS, the analysis is carried out under the assumption that dS is constant across sites [74]. The existence of heterogeneous dS rates among sites in a sequence violates this assumption, which leads to overestimation of dN/dS [78].

It is also important to note that there are a variety of well-documented biological processes where synonymous mutations can have strong effects on fitness. For instance, synonymous mutations can affect splicing events, RNA stability, protein elongation rates, and protein fold-
ing [24, 77, 41]. Nonetheless, the majority of synonymous substitutions are considered to be neutral, and on average reflect the underlying mutation rate of a protein-coding gene; hence, most studies use dN/dS ratio to infer the direction of selection conferred by non-synonymous substitutions.

7. Overlapping reading frames

Overlapping genes can also cause problems for inferring selection. If a nucleotide occurs in two or more different open reading frames, then it is difficult to categorize a mutation of that nucleotide into nonsynonymous and synonymous categories. Consequently, two different forces can be acting on a nucleotide position and it would not be possible to tease out the effect of each force independently [87].

1.4 Rationale and Hypothesis

Viral infection requires the interaction between virus proteins and host cell receptors. The interaction between a virus and its host drives an evolutionary ‘arms race’ that is assumed to drive accelerated rates of evolution [102]. As the virus gains an evolutionary advantage over its host, the host is under increased selection to accumulate mutations in response (positive selection). Some well-known examples of this positive selection involve proteins that are exposed on the surface of the virus protein (surface-exposed, SE). For example, HIV-1 env encodes envelope glycoproteins that are responsible for binding the host cell receptor, and are a primary target of neutralizing antibodies. HIV-1 env has one of the fastest rates of adaptation ever recorded in any protein [105]. Similarly, influenza hemagglutinin is a surface glycoprotein responsible for binding the host cell receptor. Measures of positive selection are stronger in hemagglutinin than other proteins encoded by this virus [61]. Thus, there is a general assumption that surface-exposed virus proteins are generally under stronger positive selection. This assumption has not yet been tested comprehensively across multiple genes and viruses, however. We hypothesize that human virus genes encoding surface-exposed virus proteins are under higher rates of adaptation than other coding regions in the virus genome. To test this hypothesis, we conducted a systematic analysis of selection across a large number of virus
genomes to determine whether being SE is a feature of proteins that distinctively influences the evolutionary rate of virus proteins.
Chapter 2

Methods

2.1 Method Overview

To evaluate our hypothesis that genes encoding surface-exposed proteins in human viruses are under a higher adaptation than other coding regions in the genome, we needed a large collection of gene sequences representing both surface-exposed and non-surface exposed proteins. We wanted to test this hypothesis for a broad range of viruses because the effect of surface exposure should apply to any host-virus interaction. We chose to focus on human viruses, not only because of their impact on human morbidity and mortality, but also because they tend to have the most sequences available in public databases.

In order to apply the FUBAR method to measure non-synonymous (dN) and synonymous (dS) substitution rates, and to generate evolutionary fingerprints, we needed codon-aware alignments for each set of homologous gene sequences. We also needed an automated method for annotating each gene for surface exposure.
2.2 Data processing

2.2.1 Data Collection

We queried the National Center for Biotechnology Information (NCBI) Viral Genome database [14] for all complete reference genomes of human viruses (https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=10239&host=human). In NCBI terminology, a reference sequence (RefSeq) is a curated non-redundant and completely annotated sequence that serves as a standard reference for the group it represents [82]. We downloaded the resulting table as a tab-separated values (TSV) file. This table comprised the following fields for every reference genome:

- Genbank accession for RefSeq record;
- host (e.g., ‘human, vertebrates’);
- RefSeq type (complete or incomplete);
- number of proteins;
- number of neighbor genomes.

A neighbour genome is a validated sequence record that is related to a given RefSeq, but is not necessarily fully annotated [14]. Sets of neighbour genomes are curated by NCBI to capture the extent of genetic variation among related genomes.

We used bash scripts to generate additional versions of this file containing only segmented and non-segmented viruses based on the file structure. Next, we imported the original table and the two derived tables into R for processing. Based on the information in these tables, we excluded viruses with incomplete genomes, as indicated by the NCBI annotation of the genome record, and with fewer than 100 neighbour genomes. After this filtering step, we generated a file listing of accession numbers corresponding to the virus RefSeq records.

We also downloaded the viral genome neighbours table from the NCBI viral genome database, which contains the accession numbers for all neighbour genome associated with the RefSeq records. The neighbour table was imported into R and used to obtain extract all neighbour genome accessions for the subset of RefSeq entries by matching the reference /
representative accession numbers. The neighbours table also included information on the taxonomic assignment of each virus (e.g., “Poxviridae, Orthopoxvirus, Cowpox virus”), which we used to identify the family each virus belonged to. We used a Python script — including functions from the BioPython Entrez, SeqIO and SeqFeature submodules [19] — to retrieve the full genome sequences associated with the accession numbers for all reference and neighbour genomes, for every virus. This script communicates with the NCBI Entrez’s application program interface (API) that enables users to automate database transactions. The genome sequences were written to files in the FASTA format.

We also used a similar Python script to retrieve all Genbank annotations of coding sequences (CDS) associated with every reference genome. If the Genbank record had a feature type of ‘mat_peptide’, it was a polyprotein and as such we retrieved the coordinates for every mature peptide coded by the sequence, and cut out the corresponding coordinates from the genome sequence to retrieve the reference genes for the mature peptide product of the polyprotein. The sequence names in the resulting FASTA file were composed of the reference genome accession number, the protein accession number, gene product name, strand, and nucleotide coordinates in the reference genome.

We note that a RefSeq entry in the NCBI Virus database does not necessarily represent a virus species. Although many RefSeq entries serve as the reference sequence for a virus at the level of species, other entries would be considered a lineage or subtype of a virus species, e.g., ‘Hepatitis C virus genotype 1’. Therefore we will refrain from using the term ‘species’ when referring to a given virus.

### 2.2.2 Downsampling neighbour genomes

From each virus, we downsampled the reference and neighbour sequences to a target number of 100. This number was small enough for the selection analyses for hundreds of alignments to be computationally feasible, while being large enough for sufficient signal to estimate patterns of selection.

To maximize the amount of evolutionary signal in the downsampled data, we reconstructed
a phylogenetic tree relating the genome sequences for each virus. Reconstructing a phylogeny requires a multiple sequence alignment, which is a complex problem with a computing time on the order of $O(N^2L + NL^2)$ [49], where $N$ is the number of sequences and $L$ is the sequence length. Instead of carrying out a full alignment, we aligned each complete neighbour genome sequence to the reference sequence using minimap2 [59], which is a reference-based program for rapid pairwise alignment. To automate this process, we adapted a Python script that was previously developed for the CoVizu (coronavirus visualization) system (https://github.com/PoonLab/covizu/blob/master/covizu/minimap2.py) [28]. By discarding any insertions relative to the reference genome, this method can rapidly generate an approximate multiple sequence alignment.

We used the resulting alignment to reconstruct a phylogenetic tree by approximate maximum likelihood using the FastTree2 (version 2.1.11) [81]. We used a version of FastTree2 compiled for double precision to allow branch lengths shorter than $10^{-4}$ expected substitutions per site (i.e., one substitution in a genome of more than 10,000 nucleotides). Finally, we used another Python script (https://github.com/PoonLab/PoonLab/blob/master/prunetree.py) — using the BioPython Phylo submodule [96] to parse the Newick tree string generated by FastTree — to progressively cut the shortest tips of the phylogenetic tree, until we were left with 100 terminal nodes each representing a sampled genome. This procedure maximizes the genetic variation represented by a given number of genomes in the sample. To accelerate this step, we used the open source implementation of the message passing interface (Open MPI version 2.1.1) to process batches of input files in parallel on multiple cores of the processor.

