Prediction of interacting motifs within the protein subunits of picornavirus capsids

A thesis submitted in partial fulfilment of the requirement for the degree

of

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework / Thesis

in

Bioinformatics and Computational, Molecular Biology

In the Department of Biochemistry and Microbiology

Faculty of Science

by

Caroline Jane Ross

February 2015

Abstract

The *Picornaviridae* family contains a number of pathogens which are economically important including Poliovirus, Coxsakievirus, Hepatitis A Virus, and Foot-and-Mouth-Disease-Virus. Recently the emergence of novel picornaviruses associated with gastrointestinal, neurological and respiratory diseases in humans has been reported. Although effective vaccines for viruses such as FMDV, PV and HAV have been developed there are currently no antivirals available for the treatment of picornavirus infections. Picornaviruses proteins are classified as: the structural proteins VP1, VP2, VP3 and VP4 which form the subunits of the viral capsid and the replication proteins which function as proteases, RNA-polymerases, primers and membrane binding proteins. Although the host specificity and viral pathogenicity varies across members of the family, the icosahedral capsid is highly conserved. The capsid consists of 60 protomers, each containing a single copy of VP1, VP2 and VP3. A fourth capsid protein, VP4, resides on the internal side of the capsid. Capsid assembly is integral to life-cycle of picornaviruses; however the process is complex and not fully-understood. The overall aim of the study was to broaden the understanding of the evolution and function of the structural proteins across the *Picornaviridae* family. Firstly a comprehensive analysis of the phylogenetic relationships amongst the individual structural proteins was performed. The functions of the structural proteins were further investigated by an exhaustive motif analysis. A subsequent structural analysis of highly conserved motifs was performed with respect to representative enteroviruses, Foot-and-Mouth-Disease-Virus and Theiler's Virus. This was supplemented by the *in silico* prediction of interacting residues within the crystal structures of these protomers. Findings in this study suggest that the capsid proteins may be evolving independently from the replication proteins through possible inter-typic recombination of functional protein regions. Moreover the study predicts that protomer assembly may be facilitated through a network of multiple subunit-subunit interactions. Multiple conserved motifs and principle residues predicted to facilitate capsid subunit-subunit interactions were identified. It was also concluded that motif conservation may support the theory of inter-typic recombination between closely related virus sub-types. As capsid assembly is critical to the viral life-cycle, the principle interacting motifs may serve as novel drug targets for the antiviral treatment of picornavirus infections. Thus the findings in the study may be fundamental to the development of treatments which are more economically feasible or clinically effective than current vaccinations.

Declaration

The research described in this thesis was carried out as part of the one-year MSc coursework and research thesis programme in Bioinformatics and Computational Molecular Biology, from 15 July 2014 to 15 February 2015 under the supervision of Prof Özlem Taştan Bishop and Dr Caroline Knox.

I, Caroline Jane Ross, declare that this thesis submitted to Rhodes University is wholly my own work and has not previously been submitted for a degree at this or any other institution.

Signature…….. Date:……

Acknowledgements

1. To my supervisor, Prof Özlem Taştan Bishop, thank you for all the help and guidance in decisions concerning the methodology and data analysis of this project. I also thank you for all the patience and time you gave to help me with the writing the final report. Your help was invaluable and I have learnt so much this year.

2. To my co-supervisor, Dr Caroline Knox, thank you for all the assistance and guidance in the decisions regarding the purpose of this study. Your vast knowledge and understanding of picornaviruses helped with the identification of a knowledge gap and the interpretation of significant results. I would also like to thank you for your input into my written reports, conclusions and abstracts.

3. To Ngonidzashe Faya, thank you for your continuous assistance of the MEME and MAST analysis and the interpretation of results.

4. To Vuyani Moses, thank you for all your assistance in the methodology and interpretation of phylogenetic analyses.

5. David Brown, thank you for you continuous assistance in the management of large datasets and guidance in utilizing programs installed on the lab server.

6. The financial assistance of the National Research Foundation (NRF) towards this research is also acknowledged.

7. I also acknowledge Rhodes University for providing additional funding through the Prestigious Henderson Scholarship

Contents

List of Figures

List of Tables

List of abbreviations

AIC	Akaike information criterion
BIC	Bayesian information criterion
CDS	Coding Sequence
ER	Endoplasmic reticulum
Hsp	Heat shock protein
HY	Hydrophobic
ICTV	International Committee of Virus Taxonomy
ID	Identity
IM	Interacting Motif
IRES	Internal ribosome entry site
lnL	Log Likelihood
MAST	Motif Alignment and Search Tool
MEME	Multiple-EM for motif Elicitation
MMH	Main chain-main chain hydrogen bond
MP	Maximum Parsimony
ML	Maximum-likelihood
MSA	Multiple sequence alignment
MSH	Main chain-side chain hydrogen bond
mRNA	Messenger RNA
NJ	Neighbour Joining
NNI	Nearest Neighbour Interchange
ORF	Open reading frame
PDB	Protein Databank
PIC	Protein Interaction Calculator
Poly-A	Poly-adenosine
Res	Residue
RNA	Ribonucleic acid
SLiM	Short Linear Motif
SSH	Side chain-side chain hydrogen bond
ViPR	Virus Pathogen Database and Analysis Resource

Table of Amino Acid Abbreviations

Amino Acid	Three-Letter Abbreviation	One-Letter Abbreviation
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

Units and symbols

%	percent
aa	Amino acid
kDa	kilo Daltons
nt	nucleotide

List of webservers and applications

1. Jalview

   www.jalview.org

2. MEGA v6.06

3. Motif alignment and Search Tool (MAST)

   http://meme.nbcr.net/meme/

4. Multiple-EM for Motif Elicitation (MEME)

   http://meme.nbcr.net/meme/

5. MUSCLE

   http://www.ebi.ac.uk/Tools/msa/muscle/

6. PROMALS 3D

   http://prodata.swmed.edu/promals3d/promals3d.php

7. Protein Databank (PDB)

   www.rcsb.org

8. Protein Interaction Calculator (PIC)

   http://pic.mbu.iisc.ernet.in/

9. PyMOL

   www.pymol.org

10. Python

    https://www.python.org/

11. Virus Pathogen Database and Analysis Resource

    http://www.viprbrc.org/

1. Review of literature

1.1 Introduction

The name Picornavirus is derived from the Spanish word *pico*, meaning point. The word *pico* refers to the small size of the virions, and in particular the name Picornavirus refers to small, RNA-containing virions (Racaniello, 2007). The viruses belonging to the order *Picornavirales* share distinctive characteristics with regard to capsid morphology, genome organisation and the translation of specific viral proteins (Racaniello, 2007). The order *Picornavirales* consists of five families: *Picornaviridae, Iflaviridae, Marnaviridae, Dicistroviridae and Secoviridae* (Gall *et al.,* 2007). Although each family exhibits the typical characteristics of the *Picornavirales* order, each family have a unique host range and unique characteristics. The Picornaviridae family currently consists of 26 genera: *Aphthovirus, Aquamavirus, Avihepatovirus, Avisivirus, Cardiovirus, Cosavirus, Dicipivirus, Enterovirus, Erbovirus, Gallivirus, Hepatovirus, Hunnivirus, Kobuvirus, Megrivirus, Mischivirus, Mosavirus, Oscivirus, Parechovirus, Pasivirus, Passerivirus, Rosavirus, Salivirus, Sapelovirus, Senecavirus, Teschovirus, Tremovirus*, with a total of 46 sub-classified species (ICTV, 2014). A summary of the classification of these viruses in presented in Table 1.1. The family encompasses a number of viruses which are considered to be of high economic or clinical importance. These viruses include Poliovirus, Coxsakievirus A and B, Echovirus 71, Hepatitis A Virus, Foot-and-Mouth-Disease Virus (FMDV) (Racaniello, 2007). Viruses of the *Picornaviridae* family infect a range of individual host species including: human, bovine, porcine, equine, simian, canine, murine, avian and certain marine species (Racaniello, 2007).

Table 1.1. The classification of picornaviruses. The family currently consists of 26 classified genera and 46 species. (adapted from ICTV, 2014).

Genus	Species			
Aphthovirus	*Bovine rhinitis A virus*	*Equine rhinitis A virus*	*Foot-and-mouth disease virus*	
	Bovine rhinitis B virus			
Aquamavirus	*Aquamavirus A*			
Avihepatovirus	*Duck hepatitis A virus*			
Avisivirus	*Avisivirus A*			
Cardiovirus	*Encephalomyocarditis virus* *Theilovirus*			
Cosavirus	*Cosavirus A*			
Dicipivirus	*Cadicivirus A*			
Enterovirus	*Enterovirus A*	*Enterovirus D*	*Enterovirus G*	*Rhinovirus A*
	Enterovirus B	*Enterovirus E*	*Enterovirus H*	*Rhinovirus B*
	Enterovirus C	*Enterovirus F*	*Enterovirus J*	*Rhinovirus C*
Erbovirus	*Equine rhinitis B virus*			
Gallivirus	*Gallivirus A*			
Hepatovirus	*Hepatitis A virus*			
Hunnivirus	*Hunnivirus A*			
Kobuvirus	*Aichivirus A* *Aichivirus B* *Aichivirus C*			
Megrivirus	*Melegrivirus A*			
Mischivirus	*Mischivirus A*			
Mosavirus	*Mosavirus A*			
Oscivirus	*Oscivirus A*			
Parechovirus	*Human parechovirus*	*Ljungan virus*		
Pasivirus	*Pasivirus A*			
Passerivirus	*Passerivirus A*			
Rosavirus	*Rosavirus A*			
Salivirus	*Salivirus A*			
Sapelovirus	*Avian sapelovirus*	*Porcine sapelovirus*	*Simian sapelovirus*	
Senecavirus	*Seneca Valley virus*			
Teschovirus	*Porcine teschovirus*			
Tremovirus	*Avian encephalomyelitis virus*			

1.2 Genome organisation

Viruses of the *Picornaviridae* family have a monopartite positively sensed RNA genome, between 7-8kb in length. Although the viral RNA strand has a 5' covalently bonded viral protein (VPg), it does not have a 5' cap. The VPg protein acts as a primer, playing a role in the initiation of viral RNA synthesis (Agol, 2001). Due to the absence of the 5' cap the translation is initiated through the recognition of the viral internal ribosome entry site (IRES) by host cell ribosomes (Racaniello, 2007). The IRES is located in the 5' untranslated region (UTR) of the genome. The genome has a single open reading frame (ORF) which is initially translated into a single polyprotein, followed by a downstream 3' UTR and a poly(A)-tail. The poly(A)-tail functions to stabilise the RNA genome, protecting it from nuclease activity (Harris *et al*., 1994).

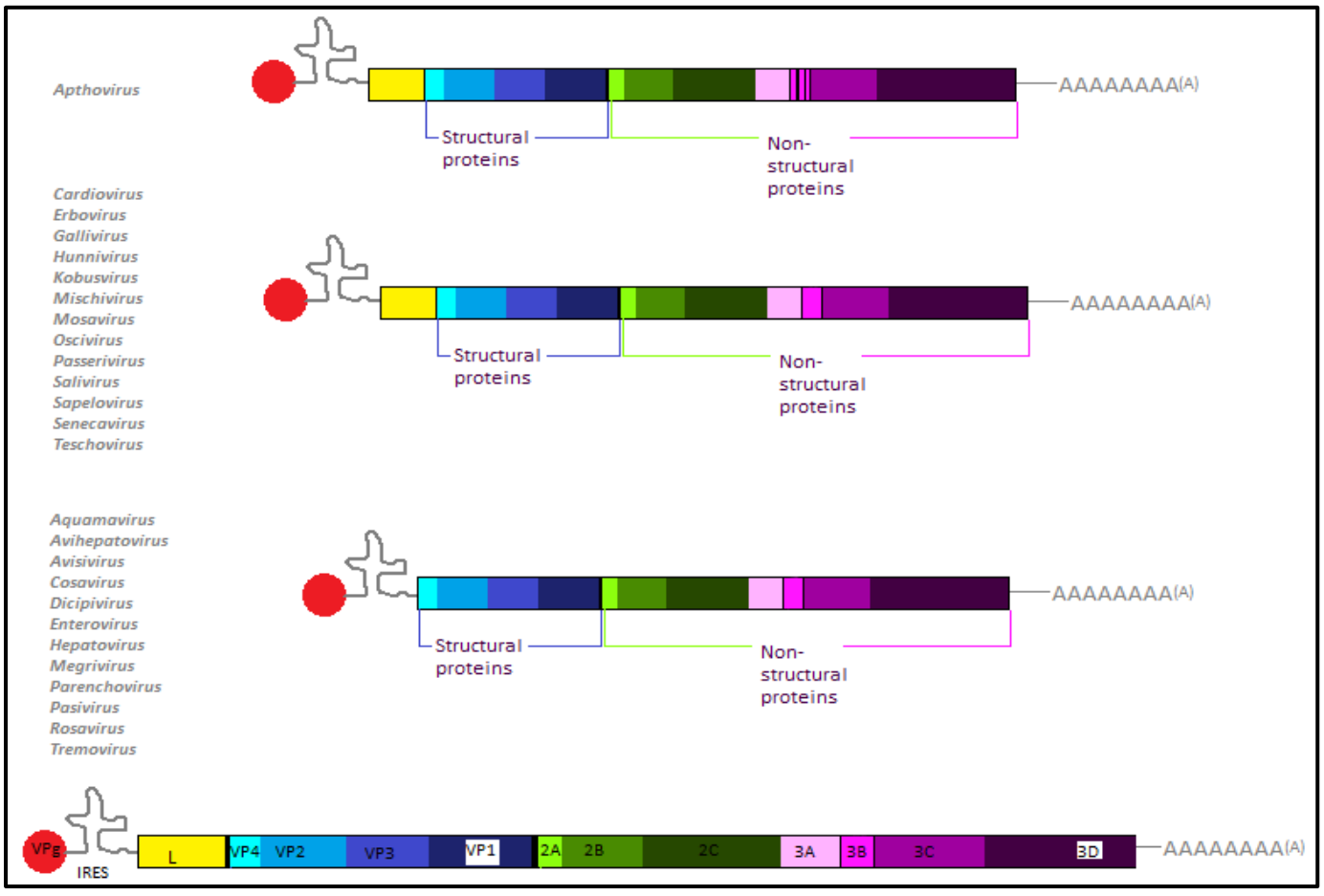


Figure 1.1. Schematic diagram of the genome organisations shared by members of the *Picornaviridae* family (adapted from Ehrenfeld *et al*., 2010) . The diagrams shows the additional L protein indicated in yellow, the 5' VPg protein and IRES, the polyprotein consisting of structural and non-structural proteins and the 3' poly(A)-tail. 2). The *Aphthovirus* genome contains three copies of viral protein 3B, as shown in the top diagram. Proteins are colour coded, as indicated by the key.

All picornaviruses encode for the same structural and non-structural proteins. However viruses belonging to the genera: *Cardiovirus, Erbovirus, Gallivirus, Hunnivirus, Kobuvirus, Mischivirus, Mosavirus, Oscivirus, Passerivirus, Salivirus, Sapelovirus, Senecavirus* and *Teschovirus* all encode for an additional non-structural protein- the 5' leader protein L. Furthermore viral genomes of the *Aphthovirus* genus encode an additional L protein, as well as tandem copies of the 3B protein. A comparison of viral genomes from all 26 genera is depicted in Figure 1.1.

1.3 Translation and processing of viral proteins

As previously stated, the genome has a single open reading frame (ORF) which is initially translated into a single polyprotein. The structural precursor protein, P1, is subsequently cleaved from the P2/P3 domains (non-structural), by the 2A viral protease (Racaniello, 2007). Consequently the precursor proteins are cleaved by the viral protease 3C and 3CD. The structural proteins VP0, VP1, VP3 form the products of precursor P1 and the non-structural/replication proteins form the products of precursors P2 and P3. The structural protein VP0 is later cleaved into VP4 and VP2, subsequent to virus assembly, to elicit maturation of the virus particle (Racaniello, 2007). The steps of proteolytic cleavage into to the final structural and non-structural proteins are depicted in Figure 1.2.

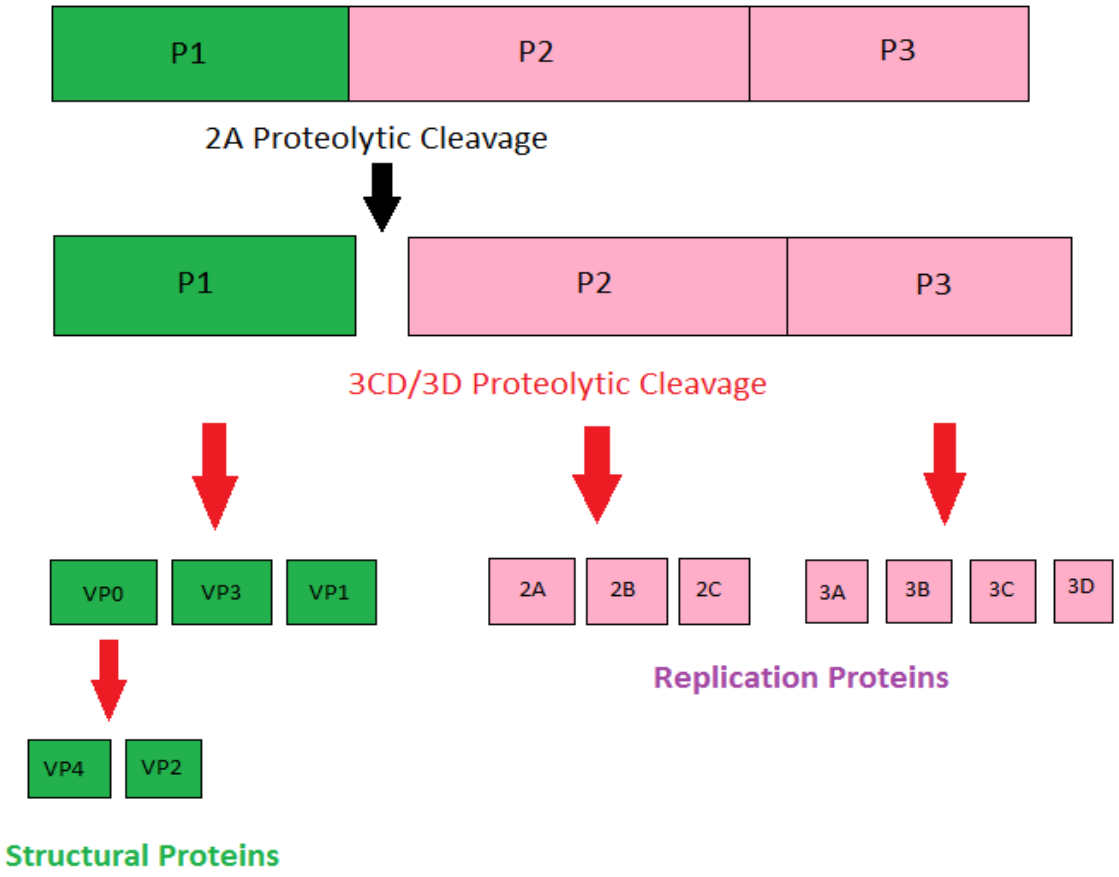


Figure 1.2. Schematic illustration of picornavirus polyprotein processing. 1) Cleavage of initial polyprotein by protease 2A yielding precursors P1 and P2-P3. 2) Cleavage of P1, P2 and P3 by viral proteases 3CD and 3C to yield the different structural and non-structural proteins. 3) Subsequent 3CD/3C cleavage of capsid protein VP0 into VP2 and VP4, resulting in a mature virus capsid (adapted from Toyoda *et al*., 1986).

1.4 Capsid Morphology of Picornaviruses

The viral capsid can be described as a non-enveloped icosahedral capsid, which has pseudo-T = 3 symmetry and is approximately 30nm in diameter. The capsid consists of 60 protomers, each of which contain three different 8-stranded beta-barrel domains. Each domain is formed as a result of the folding of one of three structural proteins, namely VP1, VP2 or VP3. A fourth structural protein, VP4, is found on the internal side of the capsid (Racaniello, 2007). This structure has been depicted in Figure 1.3. The most external and accessible surface protein is VP1. It has a complex 3D structure consisting of two anti-parallel β-sheets and is approximately 36 kDa in size. The protein is primarily responsible for binding to the host cell receptor. Furthermore the protein contains the highest number of neutralisation sites and thus

is considered to be the antigenic determinant (Collen et al., 1991). The antigenicity of VP1 is described in detail in a later section of this review.

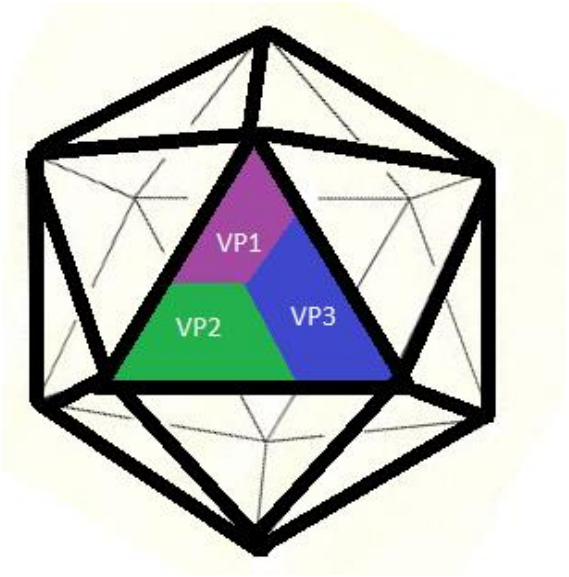


Figure 1.3. The basic icosahedral structure of a picornavirus. External structural proteins are shown: Purple) VP1, Green) VP2, Blue) VP3 (adapted from Acheson, 2007).

1.5 Antigenicity of the capsid proteins

The term antigenicity refers to the ability of the protein to induce an immune response within its host. Upon entry, the host will recognise the protein as foreign, triggering the production of antibodies against the protein. An antibody is a molecule that is produced within the B-lymphocytes of the host. Each antibody that is produced is unique and will recognise and bind a specific region, known as an epitope, of the foreign antigen/protein, in an attempt to block or neutralise the infectivity of the protein. The more antigenic the foreign protein, the greater the immune response it will elicit, thus the larger the yield of antibodies produced against the protein (Pellequer *et al*., 1991).