### 2.2.3 Obtaining Coding Sequences

To work with gene-level alignments, we needed to extract homologous gene sequences from all 100 genome sequences for a given virus. Initially, we tried using the NCBI annotation for each genome to obtain the CDSs. However, we faced many challenges due to inconsistencies in both the extent of sequence annotation, as well the use of multiple names to refer to homologous CDSs. For example, homologous gene sequences in adenovirus genomes are labelled with ‘hexon’, ‘capsid protein II’ and ‘hypothetical protein’ [67]. In fact the prevalence
of miss annotations have driven multiple collaborations to create databases of genomes with manually validated gene annotations [36, 10]

Consequently, we started with the CDS annotations for the NCBI curated reference sequence (RefSeq) of each virus. We used minimap2 to align a given RefSeq CDS to every neighbour genome in the downsampled data set. Using the resulting pairwise alignment, we ‘cut out’ the region of the neighbour genome that aligned to the RefSeq CDS (Figure 2.1), using the position that minimap2 placed the RefSeq CDS in the neighbour genome and the length of the aligned CDS [59, 28]. Then we used the alignment program MAFFT (version 7.310) with default settings to carry out pairwise alignment of the cut neighbour CDS to the RefSeq CDS. [48]. Any positions in the resulting pairwise alignment containing a gap character in the RefSeq CDS was removed. This step guaranteed that the neighbour CDS would have the same reading frame and coordinate system as the RefSeq CDS. If minimap2 failed to align a neighbour genome to a RefSeq CDS, we ran a salvage alignment using the more accurate but time-consuming MAFFT program. Since MAFFT does not automatically attempt to align one sequence to the reverse-complement of the other, we tried both versions of the neighbour genome and then chose the version that minimized the proportion of nucleotide differences in the aligned region (i.e., the p-distance). To control for the quality of CDS alignments, we discarded neighbour CDSs if more than 50% of the aligned sequence consisted of gaps.

Figure 2.1: Cutting out a coding sequence (CDS) from a neighbour (query) genome by pairwise alignment to a homologous CDS in the reference genome. An annotated CDS from the reference genome (‘Ref’, red) is aligned pairwise against a query genome (dark blue) and the aligned region is cut out of the query genome and identified as a query CDS homologous to the ref CDS used in the alignment
2.2.4 Remove Overlapping Regions

Overlapping coding sequences are problematic for calculating dN and dS values. When overlapping coding sequences have different reading frames it can be impossible to identify whether a mutation is synonymous or non-synonymous. Nucleotides in the overlapping regions code for multiple proteins, as such they are under selection pressure experienced by multiple proteins making it impossible to tease out the selection pressure of one gene from another. Therefore we removed overlapping regions from CDS before carrying out the selection analysis.

The presence of overlapping genes for each virus genome was determined from the coordinates of the respective CDS features in the NCBI RefSeq record. Specifically, we expanded the coordinate interval(s) for each CDS into a list of integers, e.g., “5:9, 11:13” becomes “5, 6, 7, 8, 9, 11, 12, 13”. We concatenated these lists across CDSs and then checked for duplicate entries. The presence of a duplicate integer identified nucleotide coordinates for overlaps between genes in the genome. These positions in the reference genome were removed from every CDS extracted from the neighbour genome by the previous step. To preserve the reading frame of the CDS, we discarded the affected positions on codon boundaries (see Figure 2.2). After removing any overlapping regions, genes with less than 50 nucleotides remaining were discarded from further analysis. However, we ran our selection and fingerprinting analysis on both sets of genes with and without overlapping regions.

2.3 Identifying Surface Exposed (SE) Genes

2.3.1 Using Gene Ontology IDs

We labeled each of the genes (CDSs) being examined as surface exposed (SE) or non-surface exposed (non-SE) based on their Gene Ontology (GO) IDs from the Gene Ontology Resources [16] [29] (http://geneontology.org/, last accessed June 2022). Specifically, we queried the UniProt database with the Genbank protein accession numbers for all CDSs in our data set, and downloaded the result as a tabular file. We imported the contents of this file into R and then parsed the text to retrieve any GO terms associated with each record.
2.3. Identifying Surface Exposed (SE) Genes

Figure 2.2: Removing overlapping regions in the genome while respecting the codon boundaries of the target coding sequence. RefSeq genes A and B are an example of partially overlapping ORFs. In this case, the overlapping region from either the beginning or the end of each gene was removed at the respective codon boundaries. RefSeq genes C and D represent a nested overlap. Here, gene D is discarded as all of its codons are overlapping sites. The non-overlapping sections of Gene C are truncated at the codon boundaries and concatenated together into a new sequence.
In non-enveloped viruses, the capsid is the outermost layer, while in enveloped viruses the envelop is the outermost layer of the virus. As such, we first grouped genes based on whether they were from an enveloped or a non-enveloped virus. For enveloped viruses, we used the presence of the gene ontology ID for ‘Virion Membrane’ representing the lipid bilayer surrounding a virion (GO:0055036). For non-enveloped viruses, we used the gene ontology ID for ”Viral capsid” (GO:0019028) to label a gene as coding a SE protein. If the respective GOID was missing for a gene then it was labelled as a non-SE protein (see Figure 2.3).

Figure 2.3: Schematic depiction of the workflow for identifying surface-exposed genes. Genes were separated by their association with enveloped or non-enveloped viruses, based on Gene Ontology terms in the UNIPROT database (top layer). For enveloped viruses, genes lacking UNIPROT entries were identified using a transmembrane protein predictor (TMbed). Genes lacking both UNIPROT entries and TMbed predictions were manually annotated. For non-enveloped viruses, genes lacking UNIPROT entries were also manually annotated.

2.3.2 Using a Transmembrane Protein Predictor

These are a number of bioinformatics tools developed that can predict the presence and location of transmembrane regions of proteins. We used the newly developed transmembrane
protein predictor TMbed which uses language model embeddings [9]. A language model enables a computer to emulate a spoken (‘natural’) language by generating character or word sequences according to a set of rules. The particular model used in TMbed is based on an artificial neural network that ‘learns’ a language from a training set of example sequences.

We obtained the predicted transmembrane regions of all reference genes from enveloped viruses. The RefSeq genes used in the selection analysis were translated into amino acid sequences and input into TMbed, which was run locally using its standalone command line program. The default three line output was parsed in Python to count the number of codon sites identified as transmembrane beta strand ($B & b$) and alpha helix ($H & h$).

2.4 Selection Analysis

The selection analysis to detect site-specific pervasive diversifying and purifying selection was carried out using the Fast Unconstrained Bayesian Approximation (FUBAR) method implemented in HyPhy (http://www.hyphy.org) [68]. Both a multiple sequence alignment (MSA) and a phylogenetic tree relating all the coding sequences in the alignment are required to run a selection analysis with FUBAR.

For each alignment, we removed any duplicate sequences (i.e., sequences that are completely identical along their whole length), and replaced any stop codons with gaps using the HyPhy batch script ‘CleanStopCodon.bf’. Next, a phylogenetic tree of MSA of each gene was constructed using the approximate maximum likelihood method in FastTree2 (version 2.1.11) compiled with double precision [81]. We increased the sampling parameter to 10 chain samples (random walks) that were iterated for 10 million steps each. Next, we discarded the first million steps as ‘burn-in’, so that the location of a random walk was not influenced by its starting point. The output of a FUBAR analysis is written to a JSON file with numerous details about the selection analysis including the codon site-specific dN and dS estimates, the posterior probability of dN being greater or less than dS for each codon, and the joint posterior probability distribution over the fixed dN and dS values over the FUBAR grid (also known as the evolutionary fingerprint). Specifically, these values were retrieved from the MLE (maximum
Figure 2.4: Diagram summarizing the workflow for the selection analysis for each gene. The multiple sequence alignment (MSA) and phylogenetic tree of the sequences for each gene are fed into FUBAR, which calculates the dN and dS values for each codon site, as well as the evolutionary fingerprint.

2.4.1 Average dN/dS ratio

We parsed the JSON output file for each CDS in the R statistical computing environment to retrieve the codon site-specific dN and dS estimates. (We note here that HyPhy labels dN as $\beta$ and dS as $\alpha$, but we will retain the more conventional notation throughout this thesis.) The mean dN/dS ratio was calculated by averaging dN and dS values across all sites, respectively, and taking the ratio of the results. For comparison, we also calculated the mean dN/dS ratio by taking the ratio of individual dN and dS estimates at each site, and then averaging the results. The distributions of mean dN/dS ratios for all SE genes and all non-SE genes were compared using the non-parametric Wilcoxon rank sum test to determine whether there was a significant difference between these groups.