In general most B-cell epitopes are composed of discontinuous regions within the protein, which are brought into the correct spacial proximity and conformation during protein folding (Pellequer *et al*., 1991). However it has been found that approximately 10% of the epitopes of a general protein will comprise of a single continuous stretch of amino acids within the polypeptide (Pellequer *et al*., 1991). This type of epitope is known as a linear B-cell epitope. Although this project focused on the prediction of functional motifs involved in the assembly of the viral capsid, the structural proteins are also primarily responsible for host cell receptor binding and the activation of the immune response. Thus it is important to distinguish

between motifs which may elicit an immune response and those which may be functional in subunit-subunit interactions.

The external capsid proteins, VP1, VP2 and VP3, have all been reported to have antigenic properties (Racaniello, 2007). However literature suggests that VP1 is the most external and immunogenic of the capsid proteins. The VP1 protein of picornaviruses is a wedged shaped protein which consists of eight stranded β-barrels, as well as connecting loops between the barrels and the N and C termini of the protein (Rueckert, 2001). VP1 has been found to play a role in host cell attachment and is considered to most external and immunodominant of the four capsid proteins (Rossman *et al*., 1985). This is supported by Eun *et al* (1998), in which the attachment of Encephalomyocarditis Virus (EMC-D) to pancreatic beta cells was facilitated by the surface protein VP1. The 3D structure of FMDV VP1 was determined through homology modelling (Liu *et al*., 2011). An analysis of this structure revealed a 23.4 kDa protein, which consisted of an internal β-sheet, comprised of consecutive hydrophobic residues and a hydrophilic surface constituted by an alpha helix and a β-sheet. A bioinformatic analysis performed on the determined 3D structure predicted possible B-cell epitope regions to range the amino acids: VP1a:22-32aa, VP1b:41-50aa, VP1c:94-105aa, VP1d:137-149aa, VP1e:196-205aa.The bioinformatics methods used to analyse the protein included Kyte–Doolittle, Karplus–Schulz, Emini and Jameson–Wolf.

The mapping of the epitopes of TMEV DA was conducted in 1985 and revealed that the major epitopes resided on the VP1 capsid protein (Nitayaphan *et al*.,1985). A study conducted by Cameron *et al* (2000), involved the comparison of the antigenicity of the nine mature TMEV DA viral proteins, including structural and non-structural proteins. The study concluded that only anti-VP1 antibodies elicited neutralisation of the virus. It has been observed that the amino acid variations within the connecting loops are responsible for the unique morphology and antigenicity of the different Picornaviruses (Ruekert, 2001). This observation is supported by Varrasso *et al* (2001), in which a high level of amino acid mutations was recorded amongst strains of Equine Rhinitis Virus (ERAV). The mutations were found be located within the connection loops, as well as the N terminus and it was suggested that these mutations confer host specific neutralisation sites within the protein. Furthermore it was found that neutralisation sites were located within the N and C termini as well as the βE-βF and βG-βH loops, with particular strong antibody recognition against the

N-terminus. There have been several studies in which the location of B-cell epitopes of different picornaviruses have been analysed.

According to Cameron *et al* (2000), of the five synthetic peptides of TMEV VP1 which were tested, $VP1_{262-276}$ elicited the greatest immune response. This study supported the earlier findings of Yausch *et al* (1995), in which major epitope regions were located within the VP1, VP2 and VP3 capsid proteins, with three epitopes belonging to VP1 and corresponding to the amino acid sequences $VP1_{12-25}$, $VP1_{146-160}$ and $VP1_{262-276.}$ Several studies into the location of neutralisation sites within proteins of the Foot-and-Mouth-Disease Virus have also supported the notation that VP1 is the most antigenic protein. According to Bittle *et al* (1982), there are two major neutralisation sites within FMDV VP1. The sites were identified to lie between the amino acid residues 131-160 and 193-204, and X-ray crystallography confirmed that both these residues were accessible at the surface of the virus. Collen (2001) reported that the VP1 capsid protein is the antigenic determinant of FMDV, with major neutralisation sites residing within the βG-βH loop between the amino acid residues 134-158. It has also been found the mutation rate of the amino acid sequence within FMDV VP1 is the highest of all viral proteins (Knowles *et al*., 2003). Van Phan *et al* (2010) reported that antigenic variation amongst serotypes is a mechanism by which FMDV evades host immunity. Thus since the mutation rate is highest amongst VP1 proteins, the findings support the immune-dominance of the protein. Furthermore the immune-dominance of VP1 can be supported by two studies, in which the immunization with VP1 induced protection against FMDV (Bachrach *et al*., 1975; Kaaden *et al*., 1977) and Enterovirus 71 (Wu *et al*., 2001).

1.6 Phylogenetic classification of picornaviruses

Members of the *Picornaviridae* family were originally classified into genera based on their serological relatedness and the physiochemical properties of viral proteins (Cooper *et al*., 1978). However since the development of sequencing techniques and the expansion in genomic sequence information, the viruses have been reclassified according to phylogenetic relationships as derived from the 3D RNA-dependent-RNA-polymerase and structural proteins. Rodrigo and Dopazo (1995), initiated this classification through an analysis of the phylogenetic relationships recovered by individual sequences of VP1, VP2, VP3 and 3D proteins from a wide range of viruses belonging to the *Enterovirus, Cardiovirus* and *Aphthovirus* genera. However, overtime virus classification has become predominantly based on the phylogeny of the RNA-polymerase, with minimal consideration of the individual structural proteins. Furthermore the rate of identification and sequencing of novel viruses has continued to increase substantially over the past two decades. Thus there is a necessity to continuously assess the phylogenetic relationships and re-classify viruses as required. Hughes (2004) identified two major clusters within the family. The *Parechovirus* genus was identified as the basal family containing two major sub-groups: 1) viruses of the *Enterovirus* and 2) viruses belonging to the *Teschovirus, Cardiovirus, Erbovirus Aphthovirus and Kobuvirus* genera. Furthermore it was found that all genera were monophyletic. These phylogenetic relationships were recovered independently from the 3D proteins and the viral polypeptides. There are several supporting studies which also found each genus to be monophyletic (Hales *et al*., 2008; Johansson *et al*., 2002; Kapoor *et al*., 2008; King *et al*., 2000). The phylogenetic relationships were further investigated with respect to the non-structural proteins 2C, 3C and 3D (Lewis-Rogers and Crandall, 2009). It was found that the tree topology recovered from the 3D polymerase was consistent with topologies of previous studies. However topologies recovered from the 2C and 3C proteins were inconsistent to each other as well as to the 3D topologies. This was particular observed with regard to the genera: *Teschovirus, Cardiovirus* and *Erbovirus*. The phylogeny of the 3D proteins also disputed the theory of host-pathogen co-phylogeny. A more recent phylogenetic analysis based on the viral 3D polymerase was reported by Phelps *et al* (2013). The analysis was performed in aid of the classification of a novel picornavirus isolated from batfish species. The study reported the clustering of the cardioviruses and cosaviruses, with closest relation to a cluster comprising of the aphthoviruses and erboviruses. The human enteroviruses, simian enteroviruses and the sapeloviruses formed a single clade, while the parechoviruses clustered with the unclassified fish picornaviruses as an out-group

As the classification of picornaviruses is predominantly based on the phylogeny of the RNA-polymerase, there is limited literature with respect to the individual phylogeny of picornavirus structural proteins. The majority of research in this regard has based on the precursor protein P1 or the highly antigenic VP1 protein. The evolution of picornaviruses is characterized by a high mutation rate (10^{-3} to 10^{-5} mutations per nucleotide) (Lewis-Rogers and Crandall, 2009). It has been suggested that this due to genetic drift resultant from the error-prone RNA polymerase, positive selection at VP1 immunogenic sites and genetic recombination between serotypes. As defined by Hu (2014), inter-typic recombination is the process of intramolecular genetic exchange between viral species which have co-infected the same cell. These conclusions of picornavirus evolution were derived from phylogenetic studies of individual genera and viruses within the family. The principle indicator of genetic recombination was reported to be the presence of distinct monophyletic groups, specifically between viral serotypes. This theory of genetic recombination was reported by Smura *et al* (2014) and Simmonds and Welch (2006). Moreover, Lukashev *et al* (2014), reported distinct changes within the phylogenetic relationships across the genomes of HEV-A serotypes. The study focused on the phylogenetic reconstruction of the viruses with respect to the VP1, 2C and 3D proteins. It was suggested that the discrepancies observed across the respective topologies was indicative of inter-typic recombination. This study supported the findings of Heath *et al* (2006), which involved an analysis of the P1 precursor proteins of FMDV serotypes. The analysis indicated the horizontal flow of sequences, with genomic regions encoding structural functionality being interchangeable amongst serotypes (Heath *et al*., 2006). The study also identified 86 possible recombinants sites out of 125 genome sequences. Evidence of inter-typic recombination amongst Enterovirus B strains was also reported by Hu et al (2014). It was also found that the phylogeny recovered from the non-structural proteins was inconsistent to that of VP1, and thus suggested that the structural proteins are evolving independently of the non-structural proteins. A more comprehensive study of the phylogeny with respect to the structural proteins was performed by Boros *et al* (2013). The study reported inconsistency with the phylogeny reconstructed according to the 3D RNA-Polymerase in previous studies. Thus it was suggested that picornavirus structural proteins may be evolving independently from the replications proteins. The study specifically investigated the phylogeny across all genera of the *Picornaviridae* family with respect to the P1 precursor protein. Boros *et al* (2013) indicated common ancestral heritage between the enteroviruses and sapeloviruses. Moreover the sapeloviruses were also found to cluster with the unclassified pigeon picornavirus. Boros *et al* (2013) also observed the close relationship

between the aphthoviruses and erboviruses, as well as the paraphyletic lineage of these viruses with the cosaviruses and cardioviruses. Previous studies with respect to the phylogeny of the structural proteins have been predominately focused on the classification of simian and human enteroviruses. The classification of simian picornaviruses was proposed based on the phylogenetic relationships between VP1 proteins (Oberste *et al.,* 2005). Furthermore the evolutionary genetics for Human Enterovirus-71 (HEV-71) was assessed through phylogenetic analysis of 628 protein VP1 sequences. It was estimated that the common ancestor emerged in 1941 and subsequently diverged into three genotypes: A, B and C. It was suggested that this evolution was resultant of selective pressure at VP1 immunogenic sites. A more recent study by Daleno *et al* (2013) comprehensively investigated the phylogeny of the VP2/VP4 precursor in RV-A, RV-B and RV-C. It was reported that the topologies indicated the distinct clustering of the viral isoforms, with RV-C distinctly clustering from RV-B. Other phylogenetic studies with respect to the structural proteins included the classification of: Human Enterovirus B (Lindberg *et al*., 2003) and human respiratory picornaviruses (Piralla *et al*., 2011).

In the recent classification of novel viruses, the phylogeny with respect to the 3D RNA polymerase and precursor protein P1 has been assessed. Thus there has not been sufficient research into the phylogenetic evolution of the individual structural proteins of recently identified viruses. In 2014 there were 16 new species identified, sequenced and classified (ICTV, 2014). The viruses include a novel strain of Rosavirus A (Lim *et al*., 2014), the Genet Fecal Theilovirus which was assigned to the *Cardiovirus* genus (Bodewes *et al*., 2014) and several chicken, duck and bird picornaviruses. It has been suggested that the assembly of the viral capsid of involves protein-protein interactions between these structural subunits. Thus it is likely that the proteins are co-evolving to retain specific sites of interaction. The next section describes the general process of virus assembly.

1.7 Viral Assembly

The assembly of picornaviruses initiates with the cleavage of the capsid precursor P1 from the P2-P3 domains. This is facilitated by the 2A viral protease. Although viral proteins play a significant role in replication and assembly, it must be noted that the viruses contain a limited genome and thus may also dependent on a range of host cellular proteins which mediate viral entry, replication and assembly. Recent studies have supported the speculation that picornaviruses utilise molecular chaperones during viral replication and assembly. The possible interaction of Hsp90 with the precursor protein P1 from members of the enterovirus genus was reported by Geller *et al*. (2012). It was suggested that Hsp90 recognises and binds P1, such that the precursor can establish correct conformation to allow for cleavage by the 3C protease (Geller *et al*., 2012). This conclusion was based on the findings that the inhibition of Hsp90 ATPase resulted in the inability of P1 to fold into the correct conformation required for proteolytic cleavage. A schematic diagram of the interaction of picornavirus P1 and Hsp90 is depicted in Figure 1.4. Subsequent cleavage of P1 is then mediated by the 3C and 3CD proteases. The sites of this cleavage are located between the VP0-VP3 and VP3-VP1 domains of P1, resulting in the formation of three structural proteins: VP1, VP3 and a precursor VP0. The interaction of a single copy of each of these three proteins results in the formation a 5s protomer. Through the suggested hydrophobic interactions between individual protomers, a 14s pentamer consisting of five protomers is then formed. Research has proposed that the N-terminal of the VP0 subunit (later processed into VP4), facilitates the protomer-protomer interactions. The proposed interaction of 12 pentamers results in the formation of a provirion capsid. It has been projected that the formation of the provirion capsid is facilitated by the interaction of VPg with the inner surface of individual pentamers. The formation of the provirion capsid is believed to be followed by the encapsidation of the RNA genome, which is proceeded by the cleavage of VP0 into VP2 and VP4 to form a mature virus particle (Racaniello, 2007). As the mechanism of viral assembly is still subject to speculation the explicit subunit-subunit interactions responsible for protomer formation have not been elucidated. Moreover the suspected role of host cellular proteins has not yet been explicated. Research has identified limited amino acids in the protein subunits which appear to have a direct effect on capsid assembly or RNA encapsidation. Couderc *et al* (1996) showed that adaptions of PV capsid proteins, specifically the VP1 T22I and VP2 S32T mutations, directly affected capsid assembly. Similarly, Kirkegaard (1990) reported that the deletion of residues 1 to 4 and 8 and 9 in PV VP1 directly affect RNA encapsidation.

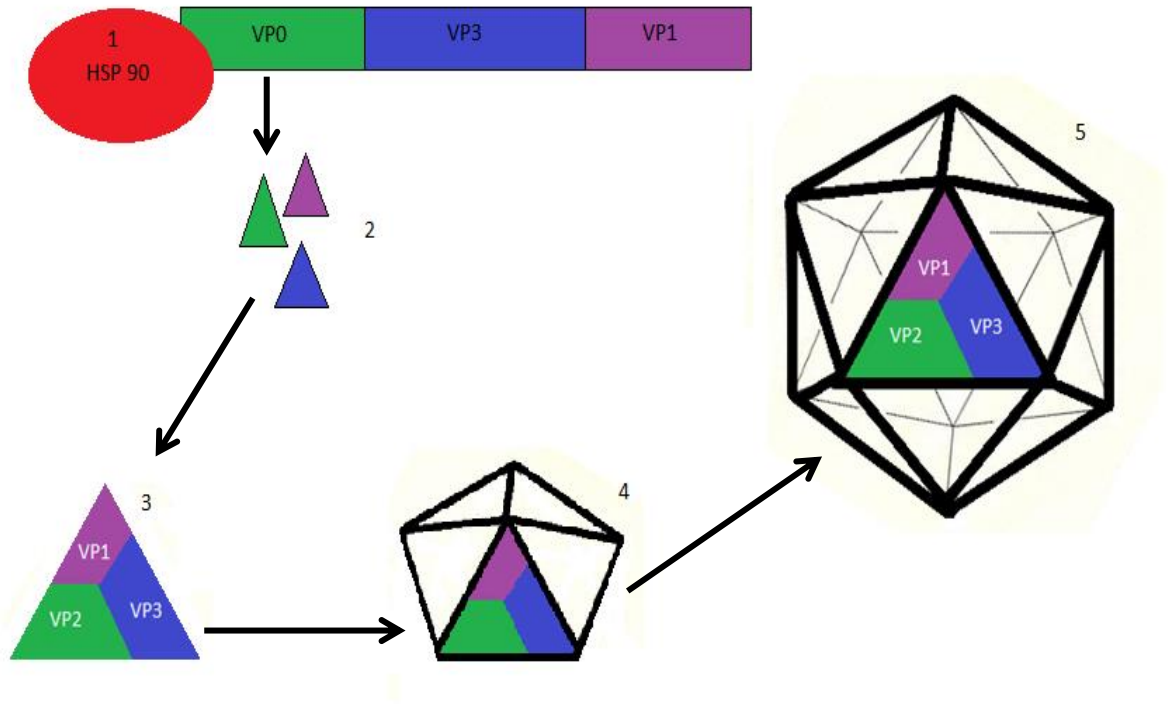


Figure 1.4. Assembly of the Picornavirus capsid. 1) Hsp90 interacts with the capsid precursor P1 to facilitate correct conformation of proteolytic cleavage. 2) Proteolytic cleavage of P1 by 3C/3CD protease. 3) Formation of viral protomer. 4) Formation of 14s pentamer 5) Assembly of provirion capsid and cleavage of the precursor VP0 to form mature virus particle (adapted from Geller *et al.*, 2012)

1.8 Project Motivation

The *Picornaviridae* family contains numerous viruses of high economic and clinical importance. Furthermore the viruses of this family have a broad range of individual hosts, thus as a whole this family of viruses has an effect on both human health and the agricultural industry. Recently the emergence of novel picornaviruses associated with gastrointestinal, neurological and respiratory diseases in humans has been reported. Although effective vaccines for viruses such as FMDV, PV and HAV have been developed there are currently no antivirals available for the treatment of picornavirus infections. Moreover the RNA genome of these viruses is susceptible to a high mutation rate as imposed by the error prone RNA polymerase. Consequently the antigenic regions of the structural proteins can effectively evolve to evade immune defence mechanisms, exponentially decreasing the effectiveness of the vaccine. This often results in the requirement of booster vaccination which may be

economically and clinically challenging. Picornavirus assembly is a complex process which has not yet been elucidated. It has been proposed that the formation of protomer and pentamer intermediate structures is fundamental to capsid assembly. However the mechanisms of protein interactions which may constitute this assembly are not fully understood. The explicit interactions between the viral subunits: VP4, VP2, VP3 and VP1, which result in protomer formation, have not been elucidated. The interactions with host cellular proteins, which may assist in protomer assembly, have also not yet been explicated. Moreover it is unknown if the mechanism of assembly is virus specific or conserved across species of the viral family. An *in silico* approach allowed for a rapid and comprehensive analysis of a large collection of *Picornaviridae* genomic sequence data, which pertained to the individual structural protein subunits. Thus the predictions of this study may provide a broadened understanding of the subunit-subunit interactions within the protomers of picornaviruses, as well as the evolutionary mechanisms of these proteins. As the structural proteins may also be reliant on the interaction with conserved host cellular proteins during assembly, viral protein regions responsible for such interactions and capsid assembly may be also be conserved across the *Picornaviridae* family. Capsid assembly is integral to the life cycle of picornaviruses. Thus this study may also assist in the identification of conserved interacting motifs or residues which could serve as possible drug targets. The development of antivirals may offer novel approaches to the treatment of picornavirus infections and thus provide a more cost effective alternative to the ephemeral vaccinations.

1.9 Aims and objectives

The overall aim of the study was to broaden the understanding of the evolution and function of the structural proteins across the *Picornaviridae* family. The study had three principle objectives. Firstly a comprehensive analysis of the phylogenetic relationships amongst the individual structural proteins was performed. The aim was to identify evolutionary patterns across sub-types of individual picornaviruses as well as determine co-host phylogenetic relationships. The study also aimed to identify correlations and discrepancies in the phylogeny of the independent structural proteins. Secondly the function of the structural proteins was further investigated by an exhaustive motif analysis performed using Multiple-EM for Motif Elicitation (MEME), a sequence analysis tool developed by Bailey and Elkan (1994). The analysis aimed to determine the conservation of motifs across the viral family, with specific identification of conserved short linear motifs (SLiMs) which may facilitate protein subunit-subunit interactions within the protomer of picornavirus capsids. Motif

conservation was assessed across the individual structural proteins of: 1) strains of individual virus types and 2) different viruses which infect the same host species. Thirdly the study aimed to predict specific subunit motif-subunit motif interactions, with identification of the principle interacting residues and the corresponding types of interactions. The study also aimed to calculate the conservation of these residues across the strains of the respective viruses. The specific objectives included an *in silico* prediction of interacting residues within representative PDB files, the mapping of predicted residues to corresponding motifs and the *in silico* analysis of interacting motifs within the subunit-subunit interface of representative crystal structures.

2. Phylogenetic analysis of *Picornaviridae* structural proteins

The *Picornaviridae* family currently consists of 26 genera, with a total of 46 sub-classified viral species (ICTV, 2014). The current classification is based on the phylogenetic relationships as derived from the 3D RNA-dependent-RNA-polymerase, a highly conserved viral replication protein (Rodrigo and Dopazo, 1995). Although the host specificity and viral pathogenicity of picornaviruses vary greatly across members of the family, the icosahedral capsid is highly conserved for all picornaviruses (Racaniello, 2007). The capsid consists of 60 protomers, each containing a single copy of structural proteins: VP1, VP2 and VP3. A fourth capsid protein, VP4, is found on the internal side of the capsid (Racaniello, 2007). In this chapter, an analysis of the phylogenetic relationships amongst the individual structural proteins is performed. The study aimed to identify any evolutionary patterns of the structural proteins of individual picornavirus species as well as determine relationships between different viruses with the same host species. Additionally, the study aimed to identify correlations and discrepancies between phylogenies across the genome. Thus phylogenetic analysis was performed in sequential order of the structural proteins corresponding to location in the viral genome. Specifically the datasets were analysed in the order of VP4, VP2, VP3 and VP1.

2.1 Introduction

Phylogenetics is the study of evolutionary relationships amongst a genetically related group of organisms (Bast 2013). These relationships, as derived from analysis of molecular sequencing data, can be statically inferred as a phylogenetic tree with characteristic branch topology (Bast 2013). This statistical inference is solely based on probability models, which represent assumptions of either the nucleotide substitution or amino acid replacement process. As these models have been derived from databases of nucleotide/amino acid sequences, they represent already seen evolutionary patterns. Thus the discipline of phylogenetics is limited to assumptions of models which may not represent the evolution of novel datasets. Over the last 30 years a collection of models with increasing complexity and accuracy regarding nucleotide/amino acid substitution have been described, however each is inclined to produce different results given the same dataset. Thus the choice amongst models

is a critical step of phylogenetic reconstruction (Cunningham *et al*., 1998). The convolution of phylogenetic reconstruction is further increased by the variety of approaches used to determine actual tree topology, given a dataset and particular evolutionary model. The next section describes the advantages and disadvantages of these different approaches with specific reference to maximum-likelihood, the approach used in this study.

2.1.1 Approaches of phylogenetic reconstruction

The most common methods of phylogenetic reconstruction include distance based methods, parsimony approaches and the maximum-likelihood approach. Distance based methods construct tree topologies which account for the evolutionary distances (expressed as the number of substitutions per site) between pairwise sequences. This approach employs a distance matrix as derived from an evolutionary model. Although this approach has the lowest computational expense, it is less precise and only appropriate for sequences with recognizable similarity. Contrary to this, maximum parsimony (MP) is a character based method which implicitly assumes a model of evolution which requires the minimum number of substitutions to explain relationships within a dataset. This method is most applicable for sequences with high similarity (Farris, 1973; Felsenstein, 1973; Yang, 1998; Steel and Penny, 2000). The third approach, maximum-likelihood is also dependent on an explicit evolutionary model. However, unlike distance methods which only account for a single parameter (substitutions per site), the maximum-likelihood (ML) approach accounts for all phylogenetic parameters (substitutions per site, tree topology, branch length, among-site rate variation, base frequency and the presence of invariant sites) (Felsenstein, 1981). Likelihood is defined as a quantity which is proportional to the probability of observing the data, given a specific model. Thus for an evolutionary model, the probability that the observations in the data would actually have been observed can be calculated as a function of that model. Through examination of this function it is possible to determine the parameters responsible for the greatest probability of observing the evolutionary pattern of the model within the given data set (Cho, 2012). More specifically, in this approach the nucleotide/amino acid bases of all sequences at each site are independently considered and the log-likelihood of a given topology for each set of individual bases is computed. This is followed by the summation of likelihoods at all given sites, which is further maximised to estimate the branch length of the tree. This procedure is repeated for all possible topologies with the resultant tree showing the highest product of site likelihood. It must be noted that the topology likelihood calculations are dependent on the evolutionary model and thus the use of the correct model is emphasized

(Cho, 2012). This is one of the major shortcomings of the ML approach, which is further substantiated by the limitations to derive information from sites under parsimony and thus information regarding these sites are purely consequential of the model used (Cho, 2012). However in comparison to distance and MP methods, ML is advantageous as it converges to the true tree with an increase in the size of the data set. Felsentein (1981), displayed that MP results are inconsistent with respect to data set size and in cases where unequal evolutionary rates are present within the sequences. In addition ML phylogenies have been found to be more consistent in the grouping of short sequences. The ML approach is statistically well understood and substantiated. Unlike other methods, it allows for the evaluation of all topologies and branch lengths, thus increasing the prospect of a correctly reconstructed phylogeny (Cho, 2012). Furthermore, both distance and MP methods are only considerate of pairwise relationships and are incapable of considering evolutionary relationships within multiple sequence alignments (MSA). Therefore substantial evolutionary information is unaccounted for (Le and Gascuel, 2008). This is not the case in ML approaches, thus offering another advantage over the distance and MP based approaches. Although the development of ML based approaches has improved the accuracy of phylogenetic reconstruction, the reality of evolutionary processes is multi-faceted and the accuracy of probability calculations is limited to the availability of current evolutionary models. The next section describes the limitations and advancements in the development of evolutionary models.

2.1.2 Evolutionary models used in phylogenetic reconstruction

The basic definition of an evolutionary model is a model which represents nucleotide/amino acid substitution and may incorporate a set of assumptions with regard to properties of the given dataset. These properties include: 1) the unequal or equal frequency of bases/residues within the sequences (+F), 2) the proportion of invariable sites within the dataset (+I), 3) the heterogeneity of among-site evolutionary rates which is accounted for with gamma distribution (+G) and 4) the unequal or equal distribution of nucleotide/amino acid substitution rates (Posada and Crandall, 2001). With regard to protein sequences, the residue substitution conformation is directly inferred from the incorporation of amino acid replacement matrices, which specifically represent the instantaneous rate of substitution from one residue to another, as derived by analysis of datasets of known protein sequences.

The earlier replacement matrices (Dayhoff *et al*., 1972) where derived from the consideration of only closely related sequence pairs (>85% identity). Essentially the phylogeny between available sequences was inferred by MP and the average number of amino acid changes per

pair closely related pairs were counted. The Dayhoff (1978) and Jones, Taylor, and Thornton (1992) (JTT) also incorporated a similar counting approach to determine their replacement matrices. However the JTT replacement matrix is derived from a much larger dataset of protein sequences. The replacement matrices derived from these counting approaches do not consider MSA, nor do they incorporate evolutionary information for sequences with less than 85% identity. Thus they are significantly limited.

The first attempts to derive replacement matrices from MSA involved a ML approach (Adachi and Hasegawa, 1996; Yang *et al*., 1998; Adachi *et al*., 2000). However due to large computational expense the datasets used were small and limited to protein sequences from 20, 23 and 10 species respectively. The major advancement in the development of replacement matrices derived from MSA was achieved by Whelan and Goldman (2001). Their development of the WAG matrix involved the much larger BRKALN database of 182 alignments and 900000 residues. Initially phylogeny was inferred by neighbour-joining (NJ) method, followed by the estimation of branch lengths by ML. The optimal replacement matrix was then derived by ML, based on the inferred phylogeny. The WAG matrix showed clear improvements over both the JTT and Dayhoff matrices, resulting in higher likelihood values for inferred phylogenies. This replacement matrix was further advance by Le and Gascuel (2008), in the development of the LG replacement matrix.

The LG matrix is considered a general matrix, as it is derived from protein sequences from three kingdoms of life. Thus this matrix does not represent residue replacement of a specific protein family or domain of life and is therefore more robust and performs well for many different collections of sequences. The matrix is derived from a larger and more diverse dataset than WAG, specifically the Pfam database (2002). It is well known that sites of a given protein do not evolve at a constant rate. Functional or structural constraints limit the rate of certain sites, while non-structural sites, often integrated in turns, evolve faster due to low evolutionary pressure. The major advancement of LG over WAG, includes the consideration of this among-site rate variation and invariant sites in the likelihood calculations and replacement rate estimations (Le and Gascuel, 2008).

In contrast to the general LG model, Dimmic *et* al (2002) had developed the reverse transcriptase (rtREV) amino acid substitution matrix. Optimized as a matrix for ML phylogenetic analysis on the dataset of 33 amino acid sequences from the retroviral POL proteins, the matrix is specifically applicable to the phylogenetic analysis of the rapidly

mutating RNA retroviral POL proteins. The replication of RNA viruses is distinctively associated with high mutation rates, short generation times and large numbers of progeny. Furthermore, horizontal and vertical gene transfer as well as systemic infection result in frequent population bottlenecks which encourage the development of a local population for founding genetic drift (Amos and Harwood, 1998). This, together with the continuous colonization of new host populations, allows for an increase in selection effects and a more complex environment unique to RNA virus proteins. The rtREV matrix was developed from amino acid sequences from the lentiviruses, spumaviruses, betaretroviruses and the gammaretroviruses. Although applicable to the POL proteins, the matrix was found to be incompatible with phylogenetic reconstruction of the retroviral GAG proteins, thus indicating the high specificity of matrices required for the phylogenetic analysis of RNA viral proteins (Dimmic *et al*., 2002).

2.1.3 A description of the MEGA v6.0 approach

MEGA is a multifaceted program, which allows for phylogenetic reconstruction through either distance, MP or ML approaches. Given a multiple sequence alignment, MEGA initially estimates the goodness-of-fit of different substitution models with and without the assumption of the existence of discrete Gamma distribution. With respect to protein sequence, this results in the evaluation of 48 amino acid substitution models. Furthermore, for each of these models, MEGA calculates the values of Gamma distribution, the proportion of invariant sites and the different substitution rates between residues within the dataset. Dependent on the type of model, the observed or assumed amino acid frequency values are also calculated. The goodness-of-fit is measured by the Bayesian information criterion (BIC) and Akaike information criterion (AIC), which is further substantiated by the log likelihood (*lnL*) of each model. MEGA also offers three options with regard to the treatment of ambiguous regions within the alignment: 1) complete deletion in which all sites containing gaps are ignored. This option is preferred because different regions amino acid sequences may evolve under different evolutionary stimuli and thus ambiguous sites are best ignored. 2) Partial deletion, in which the threshold of ambiguous regions can be stipulated. 3) Pairwise deletion, where all ambiguous sites are initially considered but removed as required during pairwise distance calculations. This option is most appropriate when gaps are randomly dispersed amongst the alignment as only those gaps involved in the pairwise comparison are removed (Tamura *et al*., 2013). Followed by model selection, phylogenetic trees constructed by the ML method involve the construction of an initial tree using a fast but suboptimal

method such as Neighbor-Joining, the branch lengths of which are adjusted to maximize the likelihood of the dataset for the given topology. Variants of this topology are then created according to the nearest neighbour interchange method (NNI), in the search for topologies which improve the fit of the data. Consequently the branch lengths of these topologies are optimized to determine the greatest likelihood. This search is exhausted until no greater likelihood can be found. The final phylogenetic tree is evaluated by bootstrap analysis and consensus tree construction.

2.2. Methods and materials

2.2.1. Sequence retrieval and dataset

Translated protein sequences corresponding to annotated coding sequences (CDSs) of all available viral genomes of the *Picornaviridae* family were downloaded, in Protein FASTA format, from the Virus Pathogen Database and Analysis Resource (ViPR). All duplicate genome sequences were excluded from the query, yielding protein sets from 2185 individual viruses. Scripting (Appendix 1.1) was used to extract and group individual structural protein sequences corresponding to VP1, VP2, VP3 and VP4. Only sequences with standardised annotations were included. Table 2.1 presents the total number of sequences per structural protein group. Python scripting, incorporating iterative pairwise alignments and percentage identity calculations, was used to subsequently filter each protein group (Appendix 1.2). The resultant sizes of the respective datasets were too large for feasible phylogenetic reconstruction, thus similar sequences (> 80% identity) were removed from each dataset. The final size of each dataset is also presented in Table 2.1.

Table 2.1. Respective sizes of each picornavirus structural protein dataset for phylogenetic analysis. Protein sequences corresponding to 2185 individual picornaviruses were downloaded from ViPR. The structural proteins were individual extracted and grouped, with all redundant sequences with greater than 80% identity removed.

Structural Protein Dataset	Total Number of Sequences Extracted	Total Number of Sequences After Filtration
VP1	1965	209
VP2	1884	80
VP3	1965	129
VP4	1804	53

Following data filtration, all sequences were uniformly renamed in the form of Host|Virus|Strain by scripting (Appendix 1.3). The abbreviations and corresponding full labels are presented in Table 2.2, while detailed text files of proteins sequences for each structural group can be found in FASTA format in Appendix 2.

Table 2.2. Abbreviations of sequence headers. Abbreviations and corresponding full-names of respective viruses and hosts. The abbreviations were used to uniformly rename all sequence headers for phylogenetic and motif analysis.

Viruses		**Hosts**	
Representative abbreviation used in this study	**Virus as labelled in ViPR**	**Representative abbreviation used in this study**	**Host as labelled in ViPR**
AV	Aichivirus	**Ts**	Tortoise
CoSV	Cosavirus	**Pi**	Pigeon
EMV	Encephalomyelitisvirus	**Sl**	Seal
EMCV	Encephalomyocarditisvirus	**M**	Mouse
EV	Enterovirus	**C**	Canine
RAV	Equine Rhinitis A Virus	**Cp,**	Caprine
RBV	Equine Rhinitis B Virus	**Tc**	Tick
FMiPV	Fatheadminnow Picornavirus	**Al**	Alpaca
FMDV	Foot-and-Mouth-Disease -Virus	**E**	Equine
HAV	Hepatitis A Virus	**Cz**	Chimpanzee
HPeV	Human Parechovirus	**O**	Ovine
HuV	Hunnivirus	**Tg**	Tiger
IaioPiV	Ia io picornavirus	**H**	Human
AV	Kobuvirus	**P**	Porcine
LV	Ljunganvirus	**F**	Feline
MiniPiV	Miniopterusschreibersii Picornavirus	**Ck**	Chicken
OHUV	Ovine Hungarovirus	**Th**	Thrush
PaV	Pasivirus	**Mi**	Minnow
PiV	Picornavirus	**Bf**	Buffalo
RfV	Rafivirus	**S**	Simian
RV	Rhinovirus	**Mk**	Monkey
SV	Sapelovirus	**B**	Bat
SiV	Sicinivirus	**Br**	Boar
TeschV	Teschovirus	**A**	Avian
TMEV	Theiler's Murine Encephalomyelitis Virus	**R**	Rat
ThV	Theilovirus	**Bo**	Bovine
TV	Turdivirus	**Tk**	Turkey
		U	Unknown Host

2.2.2. Multiple Sequence Alignments

Multiple sequence alignments (MSAs) were performed individually for all structural protein datasets. This was facilitated on a broad scale through Bio-Python and MUSCLE (Edgar, 2004). Structural alignments were also performed using PROMALS3D (Pei *et al*., 2008). The crystal structures were obtained from the Protein Databank (PDB) for each respective dataset and are shown in Table 2.3. Default settings were used for all MSAs. Upon comparison of the alignment of conserved protein regions, the PROMALS3D alignments had greater accuracy and thus were chosen for phylogenetic analysis. The alignments were further edited in Jalview v2.7 (Waterhouse et al. 2009). This allowed for the removal of large ambiguous regions prior to phylogenetic analysis as well as the manual alignment of regions which were inadequately aligned by PROMALS3D. Due to the experimental limitation of crystallization processes missing residues are not uncommon in PDB files. Therefore the sequences corresponding to those structures used by PROMALS3D were also removed prior to phylogenetic analysis. The resultant PROMALS3D alignments for each respective protein dataset can be found in Appendices 3.1-3.4. It must be noted that the viral proteins are products of proteolytic cleavage of a single polyprotein, thus methionine is not always the first residue of the sequence.

Table 2.3. Crystal structures used for PROMALS3D alignments of picornavirus capsid proteins

Capsid Protein	PDB of crystal structure with chain identifier
VP1	1TME_1; 1AYM_1; 1HXS_1; 1QQP_1; 3VBH_A
VP2	1TME_2, 1AYM_2; 1POV_0; 1QQP_2; 4G3B_0
VP3	1TME_3, 1AYM_3; 1QQP_3; 2MEV_3; 3VBH_C
VP4	1C8M_4; 1POV_0, 1TME_4; 4CDQ_D,4GMP_0

2.2.3. Phylogenetic Analysis

Phylogenetic reconstruction was performed individually for each of the structural protein datasets. In each case MEGA v6.06 (Tamura *et al*., 2013) software was used. Initial evolutionary model tests were performed for each dataset at complete deletion as well as partial deletion with site coverage cut-off set at 95% and 90% respectively. The best three evolutionary models, for each dataset, were subsequently selected based on BIC scores (Table 2.4; Table 2.5; Table 2.6; Table 2.7). For each dataset, evolutionary history was inferred by using the ML approach based on all three models at 100%, 95% and 90% site coverage cut-offs respectively. NNI was chosen as the ML heuristic method with strong

branch swap filtration. The phylogeny was tested by bootstrap method with the number of replicates set to 1000. Due to the high degree of variation within the VP1 sequences, additional phylogenies of this dataset were also reconstructed using the NJ method with pairwise deletion. Evolutionary distances were computed using the JTT (Jones *et al*., 1992) and Dayhoff (Dayhoff, 1978) matrix-based methods respectively.

2.3. Results and Discussion

2.3.1. Phylogenetic analysis of Picornaviridae VP4 capsid subunit proteins

2.3.1.1. Evolutionary model selection

Phylogenetic analysis involved a total of 53 amino acid sequences which were representative of the VP4 capsid protein from viruses across the *Picornaviridae* family. The MSA was performed using PROMALS3D and incorporated the crystal structures in Table 2.3. According to the BIC scores (Table 2.4) MEGA 6.06 predicted the best three evolutionary models as LG+G, LG+G+I, and rtREV+G, respectively for all positions with less than 100%, 95% and 90% coverage eliminated.

Table 2.4. The best-fit evolutionary models for phylogenetic reconstruction of *Picornaviridae* VP4 amino acid sequences. The model tests were performed using MEGA v6.06 software, at site coverage cut-offs of 100%, 95% and 90%. The dataset contained 53 sequences from across the *Picornaviridae* family.

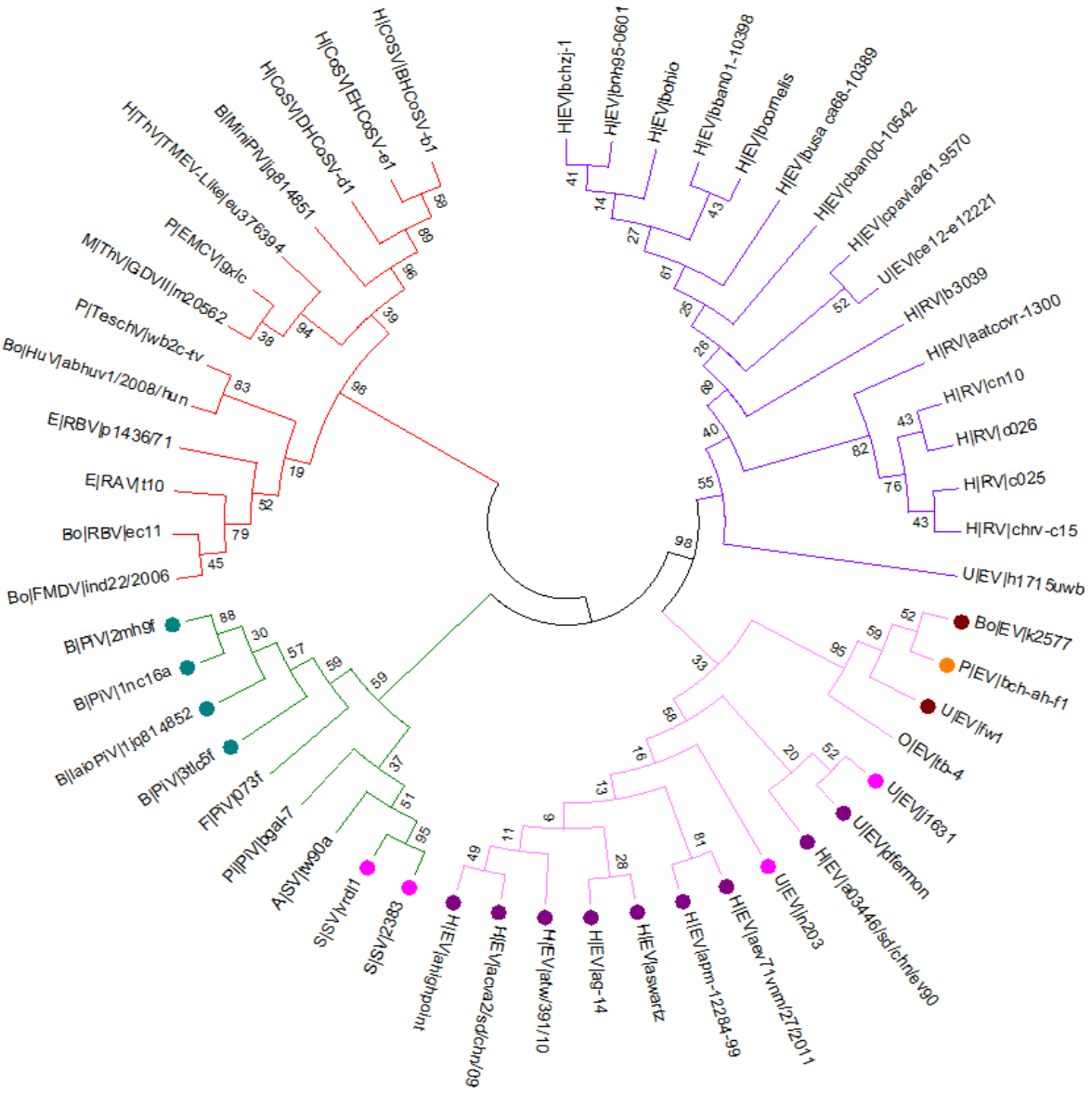
Site Coverage Cut-off (%)	Model	Model Reference	BIC Score	AIC Score	*lnL*
100	LG+G	Le and Gascuel, 2008	5241.077	4695.484	-2217.744
	LG+G+I	Le and Gascuel, 2008	5242.593	4655.632	-2214.702
	rtREV+G	Dimmic *et al*., 2002	5258.226	4676.632	-2226.318
95	LG+G	Le and Gascuel, 2008	6995.925	6379.893	-3079.544
	LG+G+I	Le and Gascuel, 2008	6997.947	6376.191	-3090.027
	rtREV+G	Dimmic *et al*., 2002	7016.891	6400.859	-3087.995
90	LG+G	Le and Gascuel, 2008	8508.185	7872.680	-3826.650
	LG+G+I	Le and Gascuel, 2008	8508.818	7867.389	-3847.587
	rtREV+G	Dimmic *et al*., 2002	8550.058	7911.359	-3844.919

The results of model prediction indicated favourable phylogenetic reconstruction at complete deletion (100% cut-off), as indicated by the incessant increase in BIC and AIC scores at 95%

and 90% cut-off respectively. Furthermore a decrease in the *lnL* values with incessant decrease of cut-off threshold also supports phylogenetic reconstruction at the complete deletion level.