2.4.2 Percentage of Positively & Negatively selected sites

FUBAR estimates the posterior probability that dS is greater than or less than dN for a given site, by determining how often that site was assigned to the respective sector of the dN/dS grid in the random walks. For each codon site, if the probability that dN is greater than dS is $> 0.8$, likelihood estimate) field of the JSON file.
it was identified as a positively selected site; if probability that $dS$ is greater than $dN$ is $>0.8$, then it was a negatively selected site. The numbers of ‘significantly’ positively- or negatively-selected sites were each normalized by the total number of codons for each gene to obtain the percentages of positive and negative sites. Again, we used the Wilcoxon rank sum test to compare the distributions of these values between SE and non-SE groups of genes in R.

### 2.4.3 Evolutionary Fingerprint

FUBAR estimates the distribution of selection across all codon sites in the form of posterior probabilities over 400 combinations of fixed $dN$ and $dS$ values on a $20 \times 20$ grid, which is reported in the JSON file [68]. By importing these posterior probabilities into R, we could generate coloured heatmaps and perspective plots of the posterior distribution over this grid, which served as a visual representation of the evolutionary fingerprint [53].

We then needed to quantify the difference between evolutionary fingerprints, i.e., to determine numerically how similar or dissimilar one fingerprint is to another. If we assume that every posterior probability is an independent outcome, then the evolutionary fingerprint of a gene can be represented as a vector of 400 posterior probabilities. In other words, it is a vector with a length and direction in a high-dimensional space. Since these probabilities are being treated as independent measurements, we can calculate the Pearson correlation coefficient ($r$) between the 400 values of two fingerprints to measure how similar they are:

$$
 r_{x,y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}
$$

(2.1)

We used $1 - r$ as a measure of the distance between two fingerprints as described by [69] such that a difference of 0 means perfect correlation. This is mathematically equivalent to calculating the cosine similarity between two vectors:

$$
 \cos \theta = \frac{x \cdot y}{\|x\| \|y\|}
$$

recentered on $(\bar{x}, \bar{y})$, where $x \cdot y$ is the dot product $\sum_i x_i y_i$, and $\|x\|$ is the Euclidean norm.
\[ \sqrt{x_1^2 + \ldots + x_n^2}. \]

### 2.5 Data visualization

We calculated the pairwise matrix of distances between every pair of genes in the study, and then visualised this matrix using two dimensionality reduction methods: Principal Component Analysis (PCA) and \( t \)-distributed stochastic neighbour embedding (\( t \)-SNE [101]). PCA rotates the distribution of data points to maximize the variation along the smallest number of axes (principal components), whereas \( t \)-SNE maps the data points to two or three dimensions in such a way that preserves clusters of similar points. However, we will only display results with PCA plots here. To compare the evolutionary fingerprints of SE and non-SE genes, the points representing genes in a PCA plot was coloured based on whether they encode SE or non-SE proteins.

We compared the distribution of enveloped and non-enveloped viruses on the PCA. A manually curated database of enveloped and non-enveloped viruses was created and used to colour the PCA. Also, the distributions of different virus families were each examined by colouring the respective gene sets on a PCA by family, according to the NCBI taxonomy field for the RefSeq virus genome records.

For each gene, the potentially confounding effect of sequence diversity among sequences on natural selection was examined. Amino acid entropy and tree length (total branch length of the phylogenetic tree) were used as a measure of sequence diversity of each gene. We coloured the genes on PCA based on amino acid entropy as well as tree length to examine their distribution.
Chapter 3

Results

3.1 Virus genome data collection

For a systematic study, we examined all 508 human virus genomes available in the NCBI Viral Genome Resource database [14] (last accessed June 2022). We filtered these reference virus genomes so that there were more than 100 curated neighbour genomes per reference (see Methods). The resulting median number of neighbour genomes for each selected virus was 366 (interquartile range, IQR = 158 to 1170) genome sequences. After applying the rest of our filtering criteria to these data, the final data set comprised the genome sequences of 52 viruses belonging to 21 virus families and 15 orders (Figure 3.1; see also Appendix, Figure A.1). Rotavirus A has the most number of curated neighbour sequences available with 14,888 genomes. The severe fever with thrombocytopenia syndrome (SFTS) virus HB29, also called Dabie bandavirus, had the fewest neighbour genomes with only 100 sequences. The availability of neighbour genomes was consistent with the amount of time that the virus has been sampled from human subjects and studied. For example, SFTS virus was first reported in central China in 2009 [113], whereas rotaviruses have circulated in the human population for decades at least [98].

We generated multiple sequence alignments of the reference and neighbour genomes for
each virus, and then reconstructed a phylogeny for each set of genomes. Next, we ‘pruned’ the shortest branches of the tree to reduce the number of genomes down to 100 sequences (as described in Methods, 2.2.2, *Downsampling neighbour genomes*). This normalized the sample size across genes and viruses, while maximizing the amount of genetic variation available for detecting patterns of selection.

Figure 3.1: Pie charts depicting the distribution of all 21 virus families. The inner pie chart is based on the relative frequencies of 52 viruses distributed among the virus families, while the outer pie chart shows the relative frequencies of 670 genes among virus families for all data in this study.
3.1.1 Obtaining coding sequences from Genome

Using each protein coding sequence (CDS) in the reference genome nucleotide sequence of a given virus species, we used pairwise alignment to locate and cut out the homologous CDS sequences from neighbour genome. 112 out of 670 genes were polyproteins (encoding multiple products that become mature peptides with different functions). In these cases we used the ‘mat.peptide’ field to extract coding sequences. We also removed overlapping regions, in which nucleotides occurred in more than one open reading frame, from every CDS, resulting in 802 CDSs from the original total of 938 reference CDSs.

Four polyproteins lacked any annotation of mature peptides (the sub-sequence equivalent of CDSs) in the Genbank database and were excluded. In addition, 128 (15.9%) CDSs with fewer than 50 homologous nucleotide sequences out of the 100 genomes for a given virus species were removed, leaving a total of 670 genes for our analysis. The majority of CDSs discarded at this step was either due to poor alignment of the reference CDS to the neighbour genome, or due to a large fraction (or the entirety) of the CDS being affected by overlapping open reading frames. For example, our requirement for a minimal amount of overlap excluded the majority of genes encoded by the human immunodeficiency virus (HIV) type 1, i.e., leaving four out of 10 known genes. Despite the exclusion of overlapping regions from our analysis, the mean codon lengths for gene alignments with and without overlapping regions were not significantly different (Student’s t test, $P > 0.49$; Figure 3.2).

The median number of genes per virus was 5 (IQR = 3 to 9) genes. However, there was very long right tail in this distribution, with a maximum of 163 genes associated with the human betaherpesvirus 5 genome. In fact, the top three viruses with respect to the number of genes all belonged to the *Herpesviridae* family (including human herpesvirus 4 type 2 with 78 genes, and human alphaherpesvirus 3 with 72 genes). This over-representation of the *Herpesviridae* family (dark blue) at the gene level is visually conspicuous in the outermost pie of Figure 3.1.
Figure 3.2: Overlapping histograms representing the distributions of the log-transformed lengths (number of codons) of the 670 gene sequences in this study. We use colour to distinguish between the distributions for gene lengths with overlapping regions (blue), and the lengths after removing the overlapping regions (red), as described in Methods.
3.1.2 Distinguishing surface exposed genes

The UNIPROT database [20], which stores structural and functional (e.g., Gene Ontology or GO) annotations for protein sequences, had records available for 408 (60.9%) of the 670 genes in this study. For enveloped viruses, we used the presence of GO IDs for ‘virion membrane’ (GO:0055036) or ‘viral envelope’ (GO:0019031) as an identifier of SE genes. By this method, we categorized 59 of the 310 genes (18.9%) from enveloped viruses as being SE. In contrast, we used the GO ID ‘viral capsid’ (GO:0019028) to identify 23 (21.4%) SE genes from 98 genes representing non-enveloped viruses.