2.3.1.2. Phylogenetic reconstruction

Phylogenetic analysis is a complex process, and branch topology is subject to change with the use of different evolutionary models. Thus for precision and comparability analysis, phylogenetic reconstruction was performed independently for all models at the complete and partial deletion levels of 95% and 90%. Bootstrap analysis and consensus tree construction, revealed that the LG+G+I at a 90% cut-off threshold performed the best. The parameters for gamma distribution and invariant sites, as calculated by MEGA v6.06, were 2.7336 and 2.8922% respectively. Although the BIC, AIC and *lnL* scores strongly supported complete deletion, VP4 sequences are considerably short (approx., 60 amino acids). Therefore the exclusion of a significant proportion of sites at 100% threshold could have been detrimental to phylogenetic reconstruction. The topology inferred at 90% by the LG+G+I model (Figure 2.1), indicates the existence of three major clusters. Firstly a statistically significant out-group (bootstrap of 98) containing the viruses from the genera: *Aphthovirus*, *erbovirus*, *teschovirus*, *hunnivirus*, *cosavirus* as well as the unclassified bat picornavirus MiniPiV JQ-814851 (Sequence header: B|MiniPiV|jq814851). And secondly, a main cluster (super-group) consisting of two major sub-clusters: I) a group containing the sapeloviruses with the pigeon, bat and feline picornaviruses (bootstrap of 59). II) A cluster consisting of the enteroviruses (bootstrap of 98). Although the corresponding bootstrap values of topologies inferred at different levels of deletion using models: LG+G and rtREV+G, were significantly lower (Appendices 4.1.1- 4.1.9), the overall clustering and out-grouping was consistent with that in Figure 2.1. There is limited literature with respect to the individual phylogeny of picornavirus VP4 proteins, with most research based on the precursor protein P1 or the highly antigenic VP1 protein. However Daleno *et al* (2013) comprehensively investigated the phylogeny of the VP2/VP4 precursor in RV-A, RV-B and RV-C. The reported topologies indicated the distinct clustering of the viral isoforms, with RV-C distinctly clustering from RV-B. Thus the distinguished grouping of RV-B from RV-C in this study is congruent with Daleno *et al* (2013). However no comparison can be made for the clustering of RV-B with EV-C and EV-B protein sequences. Furthermore Boros *et al* (2013) specifically investigated the phylogeny across all genera of the *Picornaviridae* family with respect to the P1 precursor protein, results of which directly correlate to the clustering patterns found in this study.



.

Figure 2.1. Phylogenetic tree topology of 53 amino acid sequences corresponding to the *Picornaviridae* capsid VP4 protein. Evolutionary history was inferred by MEGA v6.06 software using the ML method based on the LG model. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 2.7336). The model also allowed for invariant sites (+I, 2.8922%). The reconstruction was performed with partial deletion and all sites with less than 90% coverage were eliminated. Significant clusters are colour coded: Red) Out-group. Green) Cluster I of the super-group. Purple) and Pink) indicate the respective monophyletic sub-groupings of Cluster II of the super-group. Specific host species are indicated by coloured bullets: Purple) Human. Pink) Simian. Turquoise) Bat. Maroon) Bovine. Orange) Porcine.

Boros *et al* (2013) indicated common ancestral heritage between the enteroviruses and sapeloviruses. Moreover the sapeloviruses were also found to cluster with the unclassified pigeon picornavirus. Boros *et al* (2013) also observed the close relationship between the

aphthoviruses and erboviruses, as well as the paraphyletic lineage of these viruses with the cosaviruses and cardioviruses. In contrast to capsid phylogeny analysis, Phelps *et al* (2013) performed a recent analysis of the 3D RNA polymerase proteins. The resultant tree topology was reported to indicate congruency of the overall sub-groupings of genera. However significant differences were observed within the clade containing enteroviruses and sapeloviruses. This disagreement of the phylogenetic relationships between the structural and non-structural proteins has also been observed in previous studies (Lukashev *et al*., 2014; Heath *et al*., 2006; Boros *et al*., 2013). Distinct changes within the phylogenetic relationships across a genome are indicative of inter-typic recombination. The presence of distinct monophyletic groups observed in this study may indicate recombination between closely related species, specifically between viral serotypes. This is supportive of the studies reported by (Smura *et al*., 2014; Simmonds and Welch, 2006). The unique clustering of VP4 sequences from 3D sequences, within the monophyletic sub-groupings does however indicate different recombination patterns within different picornavirus proteins. The overall clustering of VP4 sequences in this study also disputes host-pathogen co-phylogeny. This result is supported by the phylogeny of the 3D proteins inferred by Rogers and Crandall (2009).

2.3.1.2.1. Analysis of the evolutionary relationships within the out-group of VP4 sequences

The inferred topology indicated the presence of two separate monophyletic clusters within the out-group (Figure 2.1: Red). It must be noted that protein sequences from viruses of the same genus (as previously classified by phylogenies derived from the 3D RNA polymerase) clustered together. The first cluster contained aphthoviruses: FMDV, BRBV and ERAV, the erbovirus ERBV as well as the representative teschovirus and hunnivirus sequences. Teschovirus and hunnivirus appeared to be monophyletic (bootstrap of 83), with distinct lineage from the aphthoviruses. Although bootstrap values of this splitting were low (19 of 1000 replicates), this exact lineage was consistent throughout topologies inferred by all models (Appendices 4.1.1- 4.1.9). The second monophyletic cluster contained the cardioviruses: TMEV GDVII, EMCV and Human TMEV-like cardiovirus. Once again this was distinctly split from the cosaviruses which clustered with MiniPiV JQ-814851 (bootstrap of 96). Furthermore a distinct monophyletic lineage was observed between Cosavirus E and B which were found to be paraphyletic to Cosavirus D (bootstrap of 89).

2.3.1.2.2. Analysis of the evolutionary relationships within super-group Cluster I

Analysis of Cluster I (Figure 2.1: Green), indicated the distinct lineage between the sapeloviruses, which clustered with the unclassified pigeon picornavirus BGAL-7 and a

second sub-group containing the unclassified bat and feline picornaviruses (bootstrap of 59). These results indicate a common ancestor for the bat and feline picornaviruses. However a distinct paraphyletic lineage was observed between the bat picornaviruses and the feline picornavirus. Moreover a monophyletic lineage was observed between simian SV strains (bootstrap of 95), with paraphyletic lineage to avian SV strain. This supports the hypothesized that strains from the same host would cluster into monophyletic groups. It must also be noted that unclassified pigeon picornaviruses have been reported to be closely related to the sapeloviruses and enteroviruses, thus results of this study correlate directly to previous studies (Lau *et al*., 2011).

2.3.1.2.3. Analysis of the evolutionary relationships within super-group Cluster II

The second cluster observed within the super-group consisted solely of sequences from viruses which belong to the *Enterovirus* genus. These observations support previous literature which reported the close relatedness between the enteroviruses and sapeloviruses (Lau *et al*., 2011). The presence of two major sub-groups was observed in this cluster. However the exact grouping was found to be model dependent. Topologies inferred using the LG+G+I model, showed the distinct clustering of the EV-B, EV-C, EV-H and RV-B with RV-A and RV-C (Figure 2.1: Purple) from the clustering of Porcine EV and EV-F with EV-A, EV-J and EV-D (Figure 2.1: Pink). This was contradictory to the groupings inferred by the LG+G and rtREV+G. Topologies inferred by both these models included the Porcine EV and EV-F sequences with the EV-B, EV-C and rhinoviruses, with a distinct sub-group containing EV-A, EV-J and EV-D sequences. Although these models both grouped the ungulate enteroviruses with EV-B, EV-C and the rhinoviruses, the bootstrap values with respect to this grouping were very low in both models (bootstrap values of 4-34). While the bootstrap values produced by LG+G+I with respect to the clustering of these viruses with EV-A, EV-F and EV-J ranged between (bootstrap values of 91-98) with sub-clustering bootstrap values >80 .

According to analysis of the topology inferred by the LG+G+I model, a distinct clustering of EV-B and EV-C sequences was observed. Surprisingly the representative RV-B sequence clustered with these EV-C and EV-B sequences (bootstrap of 69), and was distinctly separate from the sub-grouping of the RV-A and RV-C (bootstrap of 82). This unique clustering of RV-B was observed in topologies inferred by all models. Monophyletic relationships were observed in RV-C strains, with paraphyletic clustering to RV-A. EV-H was shown to be the basal sequence of this sub-group, most distantly related to all other viruses within the cluster. The second sub-group within this cluster indicated the distinct grouping of enteroviruses

which infect ungulate species from the human and simian enteroviruses: EV-A, EV-J and EV-D sequences (bootstrap of 95). However as evident in the clustering of the bovine and porcine enteroviruses as well as the grouping of simian EV-J with human EV-D (Figure 2.1), no monophyletic clustering was observed between strains which infect the same host species. Although monophyletic clustering between the human EV-A strains was observed, this is more likely due to the distinction of the EV-A serotypes, from the other enteroviruses. As VP4 serves as an internal capsid protein, it is more protected from selection pressure to evade host immune defences or to retain host cell specificity.

2.3.2. Phylogenetic analysis of Picornaviridae VP2 capsid subunit proteins

2.3.2.1. Evolutionary model selection

The phylogenetic reconstruction of the VP2 capsid proteins included 80 amino acid sequences. A structural alignment of the sequences was performed in PROMALS3D based on the crystal structures depicted in Table 2.3. The alignment was subsequently edited in Jalview to removed sequences corresponding to the respective structures. Large ambiguous regions were also removed prior to model selection (Appendix 3.2). According to the BIC scores (Table 2.5) MEGA 6.06 predicted the best three evolutionary models as LG+G, LG+G+I, and LG+G+F, respectively for all positions with less than 100%, 95% and 90% coverage eliminated.

Table 2.5. The best-fit evolutionary models for phylogenetic reconstruction of *Picornaviridae* VP2 amino acid sequences. The model tests were performed using MEGA v6.06 software, at site coverage cut-offs of 100%, 95% and 90%. The MSA contained 80 sequences from across the *Picornaviridae* family.

Site Coverage Cut-off (%)	Model	Model Reference	BIC Score	AIC Score	*lnL*
100	LG+G	Le and Gascuel, 2008	28348.925	27152.055	-13416.302
	LG+G+I	Le and Gascuel, 2008	28355.990	27151.567	-13415.036
	LG+G+F	Le and Gascuel, 2008	28464.595	27124.266	-13382.966
95	LG+G	Le and Gascuel, 2008	37105.367	35879.731	-17780.425
	LG+G+I	Le and Gascuel, 2008	37113.651	35880.277	-17779.679
	LG+G+F	Le and Gascuel, 2008	37843.880	35843.880	-17743.131
90	LG+G	Le and Gascuel, 2008	38661.986	37433.038	-18557.108
	LG+G+I	Le and Gascuel, 2008	38671.179	37434.471	-18556.806
	LG+G+F	Le and Gascuel, 2008	38767.527	37391.175	-18516.816

Model calculations favoured phylogenetic reconstruction at the complete deletion level. This is evident by the considerable lower BIC and AIC scores at 100% cut-off threshold. Furthermore the *lnL* score is also significantly higher at the complete deletion level. However regions of the alignment (Appendix 3.2) which contained ambiguous gaps correlated to varying regions. Any conservation observed within these varying regions paralleled to sequences of individual viral strains. Although the exclusion of these regions would simplify the dataset, the phylogeny derived from these regions may be fundamental to deriving the relatedness between these viral strains. According to the BIC scores, the LG+G model was calculated to best-fit the data at all levels of deletion of these regions. However this was not consistently supported by the AIC and *lnL* scores. Thus for sensitivity and thoroughness, phylogenetic reconstruction was performed respectively for each of the three models at each level of deletion.

2.3.2.2. Phylogenetic reconstruction

As previously stated, phylogeny was reconstructed by MEGA v6.06 using the ML approach at complete and partial deletion levels of 95% and 90%, with respect to the models LG+G, L+G+I and LG+G+F. Bootstrap analysis of the resultant topologies revealed that the LG+G at 90% deletion performed the best and is presented in Figure 2.2. This was particularly

evident in the improvement in bootstrap values at the sub-branching level as a result of using the LG+G model at 90% deletion (Appendix 4.2. and Figure 2.2). The VP2 and VP4 capsid proteins are consequent of the cleavage of precursor protein VP0. This cleavage is a late stage in viral capsid assembly and results in the formation of a mature virion (Racaniello, 2007). Thus congruency between these respective phylogenies was hypothesized. This section critically analyses the phylogeny inferred by the VP2 sequences in comparison to that of the VP4 sequences. Examination of this topology revealed significant branching of an out-group (bootstrap of 97), containing sequences corresponding to the genera *Cardiovirus, Cosavirus, Aphthovirus, Erbovirus, Tremovirus, Hepatovirus, Hunnivirus* and *Teschovirus.* This grouping was congruent with that of the VP4 amino acid sequences. However this particular analysis included both a larger number of sequences corresponding to additional virus strains, as well as sequences corresponding to Hepatitis A Virus (HAV) from the *Hepatovirus* genus and Avian Encephalomyelitis Virus (EMV) from the *Tremovirus* genus. The corresponding amino acid sequences were removed from the VP4 dataset due to their significant shortness of 21aa (HAV) and 20aa (EMV), in comparison to the average length (60aa) of the MSA. The VP2 topology also revealed a super-group (bootstrap of 98) comprising of three major clusters. Cluster III contained sequences corresponding to RV-C, RV-B, RV-A, EV-C and EV-B. Cluster II was composed of sequences corresponding to the EV-A, EV-J, EV-H serotypes as well as the porcine and bovine enteroviruses. This basal internal clustering is congruent with that of VP4. Cluster I was comprised of the representative sequences of the unclassified bat and feline picornaviruses (bootstraps 84, 93 and 96). In contrast to VP4 and phylogeny inferred for the 3D polymerase (Phelps *et al.*, 2013), the corresponding sapelovirus sequences formed an internal out-group, paraphyletic to the super-group and with closest relation to the bat and feline picornaviruses. Although the bootstrap value (bootstrap of 50) for this variation in topology was not highly significant, this internal out-group was also inferred by the LG+G+F and LG+G+I models at the 90% level of deletion. In comparison to VP4, examination of the overall external clustering of VP2 sequences revealed significant variation within topology. A comprehensive comparison and examination of the phylogenies are described in the following sections.

2.3.2.2.1. Analysis of the evolutionary relationships within the out-group of VP2 sequences

The out-group (Figure 2.2: Red), consisted of two monophyletic sub-groups. Although the bootstrap values specific to this sub-clustering were not significant (bootstrap values of16 and 20), the sub-clusters were inferred by all three models (Appendix 4.2.1- 4.2.9). Congruent

with the clustering of VP4, the teschoviruses and hunniviruses clustered together (bootstraps of 90 and 100) and was grouped with the external clustering of the aphthoviruses (bootstrap values of 69 and 99 respectively). This sub-group also contained the monophyletic cluster of HAV and EMV (bootstrap of 100). Furthermore, as in the VP4 analysis, the second monophyletic group was comprised of the cardioviruses (bootstrap values of 92, 69 and 98) and cosaviruses (bootstrapping of 50 and 100). This also supports previous findings of Boros *et al* (2013). As indicated in Figure 2.2, monophyletic clustering was observed between the human cardioviruses, supporting host specificity. There were two distinct variations between the VP4 and VP2 out-groups. While the VP2 sequence corresponding to the erbovirus, Equine Rhinitis B Virus (Sequence header: E|RBV|p1436|71), was found to be paraphyletic to the cosavirus sequences (bootstrap of 49), the corresponding VP4 sequence clustered with the aphthoviruses sequences (bootstrap of 52). Although the bootstrap values were not significantly high, this variation was observed in all three models at all levels of deletion, thus suggesting a notable discrepancy in phylogeny irrespective of the evolutionary model used. A second discrepancy was observed within the external clustering of the cosavirus sequences. The phylogeny of VP4 inferred the monophyletic cluster of CoSV-B and CoSV-E as paraphyletic to CoSV-D, while the phylogeny of VP2 inferred CoSV-E as paraphyletic to the CoSV-B and CoSV-D clade. This particular discrepancy was also observed using the LG+G+I at 90% deletion and was therefore not a consequent of the omitting invariant sites. This result supports genetic recombination between viral serotypes, while the first discrepancy may supports recombination between species from different taxonomic units.

2.3.2.2.2. Analysis of the evolutionary relationships within Cluster II

Cluster II (Figure 2.2: Light-pink) contained the EV-A, EV-J and EV-H serotypes, as well as the bovine and porcine enterovirus serotypes. This internal cluster directly correlated to that of VP4. Furthermore, as in the phylogeny of VP4, this cluster was distinctly segregated from the other virus of the *Enterovirus* genus (bootstrap values of 65 and 57). Congruency with VP4 phylogeny was also observed in the separation of the porcine and bovine enteroviruses (bootstrap of 94) from the EV-A and EV-J serotypes. The EV-A serotypes formed a significant clade (bootstrap of 100), paraphyletic to the simian sequences of the EV-J and EV-H serotypes. The phylogeny inferred for EV-H (Sequence header: U|EV|h1715uwb) was consistant across topologies inferred using all three models and was found to be incongruent with the phylogeny inferred for the corresponding VP4 protein sequence (clustered with the EV-B, EV-C and RV serotypes). This indicates a greater host specificity of the external VP2

capsid protein, compared to that of VP4. Another significant discrepancy (bootstrap of 99) was observed in the external clustering of the bovine and porcine enterovirus sequences. Again discrepancy was observed irrespective of the model used, thus further supporting the notation of inter-typic recombination between viral strains and serotypes of closely related species.

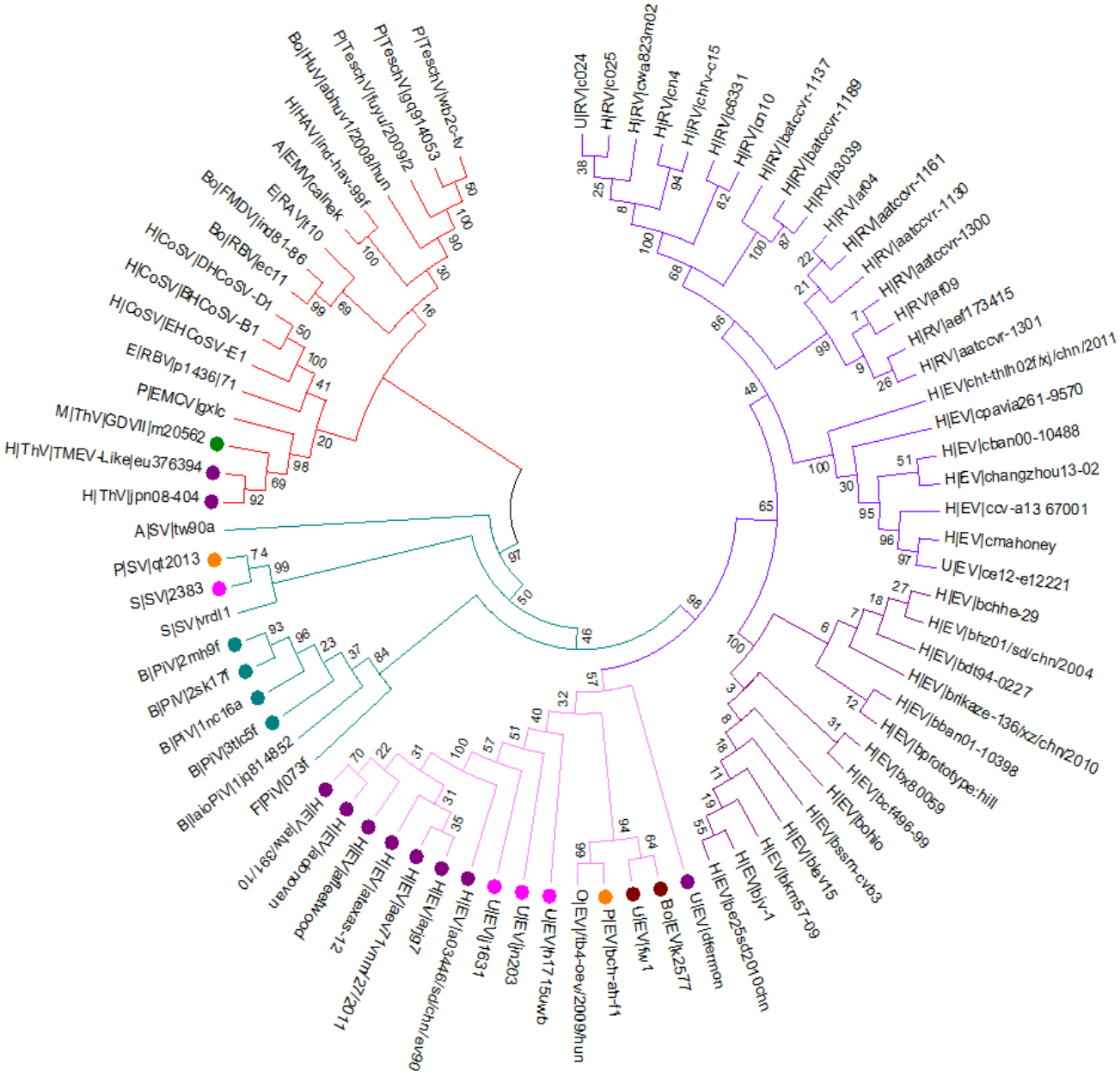


Figure 2.2. Phylogenetic tree topology of 80 amino acid sequences corresponding to the *Picornaviridae* capsid VP2 protein. Evolutionary history was inferred by MEGA v6.06 software using the ML method based on the LG model. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 0.7893). The reconstruction was performed with partial deletion and all sites with less than 90% coverage were eliminated. Significant clusters are colour coded: Red) Out-group. Green) Cluster I of the super-group. Purple) and Pink) indicate the respective monophyletic sub-groupings of Cluster II of the super-group. Specific host species are indicated by coloured bullets: Green) Murine. Purple) Human. Pink) Simian. Turquoise) Bat. Maroon) Bovine. Orange) Porcine.

2.3.2.2.3. Analysis of the evolutionary relationships within Cluster III

Cluster III (Figure 2.2: Purple) was comprised of two major sub-groups: 1) the significant monophyletic cluster of the amino acid sequences corresponding to the EV-B serotypes (bootstrap of 100) and 2) a complex cluster which comprised of the EV-C and RV serotypes.

This topology was incongruent to that inferred by VP4 sequences with respect to two conflicting results. Firstly the corresponding EV-C and EV-B VP4 sequences clustered as a single monophyletic clade, with a paraphyletic relation to the representative RV-B sequence (H|RV|b3039). Secondly the VP4 sequences representative of the RV-A and RV-C serotypes formed a segregated clade from the EV-C and EV-B sequences. In the topology inferred by the VP2 sequences, H|RV|b3039 clustered significantly within the external cluster containing other RV-B sequences (bootstrap of 100), as well as within the internal cluster comprising of RV-C, RV-B and RV-A serotypes (bootstrap of 86). In agreement with results reported by Daleno *et* al (2013), distinct external clustering was observed for the respective RV-A, RV-B and RV-C sequences (respective bootstraps of 99, 100 and 100), with closer relation between RV-B and RV-C sequences (bootstrap of 68). This opposed the closer relation between RV-A and RV-C inferred by the VP4 analysis. It must be noted that this variation was also observed in the LG+G+I model, at all levels of deletion, and was therefore not resultant of the difference in evolutionary models used to infer the topologies depicted in Figures 2.1 and 2.2. The external clustering of viral strains within the major sub-groups was not supported by bootstrap analysis, particularly evident amongst the EV-B sequences (bootstrap values < 10). Thus there is no fixed pattern of vertical speciation and the results could support evolution through horizontal gene transfer, as proposed by Phelps *et al* (2013).

2.3.3. Phylogenetic analysis of **Picornaviridae** *VP3 capsid subunit proteins*

2.3.3.1. Evolutionary model selection

The phylogenetic analysis of the VP3 dataset included a total of 129 amino acid sequences. The sequences were representative of viruses across the *Picornaviridae* family, with all sequences with greater than 80% identity removed. The MSA (Appendix 3.3) was produced by PROMALS3D and subsequently edited in Jalview. Sequences corresponding to the crystal structures (Table 2.3) were removed prior to model calculation, along with large regions of ambiguity. Evolutionary model calculations, by MEGA v6.06, were performed at the complete and partial deletion thresholds of 95% and 90%. The three best-fit models, predicted at all three levels of deletion, were LG+G+I, LG+G and LG+G+F. Analysis of the BIC, AIC and *lnL* scores (Table 2.6) indicated favourable reconstruction under complete deletion (100% threshold). However for sensitivity and thoroughness phylogenetic analysis was performed at all three deletion thresholds respectively for each of the LG+G+I, LG+G, LG+G+F evolutionary models.

Table 2.6. The best-fit evolutionary models for phylogenetic reconstruction of *Picornaviridae* VP3 amino acid sequences. The model tests were performed using MEGA v6.06 software, at site coverage cut-offs of 100%, 95% and 90%. The MSA contained 129 sequences from viruses across the *Picornaviridae* family.

Site Coverage Cut-off (%)	**Model**	**Model Reference**	**BIC Score**	**AIC Score**	***lnL***
100	LG+G+I	Le and Gascuel, 2008	46320.118	44282.438	-21881.026
	LG+G	Le and Gascuel, 2008	46320.998	44291.221	-21886.443
	LG+G+F	Le and Gascuel, 2008	46369.580	44189.665	-21816.175
95	LG+G+I	Le and Gascuel, 2008	59039.794	56949.089	-28214.941
	LG+G	Le and Gascuel, 2008	59041.075	56958.485	-28220.659
	LG+G+F	Le and Gascuel, 2008	59079.902	56843.160	-28143.598
90	LG+G+I	Le and Gascuel, 2008	66316.552	64205.916	-31843.647
	LG+G	Le and Gascuel, 2008	66320.007	64217.565	-31776.522
	LG+G+F	Le and Gascuel, 2008	66366.665	64108.567	-31776.522

2.3.3.2. Phylogenetic Reconstruction

The phylogenetic reconstruction of the VP3 amino acid sequences was more complex than that of VP4 and VP4. The VP2 and VP4 proteins are consequential to the cleavage of the precursor VP0. Hence the annotation of specific viral genomes did not distinguish between the individual VP2 and VP4 CDSs and included only the VP0, VP3 and VP1 CDSs. As a result the extraction of the VP3 and VP1 amino acid sequences included representative sequences from viruses which were not included in the VP4 and VP2 dataset. The additional viruses incorporated were: HPeV, LV, PaV, Sebokele Virus, Seal PiV, FMiPV, SiV, TV, RfV, SaKV, Salivirus and AV. The corresponding full names of these viruses can be found in Table 2.2.

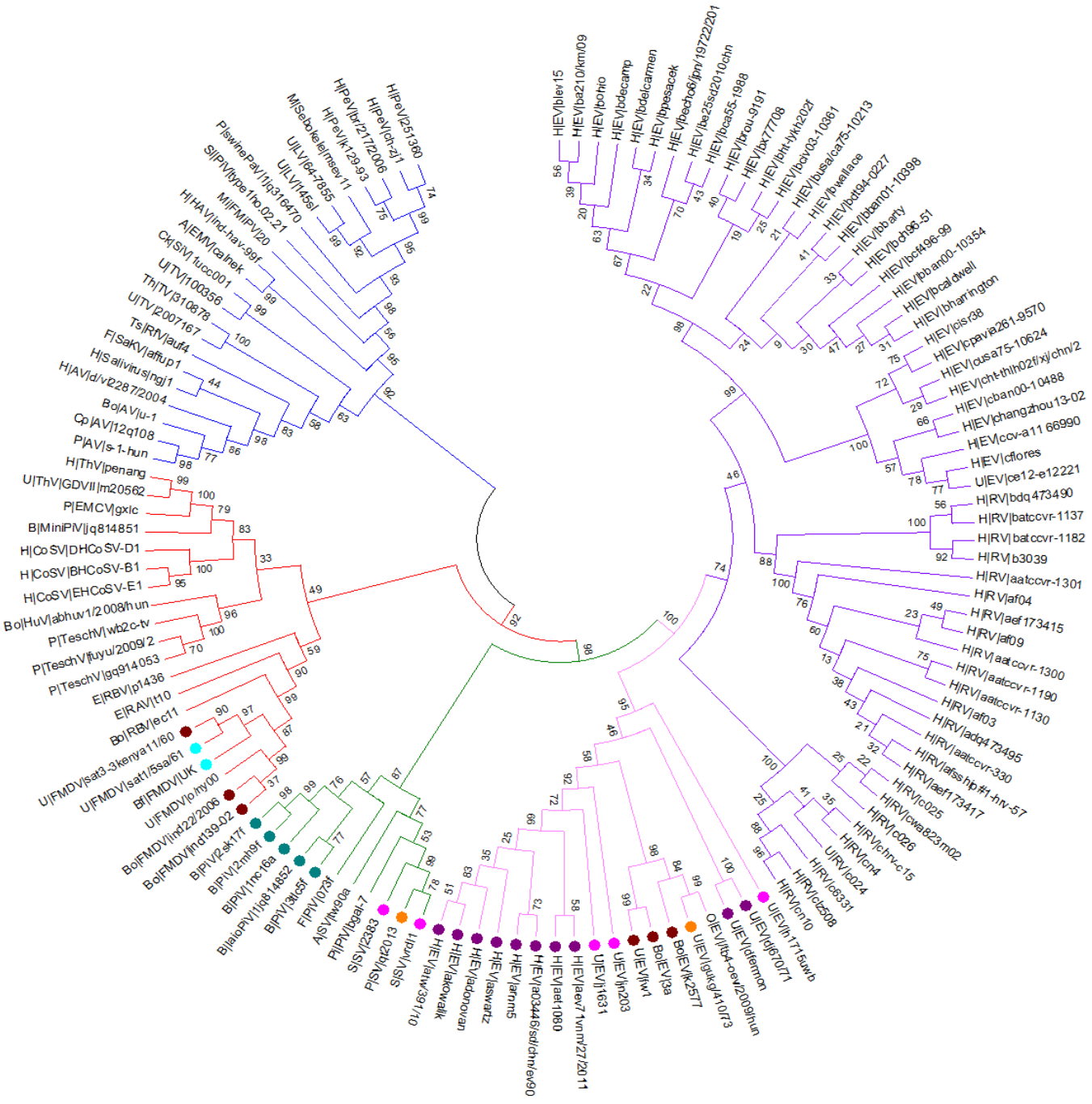


Figure 2.3.1. Phylogenetic tree topology of 129 amino acid sequences corresponding to the *Picornaviridae* capsid VP3 protein. Evolutionary history was inferred by MEGA v6.06 software using the ML method based on the LG model. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 1.8349). The model also allowed for invariant sites (+I, 1.3190%). The reconstruction was performed with partial deletion and all sites with less than 90% coverage were eliminated. Significant clusters are colour coded by branch: Blue) Out-group I. Red) Out-group II. Green) Cluster I. Pink) Cluster II. Purple) Cluster III. Specific host species are indicated by coloured bullets: Purple) Human. Pink) Simian. Turquoise) Bat. Maroon) Bovine. Orange) Porcine. Cyan) Buffalo.

A total of nine phylogenetic trees were inferred by MEGA v 6.06 using the ML approach. Bootstrap analysis and consensus tree investigation revealed that the LG+G+I model was best-fit for the dataset, with highest bootstrap values attained at the 90% deletion threshold. This tree is depicted in Figure 2.3.1, while the remaining eight trees can be found in Appendices 4.3.1- 4.3.9. It was found that the sequences corresponding to the additional viruses listed above formed a significant out-group (bootstrap of 92). Furthermore this clustering was also inferred by the LG+G and LG+G+F models (Appendix 4.3). This out-group, Out-Group I (Figure 2.3.1: Blue) is later discussed in comparison to the phylogeny of the corresponding VP1 sequences (Section 2.3.4). For comparative analysis with the phylogenies of VP2 and VP4, Out-Group I was excluded and the topology presented in Figure 2.3.2 was considered. The clustering of the VP3 amino acid sequences did not correlate to phylogenies represented by the 3D RNA-polymerase (Phelps *et al*., 2013) or the P1 precursor proteins (Boros *et al*., 2013).

2.3.3.3.1. Analysis of the evolutionary relationships within Out-Group II

The out-group depicted in Figure 2.3.2 (Red cluster) was congruent with that of the VP2 and VP4 sequences. The representative sequences of the cosaviruses and cardioviruses externally clustered together (bootstrap of 83), while the teschoviruses and hunnivirus sequence also formed an external cluster (bootstraps of 96, 100 and 70). Congruent with the clustering observed with respect to VP4 proteins, the erbovirus E|RBV|p1436 clustered externally with the aphthovirus sequences. This was incongruent with the phylogeny inferred with respect to the VP2 sequences. This discrepancy was also observed in topologies inferred by the LG+G and LG+G+F models, and was therefore not resultant from inconsistent model use across the structural protein datasets. A second discrepancy was observed between VP3 and VP4 phylogenies with respect to the grouping of the unclassified bat picornavirus MiniPiV-JQ-814851. As both the VP3 and VP4 topologies were inferred using LG+G+I model at 90% deletion, the discrepancy is not consequential of inconsistent model usage. The MiniPiV-JQ-814851 VP3 sequence clustered with the cardiovirus sequences (bootstrap of 79), while the corresponding VP4 sequence was found to be paraphyletic to the cosaviruses (bootstrap of 96). A third discrepancy was between the VP3 and VP4 phylogenies was observed in the monophyletic relationships between the cosavirus sequences. As shown in Figure 2.3.2 the monophyletic CoSV-B and CoSV-E was paraphyletic to CoSV-D (bootstrap of 95 and 100), incongruent with both the VP2 and VP4 topologies (Figures 2.1 and 2.2). This was also observed in the phylogenies inferred by the LG+G and LG+G+F models (Appendix 4.3).

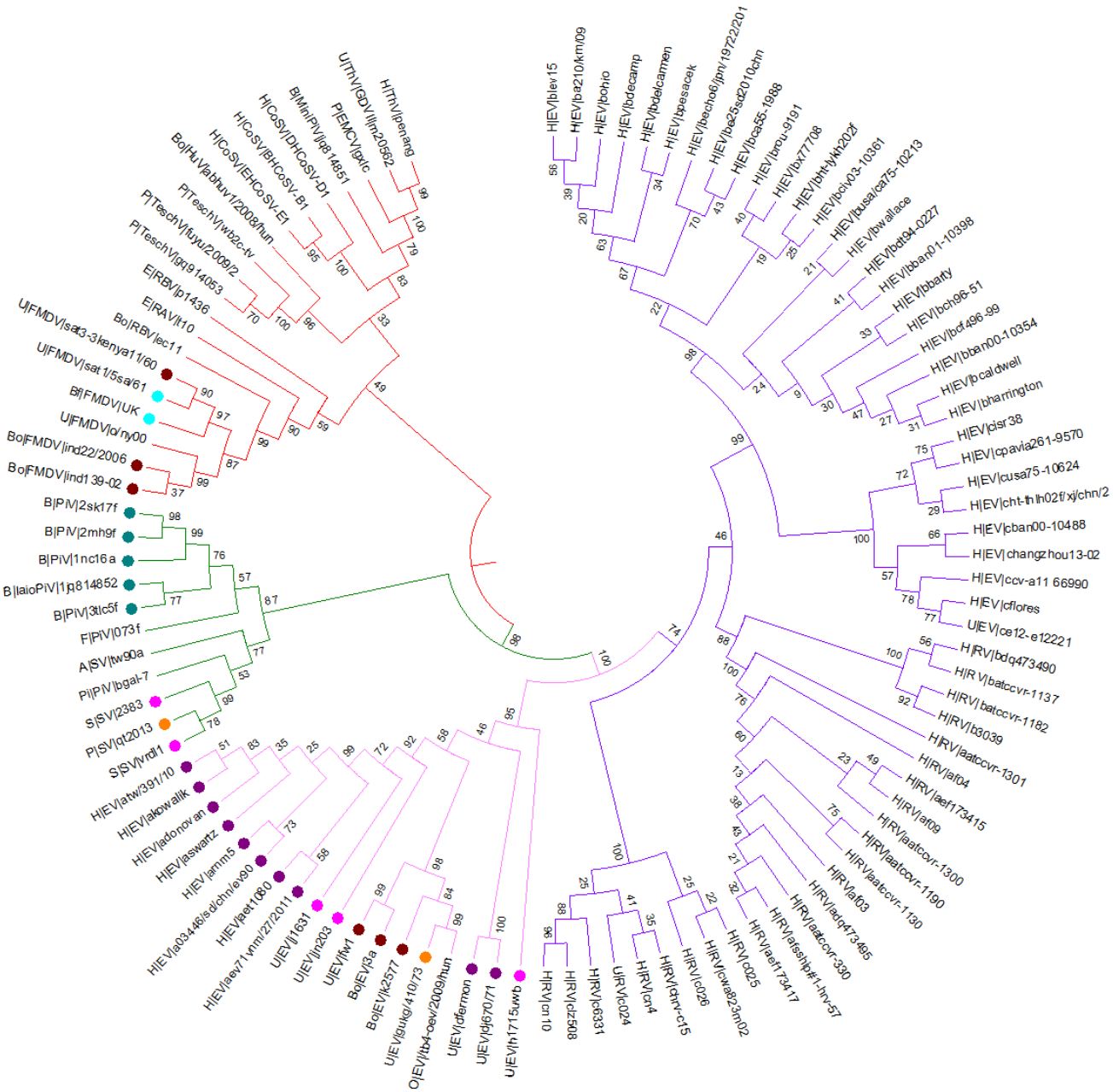


Figure 2.3.2. Phylogenetic sub-tree of the *Picornaviridae* capsid VP3 proteins. The original out-group has been excluded to contain 106 of the original 129 amino acid sequences, Evolutionary history was inferred by MEGA v6.06 software using the ML method based on the LG model. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 1.8349). The model also allowed for invariant sites (+I, 1.3190%). The reconstruction was performed with partial deletion and all sites with less than 90% coverage were eliminated. Significant clusters are colour coded by branch: Red) Out-group I. Green) Cluster I. Pink) Cluster II. Purple) Cluster III. Specific host species are indicated by coloured bullets: Purple) Human. Pink) Simian. Turquoise) Bat. Maroon) Bovine. Orange) Porcine. Cyan) Buffalo.

2.3.3.3.1. Analysis of the evolutionary relationships within Cluster I

In congruency with the phylogenies of VP2 and VP4, the super-group contained a significant cluster (bootstrap of 98), comprising of the bat and feline picornavirus sequences as well as the sapelovirus sequences. Incongruent external clustering with respect to the VP2 sequences was observed among the sapelovirus sequences (bootstraps of 78 and 99). This observation was also supported by topologies inferred by the LG+G model (Appendix 4.3).

2.3.3.3.1. Analysis of the evolutionary relationships within Cluster II

It was observed that the EV-A, EV-J, EV-H and the bovine and porcine enterovirus serotypes formed a distinct clade (bootstrap of 95). This was congruent with that of VP2 and VP4 phylogenies. Furthermore as observed in the VP2 analysis the EV-A serotypes formed a separate external clade (bootstrap of 99), with closest relation to the EV-J serotypes (bootstrap of 92). In agreement with the phylogenies of both VP2 and VP4, the respective VP3 sequences of the bovine and porcine enteroviruses also formed a distinct external clade (bootstrap of 98). Discrepancies between VP4 and VP2 were observed regarding the clustering of EV-H (U|EV|h1715uwb), with significant bootstrapping of 95 and additional support from topologies inferred by the LG+G and LG+G+F models (Appendix4.3). However this clustering correlated to that of VP1, as discussed in Section 2.3.4.

2.3.3.3.1. Analysis of the evolutionary relationships within Cluster III

As observed in the phylogenies of VP2 and VP4 sequences, the EV-B, EV-C and RV sequences formed a large complex cluster (bootstrap of 74). However comparison of the external sub-clustering was incongruent with both that inferred by VP2 and VP4 sequences respectively. In disagreement with the phylogeny of VP2 sequences, VP3 sequences corresponding to the EV-B and EV-C serotypes formed a distinct sub-cluster (bootstrap of 99). Within this sub-cluster, the EV-B and EV-C sequences formed respective monophyletic clades (bootstraps of 99 and 100 respectively). Another discrepancy with respect to both the phylogenies of VP2 and VP4 was observed with regard to the sub-grouping of the RV sequences. The representative sequences of the RV-B and RV-A serotypes formed a distinct cluster from the sequences of the RV-C serotypes (bootstraps of 88 and 100 respectively). Furthermore this topology was inferred throughout phylogenetic analyses of the VP3 sequences and is therefore not a result of model bias.

2.3.4. Phylogenetic analysis of **Picornaviridae** *VP1 capsid subunit proteins*

2.3.4.1. Evolutionary model selection

The analysis of phylogenetic relationships amongst the *Picornaviridae* VP1 capsid, involved a total of 209 amino acid sequences. Structural alignment of the sequences was facilitated by PROMALS3D (Appendix 3.4) and incorporated the crystal structures shown in Table 2.3. The MSA contained significant ambiguity, with greater sequence variation in comparison to the other capsid proteins. Therefore, along with topologies inferred by the ML approach, the phylogeny was substantiated using a distance based approach with pairwise deletion. For that reason large ambiguous regions were not excluded from the alignment prior to phylogenetic reconstruction. Model calculations, as performed by MEGA v6.06, indicated that by ML at complete deletion LG+G, LG+G+I and WAG+G were best-fit to the data. While at the 95% and 90% deletion threshold LG+G, LG+G+I and LG+G+F were calculated as best-fit. The models were primary selected according to the BIC scores, however for completeness and sensitivity phylogenies were inferred according to all models, at respective deletion thresholds, as shown in Table 2.7. The respective phylogenies inferred by the NJ distance method were based on the JTT and Dayhoff models.

Table 2.7. The best-fit evolutionary models for phylogenetic reconstruction of *Picornaviridae* VP1 amino acid sequences. The model tests were performed using MEGA v6.06 software, at site coverage cut-offs of 100%, 95% and 90%. The MSA contained 209 sequences from viruses across the *Picornaviridae* family.

Site Coverage Cut-off (%)	**Model**	**Model Reference**	**BIC Score**	**AIC Score**	***lnL***
100	LG+G	Le and Gascuel, 2008	30364.055	27151.791	-13137.210
	LG+G+I	Le and Gascuel, 2008	30369.401	27149.658	-13135.083
	WAG+G	Whelan and Goldman, 2001	30546.461	27334.198	-13228.413
95	LG+G	Le and Gascuel, 2008	75439.270	70963.204	-35050.108
	LG+G+I	Le and Gascuel, 2008	74544.624	70960.190	-35002.642
	LG+G+F	Le and Gascuel, 2008	74647.798	70960.190	-35000.169
90	LG+G	Le and Gascuel, 2008	97142.556	93463.282	-46301.217
	LG+G+I	Le and Gascuel, 2008	97148.666	93460.776	-46299.043
	LG+G+F	Le and Gascuel, 2008	97181.332	93338.365	-46219.462

2.3.4.2. Phylogenetic Reconstruction

As in VP3 analysis, the VP1 phylogenetic reconstruction contained sequences corresponding to viruses which were not included in the respective VP2 and VP4 analysis. Furthermore, due to greater variation within VP1 sequences, the filtered dataset was considerably larger than that of the other capsid proteins. The increase in variability of the VP1 sequences was expected, as this protein is considered to be the antigenic determinant as well as the main protagonist in binding of the host-cell receptor. Thus VP1 proteins are the least conserved across the virus family as they have evolved to be highly specific to the host cell (Racaniello *et al*., 2007). The analysis of the VP1 phylogeny therefore also focused on host specificity and viral pathogenicity which are discussed in the following sections.

Although model calculations favoured ML reconstruction at complete and partial deletion of 95%, bootstrap analysis of the respective topologies inferred at these levels indicated significantly low values, including bootstrap values of 0 (Appendix 4.4). Therefore these topologies have been excluded from the results analysis. Although topologies inferred at 90% deletion threshold obtained an increase in bootstrap analysis, significantly low values of less than 20 were still observed. The LG+G+I model (90% deletion), appeared to obtain the highest bootstrap values with strongest correlation to its bootstrap consensus tree. This topology, as depicted in Figure 3.4.1, was selected as the sole candidate for results analysis. Both the trees inferred by the NJ method, as based on the JTT and Dayhoff models respectively, revealed significantly lower bootstrap values of less than 10 with respect to the out-grouping clusters. Thus these topologies were also excluded from the analysis. Bootstrap consensus is an indicator of the degree to which the data fits the model of evolution. The low bootstrap values observed in this study, indicate strong disagreement with the LG model, irrespective of gamma distribution or the inclusion of invariant sites. Furthermore the JTT and Dayhoff models were found to be inapplicable and no improvement was facilitated by pairwise deletion. This iterates the complexity of phylogenetic reconstruction, and suggests that the picornavirus proteins, specifically the VP1 capsid proteins, have a distinct pattern of evolution. The replication of RNA viruses is associated with high mutation rates, short generation times and large numbers of progeny. This, together with the continuous colonization of new host populations, allows for an increase in selection effects and a more complex environment unique to RNA virus proteins. Furthermore frequent population bottlenecks, characteristic to RNA viral infections, encourage the development of quasispecies populations which found genetic drift (Amos and Harwood, 1998). As the VP1

capsid protein is the antigenic determinant, it faces greater selection pressure to mutate and evade host-cell defences while maintaining strong specificity to the host cell receptor. The strong disagreement with evolutionary models, in comparison to the other capsid proteins, suggests the presence of distinctive evolutionary pressures imposed on this protein. The next section discusses the respective correlations and distinctions observed between the VP1 proteins and the other capsid proteins. As the topology contains a degree of low bootstrap values, the tree serves only as an indicator of evolutionary relationships.