![Stacked barplots summarizing the distribution of surface-exposed (SE) and non-surface exposed (non-SE) genes with respect to enveloped and non-enveloped viruses (A) and nucleic acid (B).](image)

Figure 3.3: Stacked barplots summarizing the distribution of surface-exposed (SE) and non-surface exposed (non-SE) genes with respect to enveloped and non-enveloped viruses (A) and nucleic acid (B).

For the remaining 262 genes without a UNIPORT entry, 183 genes represented enveloped viruses and 79 genes represented non-enveloped viruses. We used the transmembrane protein predictor TMbed [9] to classify genes from enveloped viruses as SE or non-SE. The presence of any length of amino acid sequence predicted to be a transmembrane region in the protein was used to categorize 17 out of 183 genes as SE. We also categorized 126 of the genes as non-SE based on the absence of any predicted transmembrane regions, and were unable to recover any predictions from TMbed for the remaining 40 genes. We manually annotated these
40 uncategorized genes from enveloped viruses, and all genes from non-enveloped viruses without UNIPROT entries, using the records corresponding to these viruses in the ViralZone database curated by the Swiss Institute of Bioinformatics [44]. In total, we classified 125 genes as SE and 545 genes as non-SE (Figure 3.3A).

3.2 Selection analysis

We used the fast unconstrained Bayesian approximation for inferring selection (FUBAR, [52]) method to estimate site-specific non-synonymous (dN) and synonymous (dS) substitution rates in the codon-aligned sequences for each gene. The ratio of dN to dS provides a standard measure of positive selection on the proteins encoded by these genes. Thus, a ratio of one indicates that amino acid substitutions are accumulated at the same rate as nucleotide substitutions that do not change amino acids (neutral evolution). Averaging the dN/dS ratios across all codons in the gene alignment provides a crude measure of the overall amount of positive selection. However, this average can also obscure informative variation in substitution rates among sites. These average estimates are summarized in Figures 3.4 and 3.5 for non-enveloped and enveloped viruses, respectively. Overall, we found that the average dN/dS ratios rarely exceeded 1, which is consistent with most amino-acid altering mutations causing a deleterious change in the protein. In other words, we found pervasive negative (purifying) selection removing amino acid-altering mutations from the population. The three genes with a mean dN/dS exceeding one were: p12 nucleocapsid (NP_777382.1) from bovine leukemia virus, non-structural protein NS4 (NP_150079.1) from bovine coronavirus, and envelope glycoprotein (NP_040137.1) from human alphaherpesvirus 3.

To evaluate the hypothesis that SE genes are under stronger positive selection than non-SE genes in virus genomes, we compared the mean dN/dS ratios for these two groups and found no significant difference (Wilcoxon ranked sum test, \( p = 0.89 \); Figure 3.6A). Another test statistic for positive selection is to calculate the difference in dN and dS and evaluate whether \( dN - dS > 0 \). Again, we found no significant difference between non-SE and SE groups with respect to this test statistic (\( p = 0.44 \), Figure 3.6B). Finally, we examined the mean dN and dS
Figure 3.4: Stripchart summarizing the average dN/dS ratios for all genes from non-enveloped viruses, grouped by virus family. Every point represents a gene, coloured with respect to being surface-exposed (SE, red) and non-surface-exposed (non-SE, blue), respectively. Each row (strip) corresponds to a specific virus as labelled on the left, and the respective virus family labels are on the right.
Figure 3.5: Stripchart summarizing the average dN/dS ratios for all genes from en-veloped viruses, grouped by virus family. Every point represents a gene, coloured with respect to being surface-exposed (SE, red) and non-surface-exposed (non-SE, blue), respectively. Each row (strip) corresponds to a specific virus as labelled on the left, and the respective virus family labels are on the right. Vertical line segments are drawn at dN/dS=1 (neutral evolution) for reference.
rate estimates individually. The most amount of variation was seen between the mean dS of SE and non-SE genes. There was no significant difference in either dN or dS (Wilcoxon, p=0.3, p=0.3, respectively; Figure 3.6).

However, there was a significant difference between the dN/dS ratio of genes from enveloped versus non-enveloped viruses (Wilcoxon rank sum test, \( p = 1.4 \times 10^{-5} \)), as well as between genes with DNA versus RNA as their genetic material \( (p < 2.2 \times 10^{-16}, \text{Figure 3.7}) \). Since we observed significant difference in the mean dN/dS ratio for both categorizations of viruses, we examined the proportion of genes from enveloped and non-enveloped viruses that had DNA as the genetic material to see if these results were due to confounding. Overall, a relatively greater proportion of genes were associated with non-enveloped \( (n = 141) \) than enveloped \( (n = 36) \) DNA viruses in comparison to RNA viruses (Figure 3.7A). This association was statistically significant (Fisher’s exact test, odds ratio = 2.17, \( p = 1.5 \times 10^{-4} \)). In other words, a gene in a DNA virus genome was about two-fold more likely to be associated with an enveloped virus.

To further examine the statistical association of dN/dS ratios to virus characteristics, we stratified our results by enveloped and non-enveloped viruses and DNA/RNA genomes (Figure 3.8). We analyzed these data under a multi-variable regression model with an inverse link function (i.e., gamma regression), because the outcome variable (mean dN/dS ratio) is strictly positive (i.e., dN/dS > 0 for all dN \( \geq 0 \) and dS \( \geq 0 \)). The reference (intercept) levels for the regression model were RNA, non-enveloped, and non-surface exposed. We found the mean dN/dS ratio of a gene was significantly associated with both RNA/DNA (coefficient estimate \( \beta = -2.59, p < 2 \times 10^{-16} \); Table 3.1) and enveloped/non-enveloped categorical predictors \( (\beta = -0.57, p = 4.9 \times 10^{-6}) \), while being surface-exposed was not a significant predictor \( (\beta = -0.06, p = 0.61) \). The coefficient of determination \( (R^2) \) of this model was 37.2\%, which is the proportion of variation in outcomes explained by the predictors.

The above regression model ignores the inherent structure in the data, i.e., that some genes belong to the same virus genome. This is a potential issue because there is substantial variation in the number of genes among viruses. For example, human betaherpesvirus 5 (also known as
Figure 3.6: Box-and-whisker plots summarizing the distributions of (A) dN/dS ratio, (B) the difference in dN and dS, and individual (C) dN and (D) dS rate estimates across all viruses. Each point represents a gene, which are stratified into non-surface exposed (non-SE) and surface-exposed (SE) proteins. Thick line segments indicate the median for each distribution, and the upper and lower limits of the box represent the inter-quartile ranges. The $p$-values obtained from two-sample Wilcoxon rank sum tests of differences in group means are displayed for each plot.
Figure 3.7: Distributions of mean dN/dS estimates for genes stratified by virus types. (A) Stacked barplot displaying the numbers of genes with respect to nucleic acid (RNA and DNA) and non-enveloped or enveloped virus. (B) Box-and-whisker plot comparing mean dN/dS for genes in non-enveloped (indigo) and enveloped (cyan) viruses. (C) Box-and-whisker plot comparing mean dN/dS for genes in RNA (red) and DNA (blue) viruses.
Figure 3.8: Box-and-whisker plots comparing dN/dS ratios for SE and non-SE genes across categories. Genes are separated into four groups based on whether they belong to an enveloped or non-enveloped virus, and whether the genome is DNA or RNA. We report the $P$-value for Wilcoxon rank sum tests comparing SE and non-SE genes in each category above each set of plots. The $n$ value indicates total number of SE and non-SE genes in each category.
3.2. Selection analysis