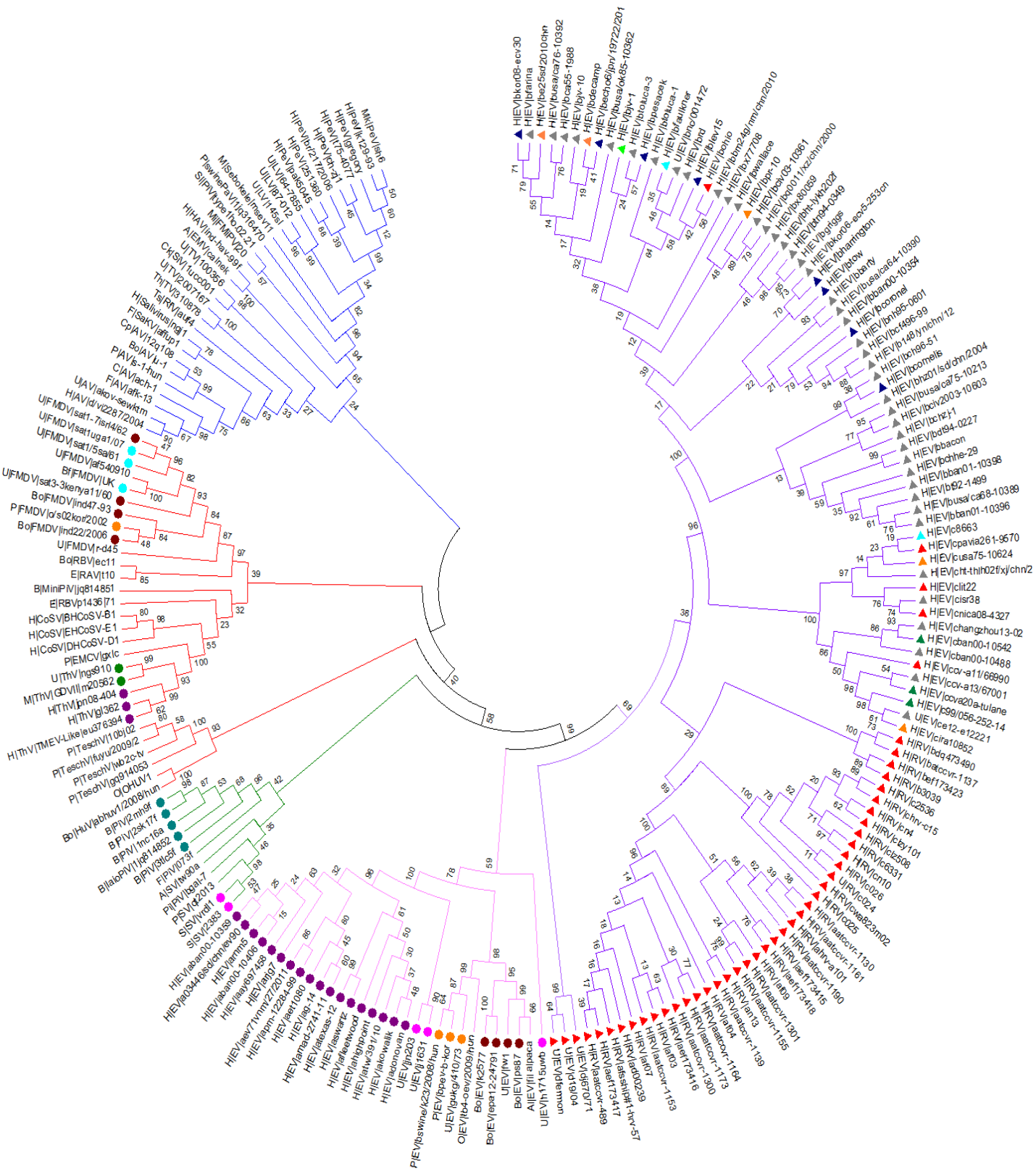

H|EV|bkor08-ecv30
H|EV|bfarina
H|EV|be25sd2010chn
H|EV|busa/ca76-10392
H|EV|busa/ca55-1988
H|EV|bjv-10
H|EV|bdecamp
H|EV|bechol6/jpn/19722/201
H|EV|busa/ok85-10362
H|EV|bjv-1
H|EV|btoluca-3
H|EV|bpesacek
H|EV|bbtoluca-1
U|EV|bfalkner
H|EV|bhc/001472
H|EV|bjrd
H|EV|bjev15
H|EV|bchnio
H|EV|bbm24gl/nm/chn/2010
H|EV|bjx77708
H|EV|bwallace
H|EV|bpP-10
H|EV|bbq0011xz/chn/2000
H|EV|bjx80059
H|EV|bht-lykh202f
H|EV|bdm94-0349
H|EV|bgrtggs
H|EV|bko06-ecv5-253cn
H|EV|bharrington
H|EV|bstow
H|EV|bbusa/ca64-10390
H|EV|bbanb00-10354
H|EV|bconnal
H|EV|bhl95-0601
H|EV|bcf496-99
H|EV|b148/ynl/chn/12
H|EV|bch96-51
H|EV|bcomelis
H|EV|bhz01/sd/chn/2004
H|EV|busa/ca75-10213
H|EV|busa/ca68-10389
H|EV|bcbv2003-10603
H|EV|bchzj-1
H|EV|bdt94-0227
H|EV|bbacon
H|EV|bchhe-29
H|EV|bbban01-10398
H|EV|b92-1499
H|EV|bban01-10396
H|EV|c8663
H|EV|cpavia261-9570
H|EV|cusa75-10624
H|EV|cht-thlh02f/xj/chn/2
H|EV|clit22
H|EV|cisr38
H|EV|cnica08-4327
H|EV|changzhou13-02
H|EV|cban00-10542
H|EV|cban00-10488
H|EV|ccv-a11/66990
H|EV|ccv-a13/67001
H|EV|ccva20a-tulane
H|EV|cb99/056-252-14
H|EV|cle12-e12221
H|EV|cira10852
H|RV|bdq473490
H|RV|batccvr-1137
H|RV|bef173423
H|RV|b3039
H|RV|c2536
H|RV|chrv-c15
H|RV|cn4
H|RV|cty101
H|RV|c6331
H|RV|cle508
H|RV|cn10
H|RV|c026
H|RV|c024
H|RV|cwa923m02
H|RV|c025
H|RV|aatccvr-1130
H|RV|aatccvr-1161
H|RV|alhrv-a101
H|RV|aatccvr-1190
H|RV|aef173416
H|RV|a69
H|RV|aatccvr-1301
H|RV|aatccvr-1155
H|RV|aatccvr-1139
H|RV|a104
H|RV|aatccvr-1164
H|RV|aatccvr-1173
H|RV|aatccvr-1300
H|RV|an13
H|RV|aef173416
H|RV|an3
H|RV|aatccvr-1153
H|RV|a407
H|RV|ad00239
H|RV|aef173417
H|RV|assshpf1-hrv-57
H|RV|aatccvr-489
U|EV|dj670/71
U|EV|dj619/04
U|EV|dhemon
U|EV|h171sw6
A|EV|fil alpaca
Bo|EV|ps87
U|EV|fw1
Bo|EV|epa12-24791
Bo|EV|K2577
O|EV|tb4-oev/2009/hun
U|EV|gukg/410/73
P|EV|bpev-b-kor
P|EV|tbswine/k23/2008/hun
U|EV|jn203
U|EV|donovan
H|EV|akowalik
H|EV|atw391/10
H|EV|lahighpoint
H|EV|allentwood
H|EV|asvartz
H|EV|atexas-12
H|EV|amad-2741-11
H|EV|jag-14
H|EV|aef1080
H|EV|apm-12284-99
H|EV|aev71vnm/27/2011
H|EV|aaw697458
H|EV|aban00-10406
H|EV|ajamm5
H|EV|a03446sd/chn/ev90
H|EV|aban00-10359
S|SV|v2383
S|SV|vrdrf
P|SV|tg4-2013
P|SV|tgar-7
A|SV|tw90a
F|PV|0731
B|aloPV1|gb814852
B|PV|3tlc5f
B|PV|1hc16a
B|PV|25k171
B|PV|2-mh8f
Bo|HuV|abhuv1/2008/hun
O|OHUV1
P|TeschV|gq914053
P|TeschV|wb2-c-tv
P|TeschV|fuyu/2009/2
P|TeschV|10bj02
H|ThV|TMEV-Like|eu376394
H|ThV|gl362
H|ThV|jpn08-404
M|ThV|GDVII|m20562
U|ThV|ngs910
P|EMCV|gx1c
H|CoSV|DHCoSV-D1
H|CoSV|EHCoSV-E1
H|CoSV|BHCoSV-B1
E|RBVp14361|71
B|MiniPiV|jq814851
E|RAV|t10
Bo|RBV|ec11
U|FMDV|r-d45
Bo|FMDV|ind22/2006
P|FMDV|o/s02kor/2002
Bo|FMDV|ind47-93
Bf|FMDV|UK
U|FMDV|sat3-3kenya11/60
U|FMDV|af540910
U|FMDV|sat1/5sal61
U|FMDV|sat1uga1/07
U|FMDV|sat1-7isr14/62
U|FMDV|d/vl2287/2004
H|AV|d/vl2287/2004
U|AV|akov-sewkm
F|AV|afk-13
C|AV|ach-1
P|AV|s-1-hun
Bo|AV|u-1
Cp|AV|12q108
F|SakV|afup1
H|SakV|afup1
Ts|RV|au44
ThTV|3 10878
U|TV|2007167
U|TV|1ucc001
U|TV|100356
A|EMV|calnet
H|HAV|ind-hav-99f
M|FMDV|PV20
PtswineParV11|q316470
S|PV|Vyrie1ho.02.21
M|SebokeleMeeV11
U|LV|145sl
U|LV|64-7855
H|PeV|ba45045
H|PeV|251-39D
H|PeV|rch-21-7/2006
H|PeV|115-4077
H|PeV|gregory
H|PeV|lk-129-93
MH|PeV|sh6

Figure 2.4.1. Phylogenetic tree topology of 209 amino acid sequences corresponding to the *Picornaviridae* capsid VP1 protein. Evolutionary history was inferred by MEGA v6.06 software using the ML method based on the LG model. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 1.1847). The model also allowed for invariant sites (+I, 0.4738%). The reconstruction was performed with partial deletion and all sites with less than 90% coverage were eliminated. Significant clusters are colour coded by branch: Blue) Out-Group I. Red) Out-Group II. Green) Cluster I. Pink) Cluster II. Purple) Cluster III. Specific host species are indicated by coloured round bullets: Purple) Human. Pink) Simian. Turquoise) Bat. Maroon) Bovine. Orange) Porcine. Cyan) Buffalo. Associated symptoms are indicated by coloured triangle bullets: Grey) Unknown. Dark Green) Poliomyelitis. Orange) Gastroenteritis. Red) Respiratory disease. Cyan) Paralytic disease. Dark Blue) Aseptic Meningitis. Light Green) Fever. Pink) HFMD

2.3.3.4.1. Analysis of the evolutionary relationships of within Out-Group I of VP3 and VP1 sequences.

The phylogenetic analysis of the VP3 and VP1 proteins included additional sequences which were representative of the following viruses: HPeV, LV, PaV, Sebokele virus, Seal PiV, FMiPV, SiV, TV, RfV, SaKV, Salivirus and AV. As the VP2 and VP4 proteins are consequential to the cleavage of the precursor VP0, the annotation of these specific viral genomes did not distinguish between the individual VP2 and VP4 CDSs and representative sequences from these viruses were not included in the VP4 and VP2 dataset. Therefore this study focused on the comparison of only the VP1 and VP3 phylogenies inferred for these viruses (Figure 2.4.2). Substantial congruency was observed between the clustering of the VP3 and VP1 sequences, with both sets of sequences forming the respective out-group. With respect to the VP3 sequences the aichivirus sequences clustered together (bootstrap of 98), with closest paraphyletic relation to the representative Salivirus and SaKV protein sequences (bootstrap of 83). This was also observed amongst the VP1 sequences with bootstrap values of 75 and 86 respectively. In both the VP3 and VP1 analysis, these sequences also formed an internal cluster with the RfV, TV and SiV representative sequences. Significant bootstrap values were observed for the monophyletic clustering of TV 2007167 with TV 310878 (bootstrap value of 100) and TV 100356 and SiV 1UCC011 (bootstrap of 99) in the VP3 analysis (Figure 2.4.2a). This corresponding VP1 sequences inferred identical topology with bootstrap values of 100 and 98 respectively (Figure 2.4.2b). Correlations to the VP1 topology reported by Boros *et al* (2013) were observed with respect to both the VP1 and VP3 phylogenies inferred in this study. Boros *et al* (2013) also reported the close relation between AV, TV and Salivirus. Further congruency was observed between VP1 and VP3 with respect to the clustering of the LV and PeV sequences as well as the monophyletic clustering of HAV with EMV (bootstraps of 99 and 100). This, together with the close relation to PaV and Seal

Picornavirus, also correlates to findings reported by Boros *et al* (2013). Incongruent clustering was observed with respect to the marine picornavirus sequences (FMiPV and Sl|PiV). The respective VP1 sequences formed a monophyletic cluster (bootstrap of 57), while the VP3 sequences where more distinctly related with paraphyletic clustering to the LV and PeV sequences (bootstraps of 56 and 98 respectively). A more significant discrepancy was observed with respect to the Sebokele Virus sequences, with the respective VP3 sequence externally clustering with the LV sequences (bootstrap of 92). The incongruences could be a result of genetic recombination, or could be a consequent of selective pressure for host cell immune evasion and receptor specificity observed in the VP1 sequences. It must also be noted that these are RNA virus proteins which are subject to a high mutation rate and genetic drift.

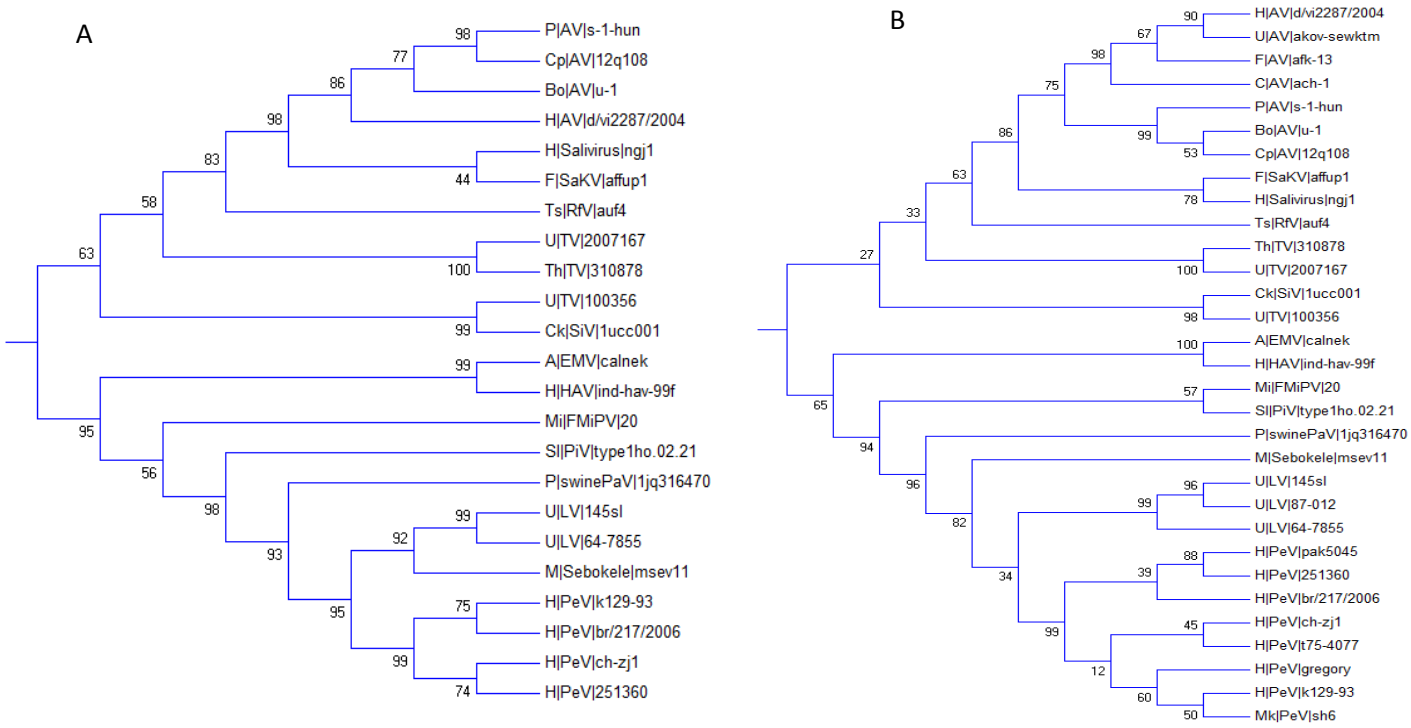


Figure 2.4.2. Phylogenetic out-groups of the respective VP1 and VP3 datasets. A) Out-group corresponding to VP3 sequences. Evolutionary history was inferred by MEGA v6.06 software using the ML method based on the LG model with 90% deletion. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 1.8349). The model also allowed for invariant sites (+I, 1.3190%). B) Out-group corresponding to VP1 sequences. Evolutionary history was inferred by MEGA v6.06 software using the ML method based on the LG model with 90% deletion. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 1.1847). The model also allowed for invariant sites (+I, 0.4738%).

2.3.3.4.2. Analysis of the evolutionary relationships of within Out-Group II of VP1 sequences.

As shown in Figure 2.4.1, Out-Group II of the VP1 sequences corresponded to the out-groups observed in the VP2, VP3 and VP4 phylogenies. The aphthoviruses clustered together, while the cardioviruses and cosaviruses clustered paraphyletic to each other. Greater congruency was observed between the VP2 and VP1 with respect to the clustering of the erbovirus (RBV p1436) with the cosaviruses. In contrast to all other capsid proteins the teschovirus and hunnivirus sequences formed an internal out-group, paraphyletic to the usually observed out-group, with furthest relation to the aphthoviruses. This may be indicative of the higher selectivity and specificity of the VP1 capsid proteins, in relation to the other capsid subunits.

Although the overall clustering of picornaviruses appeared to be independent of host species, sub-clustering of virus serotypes with the same host species was observed. This is evident in Figure 2.4.1 with respect to the sub-clustering of the cardioviruses. The human cardioviruses distinctly clustered from the murine. However this was not observed in the external clustering of FMDV serotypes, with buffalo and bovine isolates clustering together as well as porcine and bovine isolates clustering together. Additional analysis indicated that clustering was geographically dependent with the Southern African (sat) isolates clustering separately from the Indian (ind) and Korean (kor) isolates. This could support inter-typic recombination amongst viral subtypes in the same geographical location or could indicate adaption to the physical environment and climate factors.

2.3.3.4.3. Analysis of the evolutionary relationships within Cluster I

Congruent with VP2, VP4 and VP3 analyses, Cluster I contained the sapelovirus and pigeon picornavirus isolates as well as the bat and feline picornaviruses (shown in Figure 2.4.1). The clustering of the sapeloviruses with the pigeon picornaviruses directly supported the findings by Boros *et al* (2013). Corresponding with the other capsid proteins, significant external clustering was also observed with respect to the bat and feline picornaviruses (bootstrap of 96). In terms of host-cell specificity, the simian sapeloviruses viruses were found to be monophyletic, clustering paraphyletic to both the avian and porcine sapelovirus serotypes. Likewise, the bat picornaviruses clustered distinctly from the feline picornavirus serotype. Although the topology suggests a common ancestor, the bat and feline picornaviruses are still unclassified and are currently considered to be individual viral species (ICTV, 2014).

2.3.3.4.4. Analysis of the evolutionary relationships within Cluster II

In correspondence with the phylogenies of the other capsid proteins, Cluster II (Figure 2.4.1) contained the EV-A, EV-J, EV-H and bovine and porcine enterovirus serotypes, with discrepancies observed with respect to the external clustering of monophyletic strains. An additional discrepancy with respect to all other capsid proteins was observed by the absence of the EV-D serotypes from this cluster. In regard to host-cell specificity, Figure 2.4.1, clearly indicates a correlation between serotypes which infect the same host species. This correlation supports the host-cell specificity required by VP1 proteins.

2.3.3.4.5. Analysis of the evolutionary relationships within Cluster III

Cluster III contained the EV-B and EV-C serotypes, which distinctly clustered from the RV serotypes (bootstrap of 89). Within this cluster, EV-B serotypes significantly segregated from the EV-C serotypes with bootstrap values 99 and 100 respectively. Examination of the external clustering of the EV-B serotypes revealed direct correspondence with the clustering reported by Hu *et al* (2014). The exact correlations were observed with respect to the following strains grouping within the same sub-clusters: JV-1 and Toluca-1, Faulkner and Ohio, Harrington and Tow, Cornelis and CH96-51 and CCHE-29, BAN01-10398 and BAN01-10396. Similar congruency with Piralla *et al* (2013) was observed with respect to the clustering amongst the EV-C serotypes. The RV-C (bootstrap of 100) and RV-A (bootstrap of 100) also formed individual external clusters. The RV-B serotypes clustered distinctly from other RV serotypes. An internal out-group comprising of the EV-D serotypes was also observed, with closest relation to the RV serotypes (bootstrap of 69). Cluster III comprised of enteroviruses which infect human only. Human enteroviruses have been associated with a range of symptomatic illnesses, while some viruses have been isolated from healthy patients and are thought to be opportunistic infections. EV-A serotypes are commonly associated with HFDM, while EV-C serotypes are more common associated with poliomyelitis (Oberste *et* al., 2005).The majority of picornaviruses have high specificity for the gastrointestinal (GI) or respiratory tract, often using sialic acid receptors. The viruses also have secondary affinity the integrin molecules of epithelial cells and cells of the central nervous system (CNS). As VP1 is primarily involved in host-cell receptor recognition, it was hypothesised that viruses with closely related VP1 proteins may display similar pathogenicity within their hosts. However, as shown in Figure 2.5, no correlation was observed, with EV-B, EV-C and EV-A serotypes resulting in a range of diseases. This appeared to be irrespective of viral type and the sub-clustering of viral strains. This suggests that individual enteroviruses have the ability to cause a range of symptoms which may be dependent on host susceptibility. The results also suggest that the viruses have evolved specificity to different cell receptors within the GI and respiratory tract, epithelial tissue and CNS and are thus targeted to the same areas within the host, inflicting similar symptoms regardless of cell receptors.