<table>
<thead>
<tr>
<th>Model</th>
<th>Term</th>
<th>Estimate (± s.d.)</th>
<th>Effect</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed effects</td>
<td>Intercept</td>
<td>4.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DNA</td>
<td>-2.59</td>
<td>+0.28</td>
<td>&lt; 2 × 10^{-16}</td>
</tr>
<tr>
<td></td>
<td>enveloped</td>
<td>-0.57</td>
<td>+0.03</td>
<td>4.94 × 10^{-6}</td>
</tr>
<tr>
<td></td>
<td>surface-exposed</td>
<td>-0.06</td>
<td>+2.8 × 10^{-3}</td>
<td>0.6</td>
</tr>
<tr>
<td>Mixed effects</td>
<td>Intercept</td>
<td>9.00 ± 1.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DNA</td>
<td>-5.97</td>
<td>+0.22</td>
<td>1.06 × 10^{-9}</td>
</tr>
<tr>
<td></td>
<td>enveloped</td>
<td>-3.64</td>
<td>+0.08</td>
<td>9.01 × 10^{-5}</td>
</tr>
<tr>
<td></td>
<td>surface-exposed</td>
<td>-0.03</td>
<td>+3.7 × 10^{-4}</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 3.1: Summary of generalized linear model (GLM) analyses of the mean dN/dS across genes. The basic gamma GLM is reported as ‘fixed effects’, and a second model with a random intercept for genes grouped by virus is reported as ‘mixed effects’. To interpret the coefficient estimates (‘effect’), we note that the gamma regression model takes the form \( E(y) = (\alpha + \beta_1 x_1 + \ldots + \beta_n x_n)^{-1} \), where \( E(y) \) is the predicted dN/dS, \( \alpha \) is the intercept, \( x_i \) is the \( i \)-th predictor variable and \( \beta_i \) is the \( i \)-th coefficient.

human cytomegalovirus) is an enveloped DNA virus that is over-represented in the regression model because it has 175 genes, of which 163 are represented in this study. This corresponds to 24% of all genes in our data set. On the other hand, Lassa mammarenavirus is an enveloped RNA virus that only has 5 genes in total, with 3 included in this study. As such, we used a mixed-effects regression model to account for the inherent structure in our data. Specifically, we added virus-specific intercepts as a random effect term in the model (i.e., a random intercepts model), such that intercepts are drawn from a normal distribution with a mean equal to the global intercept \( \alpha \) and a standard deviation \( \sigma \) to be estimated, \( \alpha_i \sim N(\alpha, \sigma) \). The model identified substantial variation in virus-specific intercepts (\( \alpha_i \)) with an estimated variance of 1.837 (standard deviation \( \sigma = 1.36 \)) and residual variance of 0.26 (\( \sigma = 0.51 \)). The effect of the RNA/DNA categorical variable remained significant (\( \beta = -5.97, p < 1.06 \times 10^{-9} \), Table 3.1). However, the effect of the enveloped categorical variable was substantially diminished (\( \beta = -3.64, p = 9.01 \times 10^{-5} \)), and the surface variable remained non-significant (\( \beta = -0.03, p = 0.75 \)).

To account for some of the information lost by averaging the dN and dS values over the length of the gene, we also examined the percentage of codon sites under significant positive
Figure 3.9: Box-and-whisker plots summarizing the percentage of (A) negative sites, (B) positive sites and (C) sum of positive and negative sites across all genes. For each codon site, if the probability that $d_N$ is greater than $d_S$ is $> 0.8$, it was identified as a positively selected site and if probability that $d_S$ is greater than $d_N$ is $> 0.8$ as a negatively selected site. Percentage across genes were obtained by averaging out over the total number of codon sites. Each point represents a gene, which are stratified into non-surface exposed (non-SE) and surface exposed (SE) proteins.
and/or negative selection, as determined by FUBAR. We also measured the sum of these percentages (proportion of either positively or negatively selected sites) to quantify the overall extent of selection. These results are summarized in Figure 3.9. The overall mean percentage of negatively selected sites 39% was substantially greater than positively selected sites 1.8%. We found no significant difference between non-SE and SE groups with respect to any of these percentages (Wilcoxon rank sum test, $p > 0.05$).

In addition, we compared these percentages for groups of genes stratified by DNA and RNA viruses (Figure 3.10), and by enveloped and non-enveloped viruses (Figure 3.11). We found significant differences between DNA and RNA viruses with respect to the percentages of positively and negatively selected sites, and dN itself ($p < 2.2 \times 10^{-16}$, Figure 3.10). These quantities indicated that genes in RNA viruses were under significantly more negative (purifying selection). We also found evidence of significantly more negative selection in non-enveloped viruses, although this affected only the percentage of negatively selected sites and estimates of dN (Figure 3.11).

### 3.3 Evolutionary fingerprints

To get a more holistic view of the selection pressures experienced by the entire length of the gene, we examined the statistical distribution of codon-specific selection of each gene, which is known as the ‘evolutionary fingerprint’ [53]. We used FUBAR to fit a posterior probability distribution over a grid of 400 combinations of 20 dN and 20 dS values to create an evolutionary fingerprint.

To quantify the similarity between the evolutionary fingerprints of different genes, we calculated the cosine similarity for every pair of genes. This similarity measure is equivalent to calculating the Pearson correlation in posterior probabilities for every square in the grid. We visualized the resulting pairwise similarity matrix by using a principal components analysis to project this matrix into a reduced number of dimensions. To evaluate the hypothesis that SE genes are under stronger positive selection than other viral genes, we coloured the genes in the PCA plots as either SE (red) or non-SE (blue). If the hypothesis is supported by our data,
Figure 3.10: Box-and-whisker plots summarizing the distributions of (A) dN/dS ratio, (B) the difference in dN and dS, and individual (C) dN and (D) dS rate estimates across all viruses. Each point represents a gene, which are stratified into non-surface exposed (non-SE) and surface-exposed (SE) proteins. Thick line segments indicate the median for each distribution, and the upper and lower limits of the box represent the inter-quartile ranges. The $p$-values obtained from two-sample Wilcoxon rank sum tests of differences in group means are displayed for each plot.
Figure 3.11: Box-and-whisker plots summarizing the distributions of (A) dN/dS ratio, (B) the difference in dN and dS, and individual (C) dN and (D) dS rate estimates across all viruses. This supplements the plot of dN/dS distributions in Figure 3.7. Each point represents a gene, which are stratified into enveloped (cyan) and non-enveloped (indigo) viruses. Thick line segments indicate the median for each distribution, and the upper and lower limits of the box represent the inter-quartile ranges. The p-values obtained from two-sample Wilcoxon rank sum tests of differences in group means are displayed for each plot.
then we would expect SE genes to have evolutionary fingerprints distinct from non-SE genes, such that these points cluster together in the PCA. However, there was a lack of any visible clustering, indicating that the evolutionary fingerprints of SE genes were no more similar to one another than to non-SE genes (Figure 3.12).

To better visualise the difference in fingerprints we selected representative SE and non-SE genes from opposite ends of the PCA with respect to the first principal component (explaining 77.2\% of the variation, Figure 3.12D). These proteins were human alphaherpesvirus 3 envelope glycoprotein B (NP_040154), human metapneumovirus fusion protein (YP_009513268), human betaherpesvirus 5 protein UL30A (YP_007947989), and simian agent 10 large polymerase protein (YP_009505441). Figure 3.13 displays the evolutionary fingerprints for these proteins as heatmaps. We can see that the heatmaps corresponding to genes from the extreme left side of the PCA (Figure 3.13A, C) are visibly more similar to each other. Moreover, the heatmaps for genes from the right side of the PCA (Figure 3.13B, D) are more similar to one another, irrespective of being SE or non-SE genes. Furthermore, heatmaps A and C are not much different from the uniform prior probability over the dN/dS grid assumed by the FUBAR analysis, whereas heatmaps B and D indicate strong purifying selection on the respective genes. We provide additional examples of heatmaps for protein-coding genes associated with measles virus in Appendix A (Figures A.1).

In addition to examining the similarity between specific evolutionary fingerprints of SE and non-SE genes, we also examined the distribution of different virus families in the PCA. Figure 3.14 displays the PCA plots for the nine viruses with the largest numbers of genes in our data. In some cases, the genes for specific virus families localized to either the left or right side of the PCA. Plots for the remaining viruses are provided in the Appendix A (Figures A.2 and A.3).

As such we examined qualities specific to virus families, such as being enveloped or non-enveloped and having DNA or RNA genetic material (Figure 3.15). For instance, Figure 3.15A demonstrates that the clustering of fingerprints by virus in the PCA (Figure 3.14) is potentially confounded by differences between DNA and RNA viruses.