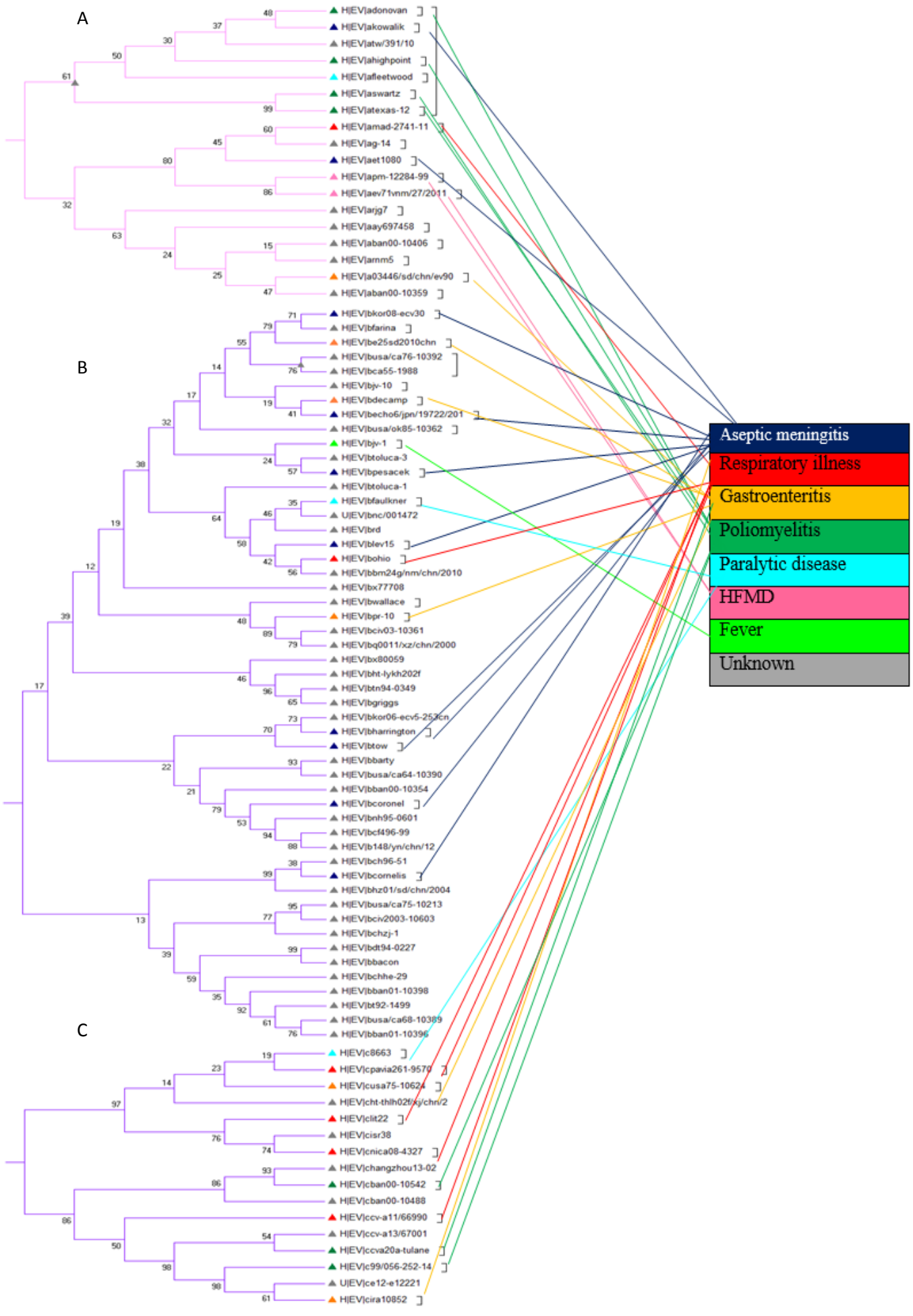

A
B
C
H|EV|adonovan
H|EV|akowalik
H|EV|atw/391/10
H|EV|ahighpoint
H|EV|afleetwood
H|EV|aswartz
H|EV|atexas-12
H|EV|amad-2741-11
H|EV|ag-14
H|EV|aet1080
H|EV|apm-12284-99
H|EV|aev71vnm/27/2011
H|EV|arjg7
H|EV|aay697458
H|EV|aban00-10406
H|EV|arnm5
H|EV|a03446/sd/chn/ev90
H|EV|aban00-10359
H|EV|bkor08-ecv30
H|EV|bfarina
H|EV|be25sd2010chn
H|EV|busa/ca76-10392
H|EV|bca55-1988
H|EV|bjv-10
H|EV|bdecamp
H|EV|becho6/jpn/19722/201
H|EV|busa/ok85-10362
H|EV|bjv-1
H|EV|btoluca-3
H|EV|bpesacek
H|EV|btoluca-1
H|EV|bfaulkner
U|EV|bnc/001472
H|EV|brd
H|EV|blev15
H|EV|bohio
H|EV|bbm24g/nm/chn/2010
H|EV|bx77708
H|EV|bwallace
H|EV|bpr-10
H|EV|bciv03-10361
H|EV|bq0011/xz/chn/2000
H|EV|bx80059
H|EV|bht-lykh202f
H|EV|btn94-0349
H|EV|bgriggs
H|EV|bkor06-ecv5-253cn
H|EV|bharrington
H|EV|btow
H|EV|bbarty
H|EV|busa/ca64-10390
H|EV|bban00-10354
H|EV|bcoronel
H|EV|bnh95-0601
H|EV|bcf496-99
H|EV|b148/yn/chn/12
H|EV|bch96-51
H|EV|bcornelis
H|EV|bhz01/sd/chn/2004
H|EV|busa/ca75-10213
H|EV|bciv2003-10603
H|EV|bchzj-1
H|EV|bdt94-0227
H|EV|bbacon
H|EV|bchhe-29
H|EV|bban01-10398
H|EV|bt92-1499
H|EV|busa/ca68-10389
H|EV|bban01-10396
H|EV|c8663
H|EV|cpavia261-9570
H|EV|cusa75-10624
H|EV|cht-thlh02fkj/chn/2
H|EV|clit22
H|EV|cisr38
H|EV|cnica08-4327
H|EV|changzhou13-02
H|EV|cban00-10542
H|EV|cban00-10488
H|EV|ccv-a11/66990
H|EV|ccv-a13/67001
H|EV|ccva20a-tulane
H|EV|c99/056-252-14
U|EV|ce12-e12221
H|EV|cira10852
Aseptic meningitis
Respiratory illness
Gastroenteritis
Poliomyelitis
Paralytic disease
HFMD
Fever
Unknown

Figure 2.4.3. Correlation between phylogenetic clustering of VP1 sequences and associated symptoms of EV-A, EV-B and EV-C serotypes. Evolutionary history was inferred by MEGA v6.06 software using the ML method based on the LG model. A discrete gamma distribution was used to model evolutionary rate differences among sites (+G, parameter = 1.1847). The model also allowed for invariant sites (+I, 0.4738%). The reconstruction was performed with partial deletion and all sites with less than 90% coverage were eliminated. A) EV-A cluster. B) EV-B Cluster. C) EV-C cluster. Associated symptoms are indicated by coloured triangle bullets: Grey) Unknown. Dark Green) Poliomyelitis. Orange) Gastroenteritis. Red) Respiratory disease. Cyan) Paralytic disease. Dark Blue) Aseptic Meningitis. Light Green) Fever. Pink) HFMD.

2.4 Conclusions

The phylogenies inferred with respect to the different capsid proteins have a substantial amount of congruency. As observed for each of the capsid proteins, the representative sequences of the cardioviruses, cosaviruses, aphthoviruses, erboviruses, teschoviruses and hunnivirus always clustered out-side of a super-grouping comprised of sequences of the sapelovirus and bat and feline picornaviruses and the enteroviruses. Furthermore the close relation between the cosavirus and cardiovirus proteins was observed across all capsid phylogenies. This too was the case for the relation between the aphthoviruses and teschoviruses and hunniviruses. The grouping of the sapelovirus sequences with the unclassified bat, feline and pigeon picornaviruses was also observed across the respective subunit phylogenies. In regard to the VP1 and VP3 sequences, direct correlation was observed with respect to the common grouping of the LV and PeV sequences, as well as the close relation between the AV, Salivirus and TV sequences. Viruses of the *Enterovirus* genus always formed a complex cluster, with sub-groupings representative of individual viral types. This group was consistently observed to have closets relation to the SV and bat, feline and pigeon picornaviruses. The topology inferred for all four of the capsid proteins was found to be directly dependent on viral type and genera and no broad host co-phylogeny observed. However a certain amount of host specificity was observed between virus sub-types which infect different species. Furthermore, the analysis of the enterovirus VP1 phylogenies with regard to viral pathogenicity revealed no direct correlation between sub-types and strains which are causative of the same diseases. Rather, it was concluded that enteroviruses have affinity for a range of epithelial and nervous system tissue which may result in systemic infections dependent on host susceptibility.

Analysis of the external clustering within the main sub-groups revealed significant discrepancy in the monophyletic relationships across the four capsid proteins. The results indicated incongruences across the topologies with respect to closely related viral species,

viral sub-types or viral strains. The greatest disagreement was pertained to the enteroviruses, particular in the clustering of the RV serotypes. It has been concluded that the incongruences observed between closely related viruses is indicative of inter-typic recombination and support suggestions by Heath *et al* (2006) that the viral capsid proteins are may be functionally interchangeable closely related virus subtypes. The bootstrap analysis of the external grouping of viral strains revealed significantly lower values with respect to those observed at the internal nodes. Thus there was no clear path of vertical speciation observed between viral strains and serotypes. Furthermore the topologies inferred in this study were not consistent with those of the replication proteins, as reported by Phelps *et al* (2013). Therefore the findings in this study support the notation that the capsid proteins are evolving independently from the replication proteins. As capsid proteins are more exposed to evolutionary pressures to evade host immune systems and vaccines, they appear to have the ability to evolve without comprising attachment to the host cell receptor. However the viral replication proteins may be more inclined to remain conserved, such that mutations do not comprise virus replication. Therefore it is possible to observe the independent evolution of the two sets of viral proteins.

The phylogenetic reconstruction with respect to each of the different capsid proteins was found to be a complex procedure. Significantly low bootstrap values observed, particular in the VP1 and VP2 datasets. This iterates the limitations of phylogenetic analysis, with most current models representing general protein databases. As the picornavirus proteins are subject to the high mutation rate and genetic drift of RNA viruses, their evolutionary pattern may be unique and thus cannot be described according to current available models. Furthermore the continuous infection of new hosts, derivation of quasispecies and genetic recombination impose evolutionary pressures unique to these viruses.

3. Short linear motif prediction

The capsid of picornaviruses is composed of 60 protomers, assembled to form an icosahedron with pseudo = T3 symmetry. Each protomer results from the interaction between its subunits: VP1, VP3 and precursor VP0 which is subsequently cleaved into VP2 and VP4 during maturation of the viral capsid. Short linear motifs (SLiMs) are well-known to function in protein-protein interactions (Diella *et al*., 2008; Neduva and Russell, 2006). Therefore the prediction of possible SLiMs within the subunits of picornavirus capsids may further the understanding of residues which are critical for virus assembly. In this chapter an exhaustive motif analysis was performed for a representative dataset of each of the subunit proteins, using the sequence analysis tool developed by Bailey and Elkan (1994): Multiple-EM for Motif Elicitation (MEME). The analysis aimed to identify motif conservation across the viral family, with specific identification of SLiMs which are conserved across: 1) strains of individual virus types and 2) different viruses which infect the same host species. Motif conservation was also compared with the phylogenetic results of Chapter 2. The conservation of functional motifs across virus sub-types may further support the theory of genetic recombination discussed in Chapter 2. Subsequent to the identification of highly conserved motifs, further analysis involved the mapping of motifs to representative crystal structures of capsid protomers. An *in silico* prediction of residues involved in protein-protein interactions within the PDB files was also performed. The objective was to identify principle residues which were predicted to play a role in motif-motif interactions as well as determine the specific residue conservation across all available strains of the respective virus sub-groups. A detailed flowchart of the methodology of this chapter is depicted in Figure 3.1.

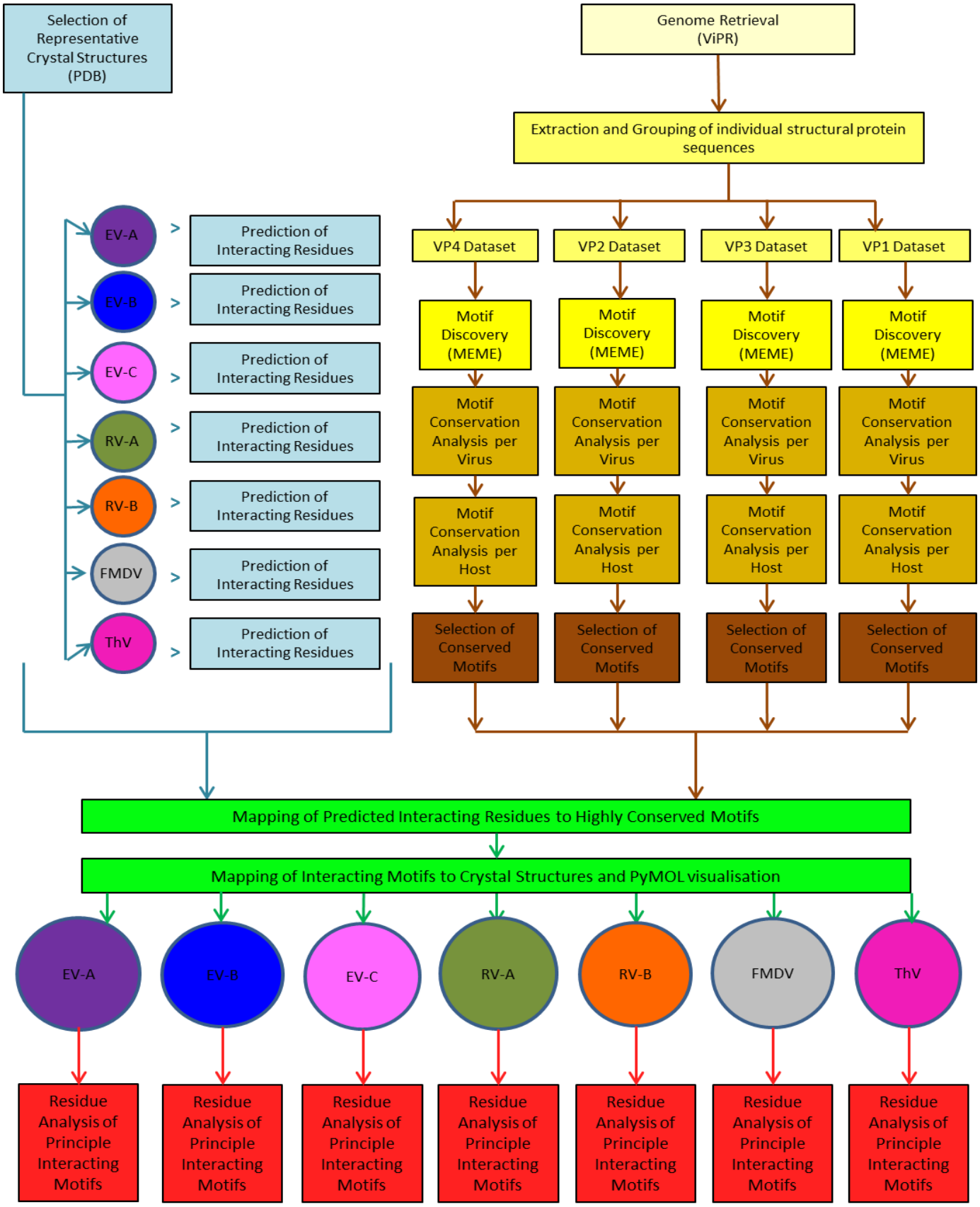


Figure 3.1. Methodology for the prediction of SLiMs which may facilitate viral subunit-subunit interactions required for assembly of promoter intermediates in capsids of picornaviruses. Crystal structures are depicted as circles. Rectangles indicate methodology processes. Blue) Prediction of interacting residues in available crystal structures of picornavirus capsid protomers. Yellows) Motif discovery. Green) Structural mapping of motifs. Red) Residue analysis of principle interacting motifs

3.1 Introduction

An amino acid motif is a subsequence pattern which has repeated occurrence in a group of related proteins. This repeated occurrence provides evidence that the motif did not occur by chance but rather serves a common biological function (Bailey and Elkan, 1994). Many protein sequences have been found to contain SLiMs which function in recognition and targeting activities, thus facilitating protein-protein interactions. These motifs are considered linear with typical length of 3-11 consecutive amino acids (Davey *et al*., 2012). To allow for the discovery of motifs which may be marginally longer than average, this study specifically searched for motifs ranging from 3-20 amino acids. Motif discovery was facilitated by MEME and was performed individually for datasets containing VP1, VP2, VP3 and VP4 protein sequences respectively.

MEME was developed by Bailey and Elkan (1994) and allows for the discovery of motifs in a group of related DNA or protein sequences. The program incorporates a two-component mixture model which allows for the statically detection of multiple motifs. Given an unaligned dataset, MEME conceptually divides the dataset into all possible n (overlapping) subsequences of length W. A probabilistic model for each motif is generated, with one component describing the set of subsequences of length W and the other component describing the background of all other positions within the sequence. The model estimates the number of occurrences of the motif in each sequence, the optimal width W and the description of the motif (Bailey and Elkan, 1994). MEME results are presented as position dependent, residue probability matrices which hypothesize the probability of the occurrence of each residue at each possible position within the motif. MEME also returns the statistical significance of each motif in the form of an E-value. More explicitly the E-value describes the expected number of motifs with the same width and site count that would be found in a similarly sized set of random sequences. Furthermore for every site of occurrence, MEME returns a p-value indicating the probability of the occurrence of a random string, generated from the frequencies of background residues (Bailey *et al*., 2009). The MEME suite also incorporates the program Motif Alignment and Search Tool (MAST) (Bailey and Gribskov, 1998), which can be used to further analyse MEME results such that any unsuitable motifs be identified and consequently removed from the results.

3.2 Methods and Materials

3.2.1 Sequence retrieval and dataset

Sequence retrieval was as described in Section 2.2.1. However as MEME analysis is not subject to the limitations concerned with phylogenetic reconstruction, the motif analysis could be performed across significantly larger datasets. Thus only redundant sequences of 100% identity were removed from each dataset. This was facilitated by scripting which incorporated iterative pairwise alignments and percentage identity calculations (Appendix 1.2). The original and final sizes of each dataset are presented in Table 3.1. The sequence headers were edited with standardized abbreviations corresponding to the format of Host|Virus|Strain. The abbreviations and corresponding full labels are presented in Table 2.2., while detailed lists of proteins sequences for each structural group can be found in Appendices 2.1.-2.4.

Table 3.1. Respective sizes of each structural protein dataset for motif analysis. Protein sequences corresponding to 2185 individual picornaviruses were downloaded from ViPR. The structural proteins were individual extracted and grouped, with all redundant sequences of 100% identity removed.

Structural Protein Dataset	Total Number of Sequences Extracted	Total Number of Sequences After Filtration
VP1	1965	1289
VP2	1884	972
VP3	1965	998
VP4	1804	451

3.2.2 Prediction of short linear motifs

Motif discovery was facilitated by MEME v4.90 and performed individually for each of the filtered structural protein datasets. The MEME parameters were set to maximise the number of motif predictions to 100, with minimum width of 3 amino acids and maximum width of 20 amino acids. Each set of MEME results were subsequently subjected to MAST for identification of any over-lapping or unsuitable motifs.

3.2.3 Analysis of motif predictions

Python scripting was used to analyse the respective MEME and MAST output files of each protein dataset (Appendix 1.4). MAST allowed for the identification of over-lapping motif regions. Thus all unsuitable motifs identified by MAST were removed from each set of

results. Scripting also facilitated the exclusion of insignificant motifs (E-value >0.05) as well as the exclusion of any sequences for which a specific motif was found to be insignificant (p-value > 0.05). Each structural protein dataset was sequentially divided into two sets of sub-groups. Firstly representative sequences of respective virus types were grouped together. Secondly, the data was re-grouped such that sequences of different virus which infect the same host species were clustered together. The respective number of sequences in each of these sub-groups is shown in Table 3.2. Python scripting was used to group the data according to the respective sequence headers, which served as a key for the host and species of individual virus sequence (Host|Virus|Strain). An iterative scripting analysis of each MEME text file was performed with respect to all identified motifs. Figure 3.2 illustrates a flowchart of the algorithm used to determine overall motif conservation across respective sub-groups. Specifically the script mapped the sequence headers of respective motif sites to identical headers within specific virus and host subgroups. Subsequent calculation of sequence specific motif conservation was calculated with respect to each sub-group. Conservation heat maps were constructed for all motifs across each sequence in each sub-group, as well as the overall conservation of each motif across each sub-group. Highly conserved motifs were selected for additional structural analysis and the *in silico* prediction of residues involved in protein-protein interactions within representative crystal structures of capsid protomers.

3.2.3 Selection of Crystal Structures

Representative crystal structures of virus capsid were obtained for the EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV viruses. Structures with the highest quality, as determined by lowest resolution and R-value, were selected and downloaded from the Protein Databank (PDB).