We used amino acid entropy of the multiple sequence alignment (MSA) of all sequence
3.3. Evolutionary fingerprints

Figure 3.12: Principal components analysis (PCA) to visualise the cosine similarity matrix of evolutionary fingerprints. (A) Scree plot showing the first five eigenvalues of the PCA, expressed as percentages of variance explained (totalling 96.9%). (B, C) 2D visualizations of components 1 and 2 (total 87.8%) and components 3 and 4 (total 8.2%), respectively. Each point represents one of 647 genes stratified into SE (red) and non-SE (blue) proteins. (D) A different version of the PCA displayed in (B), highlighting four genes that were selected for visually comparing evolutionary fingerprints in Figure 3.13.
Figure 3.13: Heatmap visualizations of the evolutionary fingerprints for four representative genes that were selected from two distinct clusters in the PCA plot, as shown in Figure 3.12D. Heatmaps (A) and (B) represent surface-exposed (SE) proteins, whereas (C) and (D) represent non-SE proteins. These heatmaps are arranged to the left and right according to the location of clusters in the PCA plot. White indicates that the posterior probability is no different from the uniform prior probability, whereas red and blue indicate higher and lower posterior values, respectively.
3.3. Evolutionary Fingerprints

Figure 3.14: Scatterplots of genes distributed with respect to the first two principal components, coloured by virus family.
Figure 3.15: PCA plots of the first two principal components, with genes coloured by (A) DNA (pink) and RNA (blue) viruses, and by (B) enveloped (cyan) and non-enveloped (indigo) viruses.

for each gene and the tree length, total branch length of the phylogenetic tree constructed for the MSA, as measures of sequence diversity. Figure 3.16 indicates that the clustering of fingerprints in the PCA plot is potentially confounded by sequence diversity. This is consistent with our observation that fingerprints associated with the left cluster in Figure 3.12D were not markedly different from the prior distribution.
Figure 3.16: PCA plots of the first two principal components, with genes coloured by (A) mean amino acid entropy of the multiple sequence alignment, and by (B) the total length of branches in the phylogeny (tree length). Both mean entropy and tree length are measures of genetic variation. Colours closer to blue on the red-blue gradient indicate higher levels of variation.
Chapter 4

Discussion

4.1 Conclusions

By systematically examining all curated human virus genome sequences in the NCBI database, we were able to study the selection pressure acting on 670 genes (177 SE and 493 non-SE) belonging to 52 viruses from 21 virus families. Contrary to our hypothesis that genes encoding surface-exposed (SE) virus proteins are under a higher adaptation than other coding regions in the viral genome, we did not find a characteristic evolutionary profile for such genes when comparing them to viral genes encoding non-SE proteins.

We quantified selection pressures by calculating codon site-specific synonymous (dS) and non-synonymous (dN) substitution rates in FUBAR. Furthermore, we compared both the average dN/dS ratio for each gene, and the percentages of positively and/or negatively selected codon sites. Finally, we examined the statistical distribution of codon-specific selection of each gene, which is known as the ‘evolutionary fingerprint’ [53]. We compared evolutionary fingerprints of all genes in our data by calculating their cosine distances and visualizing the results by PCA. Having used multiple methods to quantify and compare the selection pressure acting of SE versus non-SE genes, we can conclude that being SE is not a feature of proteins that distinctively influences the evolutionary rate of virus proteins.
4.1. Conclusions

However, we found virus-level characteristics such as having DNA versus RNA as its genetic material, and being an enveloped versus non-enveloped virus, significantly influence the extent of selection on the respective virus proteins. To better understand these results, we also looked at dN and dS values independently. It is important to note that in both comparisons (enveloped versus non-enveloped viruses, and DNA versus RNA viruses), there were no significant differences in the synonymous substitution (dS) rates of SE versus non-SE genes. This is consistent with our expectations that synonymous substitutions would be mostly selectively neutral across contexts. With the relative lack of variation in dS, the substantial variation we observed in dN/dS ratios among genes would have been largely driven by variation in dN.

Genes belonging to DNA viruses had significantly higher mean dN/dS ratios than genes from RNA viruses (mean dN/dS = 0.34 and 0.14, respectively; Wilcoxon rank sum test, p = 2.2 × 10^{-16}; Figure 3.7C). In addition, both the percentages of positive and negatively selected sites were significantly different between these groups. An increased level of purifying selection in genes belonging to RNA viruses was observed — both through the higher percentage of the negatively selected sites and the lower dN/dS ratios — compared to genes belonging to DNA viruses. This result may be explained by the higher mutation rates in RNA versus DNA viruses. Since most mutations are deleterious, more mutations would result in more purifying selection.

The other virus characteristic with a significant effect on the dN/dS ratio of SE versus non-SE genes was whether the gene belongs to an enveloped versus a non-enveloped virus. Genes from enveloped viruses had a higher dN/dS ratio than non-enveloped viruses (mean=0.3 and 0.22, respectively, Wilcoxon rank sum test, p = 1.4×10^{-5}) (Figure 3.7 B) The significant difference remained for the percentage of negatively selected sites but intriguingly not for positively selected sites.

The difference in selection observed between enveloped and non-enveloped viruses are likely confounded by DNA and RNA viruses. Genes from RNA viruses are more likely to also be associated with enveloped viruses in our data set; in other words, there were fewer genes associated with non-enveloped RNA viruses than expected by chance. Since genes of
RNA viruses have more variation due to higher mutation rates, there would be more synonymous genetic variation against which we could detect purifying selection (from the absence of nonsynonymous variation). Conversely, we may have had limited power to detect purifying selection in genes of DNA viruses, where lower mutation rates resulted in less divergence among sequences at synonymous sites.

4.2 Limitations

4.2.1 Codon Level Selection

This study used codon based methods to quantify selection pressure. However, there are a number of ways that genes can undergo adaptation without any changes at the codon level that this study would not be able to detect.

For example, viruses have evolved a number of independent antagonists to counteract the mammalian Tetherin gene (BST-2 or CD317) which inhibits the release of enveloped viruses from infected cells, such as Vpu and Nef proteins [100, 47]. One reported mechanism by which Tetherin counteracts its restriction by the Nef protein is due to a 5 amino-acid deletion, which is an example of adaptation not involving nucleotide substitutions at the codon level [60]. Another example was observed in the human Tripartite motif (TRIM)–encoding proteins, which through gene fusion it was able to acquire a novel protein domain with novel capsid specificity (Trim-Cyp) [80]. As such it is possible for the evolutionary arms race to involve mutations other than what can be characterized as nonsynonymous and synonymous substitutions at the codon level. Consequently, codon-based methods of selection analysis would fail to detect such processes.

4.2.2 Variable proportion of SE genes exposed to the outside of the virion

It is important to note that not all of the protein encoded by a gene we classify as SE is actually exposed on the surface. A protein of an enveloped virus positioned on the surface of the virus comprises of a transmembrane region and an internal region, in addition to the surface
exposed region. As such, if positively selected sites are limited to only regions of a SE gene encoding amino acids on the outside of the virion, then the signal for detecting the arms race dynamic may be masked by the lack of selection affecting the rest of the gene.

For example, the envelope protein E (YP_009430300.1) of Zika virus from the Flaviviridae virus family (Figure 4.1) has two transmembrane regions separated by a short internal region and followed by an external (SE) region attached to the membrane protein. In our analysis, the evolutionary fingerprint of the gene for Zika virus enveloped protein E was labeled as that of a SE protein — in fact, only a relatively short section of that gene sequence encodes the portion of the protein that is surface exposed.

In Figure 4.1, we can also see the proportions of SE genes that are transmembrane, *i.e.*, on the inside and outside of the envelope, varies between different proteins. In addition to the position of the protein with respect to the plasma membrane, the surface exposed proportion can also be affected by other proteins. Furthermore, these structures are not constant, and the exposure of proteins to the surface can change. For example, the configuration of the envelope E protein changes at the immature versus mature state of flaviviruses.

### 4.2.3 Functional Constraints

The lack of any detectable signal of evolutionary dynamic unique to SE genes could be potentially explained by the evolutionary constraints imposed by other cellular function on these SE proteins which limit the number of codon sites with the freedom to participate in the host-virus arms race. As such the effects of evolutionary arms race can be limited to a small number of sites on surface proteins. With the effect of evolutionary arms race limited to a small region it is possible to have gone undetected as it may have been masked but the signal from the rest of the sequence, likely experiencing purifying selection under function constraints.

Many studies have found elevated adaptive evolution at the sites of interaction with viral antagonists in both host and virus proteins involved in evolutionary arms race. Wang *et al.* [103] studied viral receptor genes on primates, and found many of the positively selected sites to overlap with the virus-receptor interacting interfaces. This supports the idea that the effects of
Figure 4.1: Flaviviruses (Zika virus) transmembrane proteins [76] (A) Number of transmembrane Falvivirus proteins including the envelope glycoprotein (E) the membrane protein (M) and the nonstructural proteins NS1 and NS2. (b) The E protein and it’s transmembrane, internal and external regions.

arms race gene would likely be limited to the few sites directly involved in contact with viral antagonists.

4.2.4 Other sources of diversifying selection

Another reason behind lack of a significant difference in dN/dS ratio between SE and non-SE genes is that there are other sources of diversifying selection not specific to being surface exposed. There has also been evidence of positive selection in nonstructural viral proteins. For instance, the study by Sironi et al. [94] identified nonstructural proteins as the preferential targets of positive selection in Zika virus. It is also interesting to note that the positively selected sites were located in functionally important regions including immune epitopes. Even though structural protein, Envelope E of Zika virus would be considered a likely target of positively selected sites as it is a major target for immune responses involved in antigenic selection [27] and its involvement in essential steps of host cell entry and infection [55]. The study by Sironi et al. found no evidence of episodic or pervasive positive selection in branches of flaviviruses.
or in the Zika virus. It is interesting that most neutralizing antibodies act against the E protein of flavivirus, and yet it is the non-neutralizing antibodies against the nonstructural protein NS1, absent from the virion, that have been found to prevent lethal infection by many of the flaviviruses including Dengue virus (DENV) and Japanese encephalitis virus (JEV) [18, 58].

4.3 Future directions

Though we tried to use systematic ways to label virus genes as SE, the analysis would benefit from a manually curated database of SE viral genes. We found labeling SE genes using only the literature challenging as such information from structural predictors should also be taken into account when creating a manual database of SE viral genes.

It would also be very interesting to account for the variation in the proportions of SE genes that are exposed on the outside of the virion. The selection pressure can be measured for only the regions of SE genes that are on the outside of the virion. For enveloped viruses, the SE regions can be extracted using transmembrane predictors such as DeepTMHMM [39] or TMbed [9] or manually using previous studies of structural annotations. However, measuring selection pressure for such short sequences will be challenging. As such, a possible alternative approach is to account for the portion of each SE genes that is positioned on the outer membrane of a virion, and to normalise the dN/dS calculations.

We could also learn a lot from identifying clusters of genes with similar evolutionary fingerprints. Average-linkage hierarchical clustering could be used, as it was in the study by Murrell et al. [69] to identify an evolutionary fingerprint unique to canonical arms race genes. Similar nested evolutionary fingerprints can also be recognized through unsupervised machine learning algorithms.
Bibliography


[16] Seth Carbon, Eric Douglass, Benjamin M. Good, Deepak R. Unni, Nomi L. Harris, Christopher J. Mungall, Siddartha Basu, Rex L. Chisholm, Robert J. Dodson, Eric Hartline, Petra Fey, Paul D. Thomas, Laurent Philippe Alou, Dustin Ebert, Michael J. Kesling, Huaiyu Mi, Anushya Muruganujan, Xiaosong Huang, Tremayne Mushyahama, Sandra A. LaBonte, Deborah A. Siegele, Giulia Antonazzo, Helen Attrill, Nick H. Brown, Phani Garapati, Steven J. Marygold, Vitor Trovisco, Gil dos Santos, Kathleen
BIBLIOGRAPHY


Appendix A

Supplementary Figures and Tables
Table A.1: List of viruses being studied.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Length</th>
<th>Virus name</th>
<th>Family</th>
<th>Order</th>
<th>Neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_001348</td>
<td>12484</td>
<td>Human alphaherpesvirus 3</td>
<td>Herpesviridae</td>
<td>Herpesvirales</td>
<td>204</td>
</tr>
<tr>
<td>NC_001355</td>
<td>7902</td>
<td>Human papillomavirus type 6b</td>
<td>Papillomaviridae</td>
<td>Zurhausenvirales</td>
<td>371</td>
</tr>
<tr>
<td>NC_001526</td>
<td>7906</td>
<td>Human papillomavirus type 16</td>
<td>Papillomaviridae</td>
<td>Zurhausenvirales</td>
<td>5035</td>
</tr>
<tr>
<td>NC_001593</td>
<td>7856</td>
<td>Human papillomavirus type 53</td>
<td>Papillomaviridae</td>
<td>Zurhausenvirales</td>
<td>120</td>
</tr>
<tr>
<td>NC_000940</td>
<td>7320</td>
<td>Cowden virus</td>
<td>Caliciviridae</td>
<td>Picornavirales</td>
<td>183</td>
</tr>
<tr>
<td>NC_002200</td>
<td>15384</td>
<td>Mumps orthorubulavirus</td>
<td>Paramyxoviridae</td>
<td>Mononegavirales</td>
<td>719</td>
</tr>
<tr>
<td>NC_001542</td>
<td>11932</td>
<td>Rabies lyssavirus</td>
<td>Rhabdoviridae</td>
<td>Mononegavirales</td>
<td>2186</td>
</tr>
<tr>
<td>NC_001781</td>
<td>15225</td>
<td>Human orthopneumovirus</td>
<td>Pneumoviridae</td>
<td>Mononegavirales</td>
<td>2034</td>
</tr>
<tr>
<td>NC_001498</td>
<td>15894</td>
<td>Measles morbillivirus</td>
<td>Paramyxoviridae</td>
<td>Mononegavirales</td>
<td>415</td>
</tr>
<tr>
<td>NC_003045</td>
<td>31028</td>
<td>Bovine coronavirus</td>
<td>Coronavirus</td>
<td>Nidovirales</td>
<td>336</td>
</tr>
<tr>
<td>NC_001802</td>
<td>9181</td>
<td>Human immunodeficiency virus 1</td>
<td>Retroviridae</td>
<td>Ortervirales</td>
<td>8750</td>
</tr>
<tr>
<td>NC_001653</td>
<td>1682</td>
<td>Hepatitis delta virus</td>
<td>Kolmioviridae</td>
<td>incertae sedis</td>
<td>671</td>
</tr>
<tr>
<td>NC_004162</td>
<td>11826</td>
<td>Chikungunya virus</td>
<td>Togaviridae</td>
<td>Martellivirales</td>
<td>929</td>
</tr>
</tbody>
</table>
Table A.1 (continued).

<table>
<thead>
<tr>
<th>Accession</th>
<th>Length</th>
<th>Virus name</th>
<th>Family</th>
<th>Order</th>
<th>Neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_004296</td>
<td>3402</td>
<td>Lassa mammarenavirus</td>
<td>Arenaviridae</td>
<td>Bunyavirales</td>
<td>344</td>
</tr>
<tr>
<td>NC_002058</td>
<td>7440</td>
<td>Enterovirus C</td>
<td>Picornaviridae</td>
<td>Picornavirales</td>
<td>872</td>
</tr>
<tr>
<td>NC_001612</td>
<td>7413</td>
<td>Enterovirus A</td>
<td>Picornaviridae</td>
<td>Picornavirales</td>
<td>2670</td>
</tr>
<tr>
<td>NC_001472</td>
<td>7389</td>
<td>Enterovirus B</td>
<td>Picornaviridae</td>
<td>Picornavirales</td>
<td>734</td>
</tr>
<tr>
<td>NC_003899</td>
<td>11675</td>
<td>Eastern equine encephalitis virus</td>
<td>Togaviridae</td>
<td>Martellivirales</td>
<td>455</td>
</tr>
<tr>
<td>NC_001547</td>
<td>11703</td>
<td>Sindbis virus</td>
<td>Togaviridae</td>
<td>Martellivirales</td>
<td>229</td>
</tr>
<tr>
<td>NC_001430</td>
<td>7390</td>
<td>Enterovirus D</td>
<td>Picornaviridae</td>
<td>Picornavirales</td>
<td>785</td>
</tr>
<tr>
<td>NC_004455</td>
<td>9518</td>
<td>Simian immunodeficiency virus SIV-mnd 2</td>
<td>Retroviridae</td>
<td>Ortervirales</td>
<td>550</td>
</tr>
<tr>
<td>NC_001437</td>
<td>10976</td>
<td>Japanese encephalitis virus</td>
<td>Flaviviridae</td>
<td>Amarillovirales</td>
<td>374</td>
</tr>
<tr>
<td>NC_001563</td>
<td>10962</td>
<td>West Nile virus</td>
<td>Flaviviridae</td>
<td>Amarillovirales</td>
<td>1894</td>
</tr>
<tr>
<td>NC_001461</td>
<td>12573</td>
<td>Bovine viral diarrhea virus 1</td>
<td>Flaviviridae</td>
<td>Amarillovirales</td>
<td>147</td>
</tr>
<tr>
<td>NC_001414</td>
<td>8419</td>
<td>Bovine leukemia virus</td>
<td>Retroviridae</td>
<td>Ortervirales</td>
<td>154</td>
</tr>
<tr>
<td>NC_004718</td>
<td>29751</td>
<td>SARS coronavirus Tor2</td>
<td>Coronaviridae</td>
<td>Nidovirales</td>
<td>368</td>
</tr>
<tr>
<td>NC_005218</td>
<td>1696</td>
<td>Hantaan orthohantavirus</td>
<td>Hantaviridae</td>
<td>Bunyavirales</td>
<td>273</td>
</tr>
</tbody>
</table>
Table A.1 (continued).

<table>
<thead>
<tr>
<th>Accession</th>
<th>Length</th>
<th>Virus name</th>
<th>Family</th>
<th>Order</th>
<th>Neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_005300</td>
<td>5366</td>
<td>Crimean-Congo hemorrhagic fever orthonairovirus</td>
<td>Nairoviridae</td>
<td>Bunyavirales</td>
<td>159</td>
</tr>
<tr>
<td>NC_001436</td>
<td>8507</td>
<td>Human T-cell leukemia virus type I</td>
<td>Retroviridae</td>
<td>Ortervirales</td>
<td>299</td>
</tr>
<tr>
<td>NC_009334</td>
<td>172764</td>
<td>Human herpesvirus 4 type 2</td>
<td>Herpesviridae</td>
<td>Herpesvirales</td>
<td>658</td>
</tr>
<tr>
<td>NC_006273</td>
<td>235646</td>
<td>Human beta herpesvirus 5</td>
<td>Herpesviridae</td>
<td>Herpesvirales</td>
<td>344</td>
</tr>
<tr>
<td>NC_014397</td>
<td>6404</td>
<td>Rift Valley fever virus</td>
<td>Phenuiviridae</td>
<td>Bunyavirales</td>
<td>150</td>
</tr>
<tr>
<td>NC_018136</td>
<td>6368</td>
<td>SFTS virus HB29</td>
<td>Phenuiviridae</td>
<td>Bunyavirales</td>
<td>100</td>
</tr>
<tr>
<td>NC_019843</td>
<td>30119</td>
<td>Middle East respiratory syndrome-related coronavirus</td>
<td>Coronaviridae</td>
<td>Nidovirales</td>
<td>672</td>
</tr>
<tr>
<td>NC_003977</td>
<td>3182</td>
<td>Hepatitis B virus</td>
<td>Hepadnaviridae</td>
<td>Blubervirales</td>
<td>12650</td>
</tr>
<tr>
<td>NC_012532</td>
<td>10794</td>
<td>Zika virus</td>
<td>Flaviviridae</td>
<td>Amarillovirales</td>
<td>768</td>
</tr>
<tr>
<td>NC_030791</td>
<td>9443</td>
<td>Hepatitis C virus genotype 7</td>
<td>Flaviviridae</td>
<td>Amarillovirales</td>
<td>2299</td>
</tr>
<tr>
<td>NC_038270</td>
<td>15462</td>
<td>Simian Agent 10</td>
<td>Paramyxoviridae</td>
<td>Mononegavirales</td>
<td>406</td>
</tr>
<tr>
<td>NC_039237</td>
<td>12513</td>
<td>Bovine viral diarrhea virus 2</td>
<td>Flaviviridae</td>
<td>Amarillovirales</td>
<td>107</td>
</tr>
<tr>
<td>NC_039199</td>
<td>13350</td>
<td>Human metapneumovirus</td>
<td>Pneumoviridae</td>
<td>Mononegavirales</td>
<td>160</td>
</tr>
</tbody>
</table>
Table A.1 (continued).

<table>
<thead>
<tr>
<th>Accession</th>
<th>Length</th>
<th>Virus name</th>
<th>Family</th>
<th>Order</th>
<th>Neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>NC_048214</td>
<td>27754</td>
<td>Duck coronavirus</td>
<td>Coronaviridae</td>
<td>Nidovirales</td>
<td>442</td>
</tr>
<tr>
<td>NC_006383</td>
<td>15948</td>
<td>Peste des petits ruminants virus</td>
<td>Paramyxoviridae</td>
<td>Mononegavirales</td>
<td>109</td>
</tr>
<tr>
<td>NC_001357</td>
<td>7857</td>
<td>Alphapapillomavirus 7</td>
<td>Papillomaviridae</td>
<td>Mononegavirales</td>
<td>281</td>
</tr>
<tr>
<td>NC_001699</td>
<td>5130</td>
<td>JC polyomavirus</td>
<td>Polyomaviridae</td>
<td>Mononegavirales</td>
<td>678</td>
</tr>
<tr>
<td>NC_001427</td>
<td>2319</td>
<td>Chicken anemia virus</td>
<td>Anellovirida</td>
<td>incertae sedis</td>
<td>301</td>
</tr>
<tr>
<td>NC_001434</td>
<td>7176</td>
<td>Orthohepeivirus A</td>
<td>Hepevirida</td>
<td>Hepelivirales</td>
<td>829</td>
</tr>
<tr>
<td>NC_001489</td>
<td>7478</td>
<td>Hepatovirus A</td>
<td>Picornaviridae</td>
<td>Picornavirales</td>
<td>138</td>
</tr>
<tr>
<td>NC_004441</td>
<td>7388</td>
<td>Porcine enterovirus 9</td>
<td>Picornaviridae</td>
<td>Picornavirales</td>
<td>103</td>
</tr>
<tr>
<td>NC_005148</td>
<td>1768</td>
<td>Porcine circovirus 2</td>
<td>Circoviridae</td>
<td>Cirlivirales</td>
<td>3117</td>
</tr>
<tr>
<td>NC_010956</td>
<td>35083</td>
<td>Human mastadenovirus D</td>
<td>Adenovirida</td>
<td>Rowavirales</td>
<td>197</td>
</tr>
<tr>
<td>NC_011203</td>
<td>35343</td>
<td>Human mastadenovirus B</td>
<td>Adenovirida</td>
<td>Rowavirales</td>
<td>256</td>
</tr>
<tr>
<td>NC_011501</td>
<td>1105</td>
<td>Rotavirus A</td>
<td>Reoviridae</td>
<td>Reovirales</td>
<td>4097</td>
</tr>
<tr>
<td>NC_007547</td>
<td>3309</td>
<td>Rotavirus C</td>
<td>Reoviridae</td>
<td>Reovirales</td>
<td>101</td>
</tr>
<tr>
<td>NC_001538</td>
<td>5153</td>
<td>Human polyomavirus 1</td>
<td>Polyomaviridae</td>
<td>Sepolyvirales</td>
<td>363</td>
</tr>
</tbody>
</table>
Figure A.1: Measles virus heatmap
Figure A.2: Scatterplots of genes distributed with respect to the first two principal components, coloured by virus family.
Figure A.3: Scatterplots of genes distributed with respect to the first two principal components, coloured by virus family.