3.2.4 Prediction of protein-protein interactions

The PDB files of each of the representative virus protomers were individually submitted to the PIC webserver (Tina *et al*., 2007) for the identification of specific amino acids predicted to participate in protein-protein interactions within a multichain protein structure. The default settings of the PIC webserver were implemented for the prediction of: 1) hydrophobic interactions within 5A, 2) main chain-main chain hydrogen bonds, 3) main chain-side chain hydrogen bonds, 4) side chain-side chain hydrogen bonds and 5) ionic interactions within 6A. The results were stored in respective text files for subsequent scripting analysis.

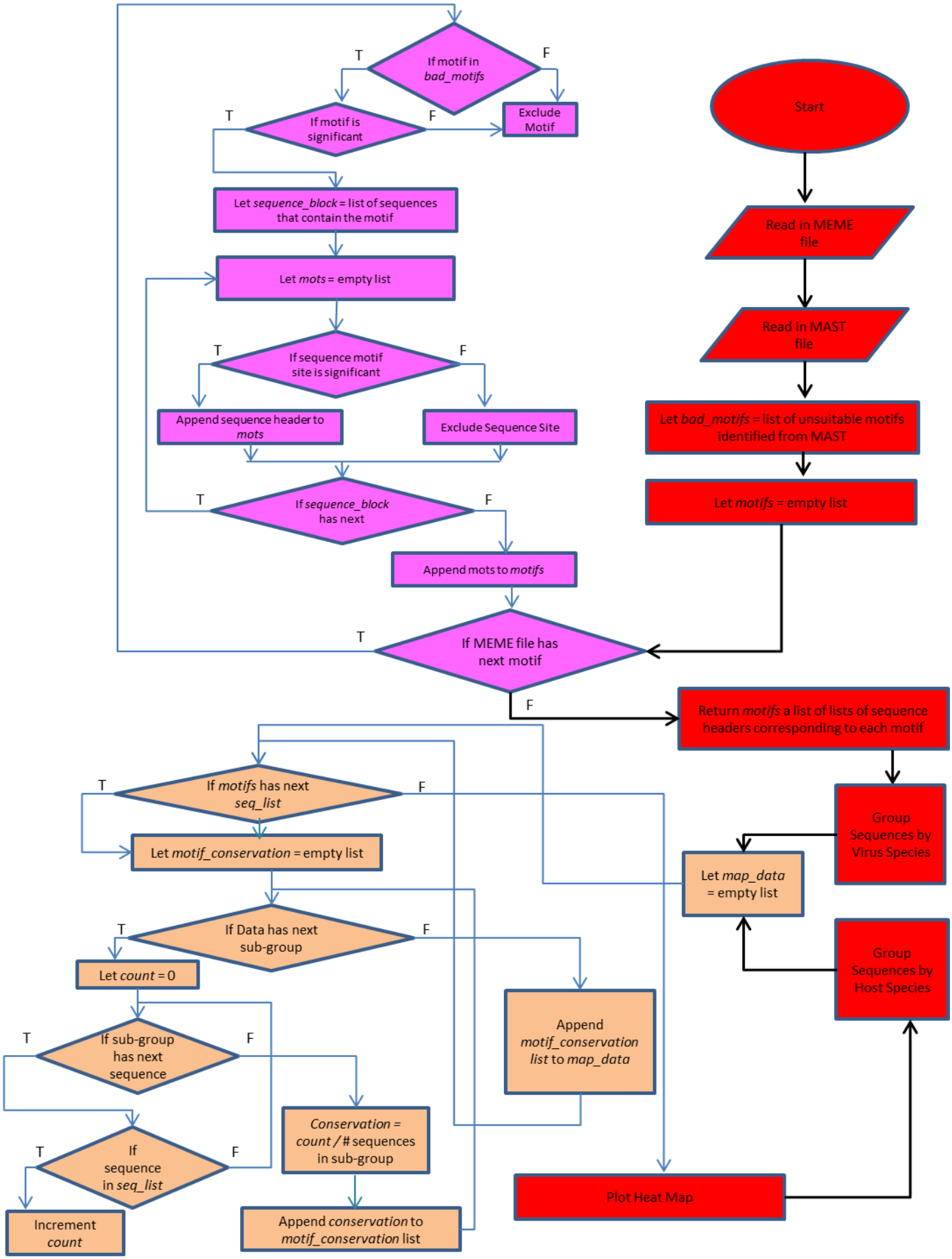


Figure 3.2. Algorithm for motif conservation calculation. The flowchart shows the algorithm for calculating overall motif conservation across individual sub-groups. Parallelogram) Input. Rectangle) Processing. Diamond) Decision (T: True; F: False). Red) Main process. Orange) Conservation calculation sub-process, iteratively performed for each sub-group. Pink) Sub-process: Retrieval of sequence specific motif sites, performed iteratively for each motif in the MEME text file.

3.2.5 Structural mapping of interacting motifs

The PIC results of each representative virus was analysed by Python scripting (Appendix 1.5). The input parameters of the script included: the MEME text files for each structural protein dataset, the sequence header of the representative structural sequence, the PIC output file corresponding to the PDB file of the structural sequence and lists of the highly conserved motifs, previously identified, in each structural protein. Given the motif ID and sequence header, the script extracted the residue positions of each motif from the relative MEME text files. An iterative analysis was then performed with respect to each predicted interaction listed in the PIC output file. For each interaction the script extracted the positions and chains of each respective residue. The position of each residue was then iteratively matched against the position of each residue in each highly conserved motif within each corresponding chain. Thus the script allowed for the identification of highly conserved motifs which contained amino acids predicted to be interacting, with further identification of particular motif-motif interactions between the subunits of each representative virus protomer. The script exhaustively identified motif-motif interactions between all combinations of the capsid subunits, specifically between: 1) VP1 and VP2 subunits, 2) VP1 and VP3 subunits, 3) VP1 and VP4 subunits, 4) VP2 and VP3 subunits, 5) VP2 and VP4 subunits and 6) VP3 and VP4 subunits. A descriptive flowchart of this algorithm is depicted in Figure 3.3. The script output included a list of subunit motif-subunit motif interactions, a list of motif specific interacting residues with the corresponding type of interaction and a list of all motifs, with respect to each structural protein, which was predicted to facilitate subunit motif-subunit motif interactions. Additional Python scripting (Appendix 1.6), with the incorporation of PyMOL commands, facilitated the mapping of the predicted interacting motifs to the respective PDB files. The mapped protomers were visualised in PyMOL (Schrodinger, 2010).

3.2.6 Analysis of motif-specific interacting residues

The total number of predicted interactions per residue within principle motifs was calculated with respect to the relative block sequence of the respective representative viruses. Python scripting and the Matplotlib was used for the generation of histogram plots of the total number of interactions per residue in each motif (Appendix 1.5). Script input included: the MEME text files for each structural protein dataset, the sequence header of the representative structural sequence, the PIC output file corresponding to the PDB file of the structural sequence and lists of the relative motifs, previously identified, in each structural protein. Given the motif ID and sequence header, the script extracted the residue positions and the sequence specific block diagram of each motif from the relative MEME text files. The

residue positions corresponding to each interaction listed in the PIC output file were iteratively matched against the residue positions of each motif. Thus the total number of predicted interactions per residue of each motif was calculated.

Residue conservation was calculated across all available strains of the respective virus sub-groups: EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV. Specifically script input included: MEME output file of a given structural protein dataset, a list of IDs of conserved motifs and a list of virus sub-groups. For every given motif, the script extracted the corresponding regular expression from the MEME output file. Subsequently, for every given virus-subgroup, the sequence specific block diagrams of all sequences representative of the relative virus and corresponding to the particular motif were extracted from the MEME output file. The number of occurrences of each residue within the regular expression across all sequence specific block diagrams was iteratively counted. This step was iteratively performed for each virus sub-group. Moreover, the script also counted the number of occurrences of residue substitutions which were exceptions to the regular expression. Thus conservation was calculated individually for each residue in each motif as a fraction of the number of sites at each position per total number of sequences (Table 3.2) within each sub-group. Matplotlib was used to construct histogram plots of residue conservation within each sub-group for each motif. Upper-case letters were used to indicate residues of the regular expression, while lower-case letters indicated the presence of residues which were not present in the regular expression of the motif. The script has been included as Appendix 1.7.

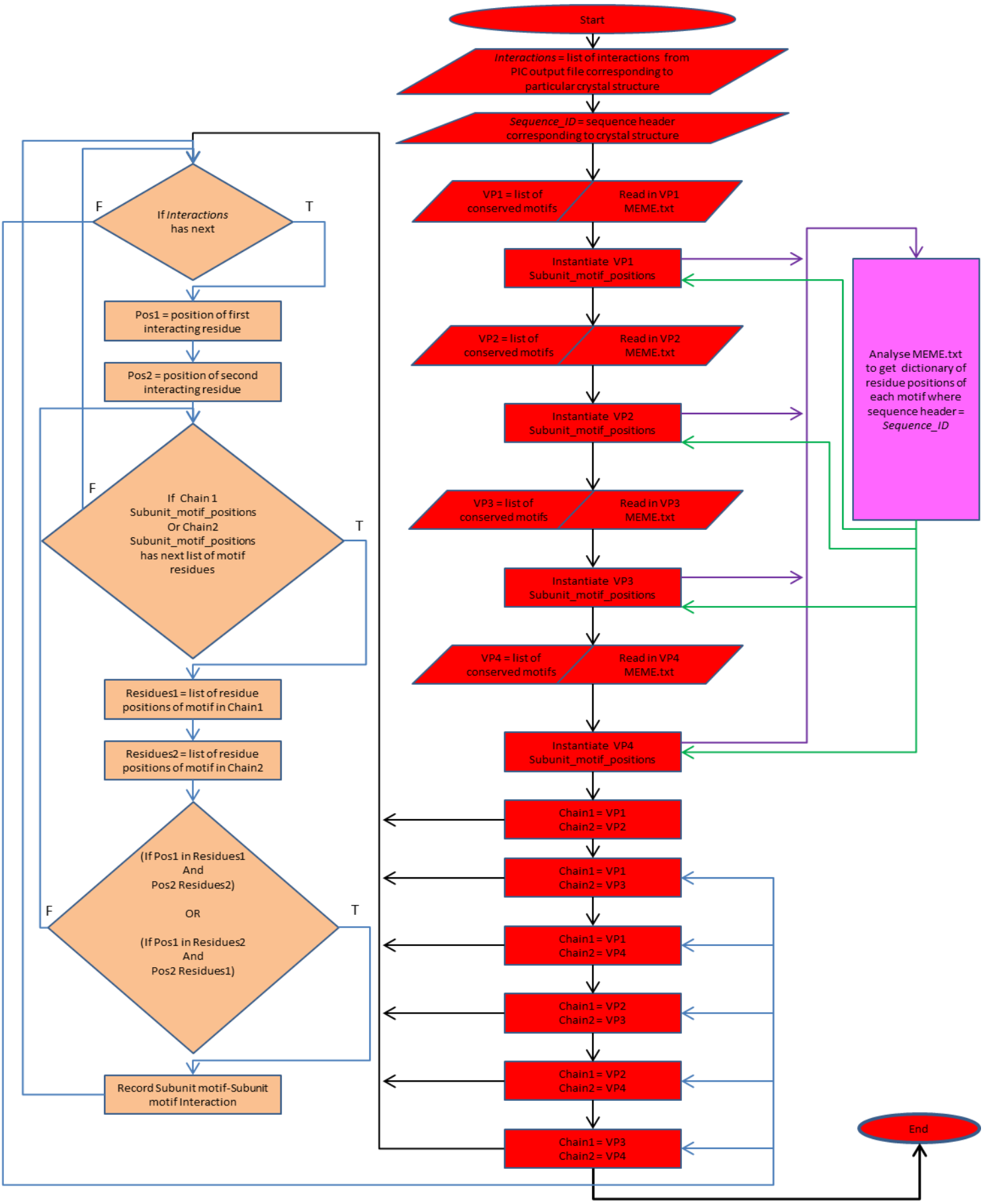


Figure 3.3. Algorithm for prediction of subunit motif-subunit motif interactions. . Parallelogram) Input. Rectangle) Processing. Diamond) Decision (T: True; F: False). Red) Main process. Pink) Sub-process: Retrieval of motif residue positions specific to the representative viral sequence. The process is iteratively called for all conserved motifs in each structrual protein dataset. Orange) Sub-process: Mapping of PIC predicted interacting residue positions within the representative crystal structure, to the residue positions of motifs within corresponding subunit proteins. The process is iteratively called for all possible subunit-subunit interactions.

Table 3.2. Respective sizes of each sub-group of each structural protein dataset. Protein sequences corresponding to 2185 individual picornaviruses were downloaded. The structural proteins were individual extracted and grouped, with all redundant sequences of 100% identity removed. Each protein dataset was sub-divided into groups corresponding to individual virus type and individual host species.

Sub-Group	Number of Sequences per Structural Protein Dataset			
	VP1	VP2	VP3	VP4
Hosts				
Avian	7	6	7	5
Bat	7	6	8	7
Canine	5	0	3	0
Feline	7	5	6	3
Murine	7	6	5	2
Primates	844	628	631	313
Ungulates	21	143	140	57
Uncommon Hosts	8	2	7	0
	VP1	**VP2**	**VP3**	**VP4**
Viruses				
AV	19	0	15	0
CoSV	3	3	3	3
EMCV	25	13	9	7
EMV	3	3	3	2
EV-A	223	147	111	61
EV-B	157	147	147	121
EV-C	254	176	181	55
Other EV	27	26		20
FMDV	268	229	224	65
FMiPV	1	0	1	0
HAV	41	23	24	0
HuV	1	1	1	1
IaioPiV	1	1	1	1
LV	5	0	4	0
MiniPiV	1	0	1	1
OHUV	1	1	1	1
PeV	32	0	32	0
PiV	13	12	13	10
RAV	8	7	7	2
RBV	2	2	2	2
RfV	1	0	1	0
RV-A	94	89	89	35
RV-B	30	31	29	22
RV-C	18	17	18	16
SaKV	1	0	1	0
Salivirus	1	0	1	0
Sebokele Virus	1	0	1	0
SiV	1	0	1	0
SV	8	7	7	6
SwinePaV	2	0	2	0
TeschV	7	5	6	4
ThV	34	32	30	16
TV	6	0	6	0

3.3 Results and Discussion

3.3.1 Analysis of motif predictions

3.3.1.1 Motif analysis of VP4 dataset

The VP4 dataset submitted to MEME consisted of 451 unaligned protein sequences which were representative of 26 individual picornaviruses isolated from a range of six host species. The MEME parameters were set to identify a total of 100 motifs with widths from 3 to 20 amino acids. The results were subsequently subjected to a MAST analysis for the identification of unsuitable or over-lapping motifs. Both the MEME and MAST result sets were successively were analysed by scripting. Motifs 13, 16, 22-31, 33-100 were identified as unsuitable and thus removed prior to conservation analysis. Furthermore scripting also facilitated the exclusion of any insignificant motifs (E-Value >0.05), as well as the exclusion of sequence specific motif sites with statistical insignificance (p-value > 0.05). Thus motifs 21 and 32 were also excluded and the conservation analysis was performed with respect to the remaining 18 significant motifs. For conservation analysis, the dataset was sub-divided according to 1) sequences representative of the same virus type and 2) sequences representative of viruses which infect the same host species. A total of 21 viral sub-groups and six host sub-groups were respectively generated. Due to the limited number of sequences from EV-D, EV-F, EV-G, EV-H, EV-J and the porcine enteroviruses, these sequences were grouped as a single sub-group of other EV. Likewise the bat picornavirus, feline picornavirus and pigeon picornavirus were grouped as PiV. A detailed list of the sub-groups with respective number of sequences is shown in Table 3.2. Heat maps illustrating the overall conservation of each motif across each sub-group (Figures 3.4 and 3.5) were constructed. Moreover the conservation with respect to each sequence within each sub-group was also mapped (Appendix 5.1). It was observed that motif conservation was virus dependent, with no obvious correlation between different viruses which infect the same host species. Figure 3.5 indicates a bias conservation of motifs across the feline and murine species as these sub-groups only contain feline picornavirus and ThV sequences respectively, and is therefore a result of conservation across these individual viral sub-groups. Furthermore the conservation of motifs 1-3 across primate species was due to the large number of EV and RV sequences. An analysis of Figure 3.4 indicates that this conservation was actually unique to the enterovirus genus. Consequent to this observation motif conservation was analysed comprehensively with respect to individual virus species.

As shown in Figure 3.4, motif conservation was observed to be specific to closely related viral species with no motif being conserved across the *Picornaviridae* family. Motif 3 (E-value = 4.3e-3304) had the highest number of sites (339) with complete conservation across the EV and RV serotypes and high conservation across the FMDV serotypes. The motif was also present in the aphtovirus RAV, hunnivirus and the bat and pigeon picornaviruses (Appendix 5.1.1). Motif 1 and 2, also with significant E-values of 1.1e-4930 and 5.4e-4059 respectively, were unique to the enterovirus genus with complete conservation across the EV and RV serotypes. Thus supporting the phylogenetic clustering presented in the previous chapter. Moreover, Motif 4 (E-value = 5.1e-1451) was completely conserved across all sequences which clustered to form a phylogenetic out-group including the aphthoviruses (FMDV and RAV), erbovirus (RBV), teschovirus, hunnivirus, the unclassified MiniPiV, cosavirus and the cardioviruses (EMCV and ThV) sequences. Motif 14 was also found to be unique to certain viruses within the out-group including the monophyletic teschovirus and hunnivirus as well as all RBV strains. Likewise motif 10 was completely conserved across the cardioviruses and RAV strains while motif 8 was completely conserved across the cardioviruses and teschovirus strains. The phylogenetic super-group was supported by motif 7 which was highly conserved across the enteroviruses, sapeloviruses and the bat picornaviruses, while the close evolutionary relationship between the sapeloviruses and the unclassified bat, feline and pigeon picornaviruses was supported by motif 9. Although motif conservation was not observed across the *Picornaviridae* family, significant findings existed in the conservation of motifs across sequences from the larger sub-groups of the enteroviruses, FMDV and ThV. This observation may provide further evidence of genetic recombination between virus subtypes, as discussed in Chapter 2. Moreover it may also support the proposal that function units of the capsid proteins may be interchangeable between closely related species (Health *et al*., 2006).

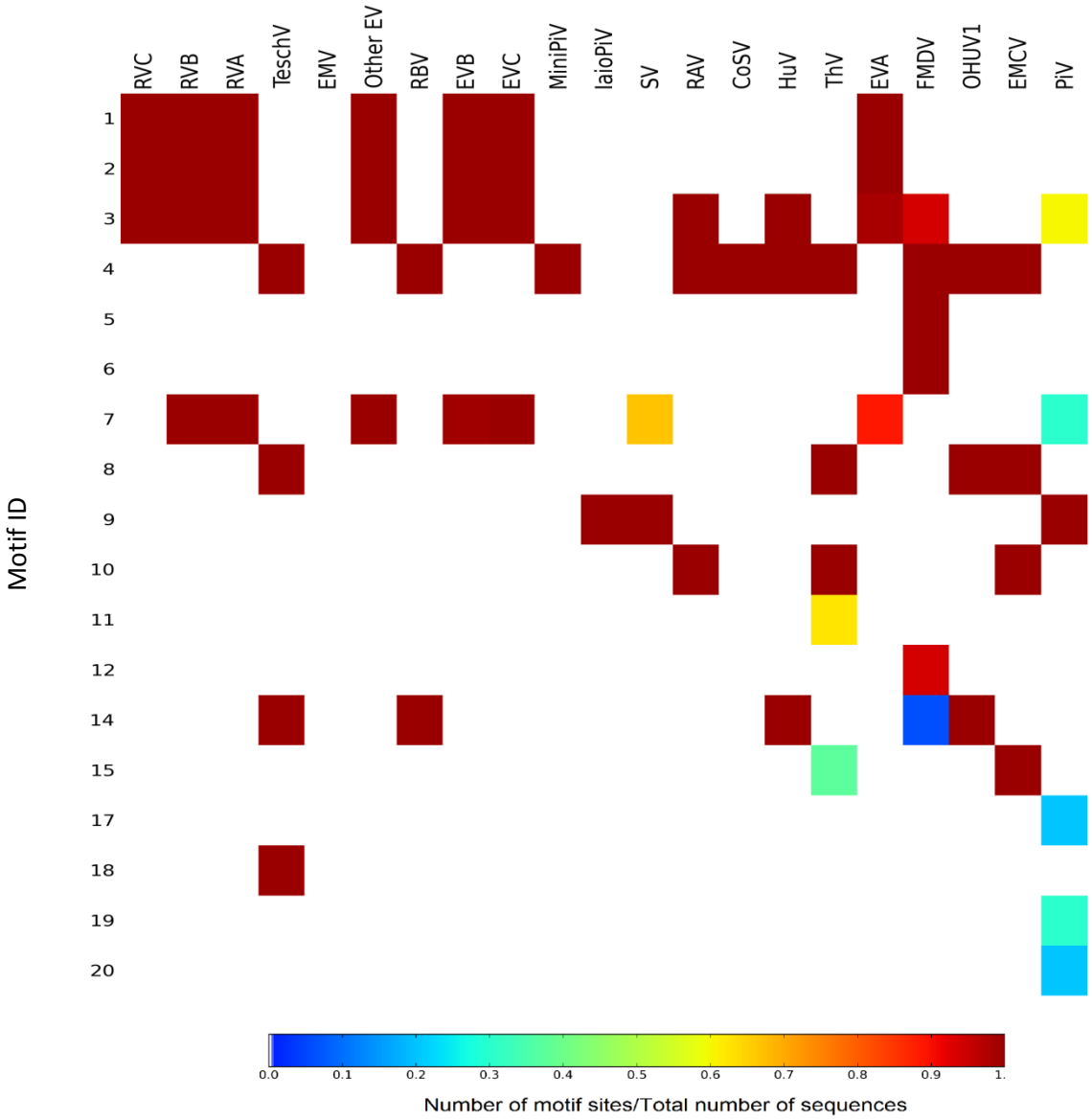


Figure 3.4. Heatmap of VP4 motif conservation across picornavirus species. All available picornavirus VP4 protein sequences were subjected to MEME for motif discovery. The conservation of each motif was calculated with respect to sub-groups comprising of sequences from the same virus species. The Heatmap was generated using the Matplotlib and Python scripting.

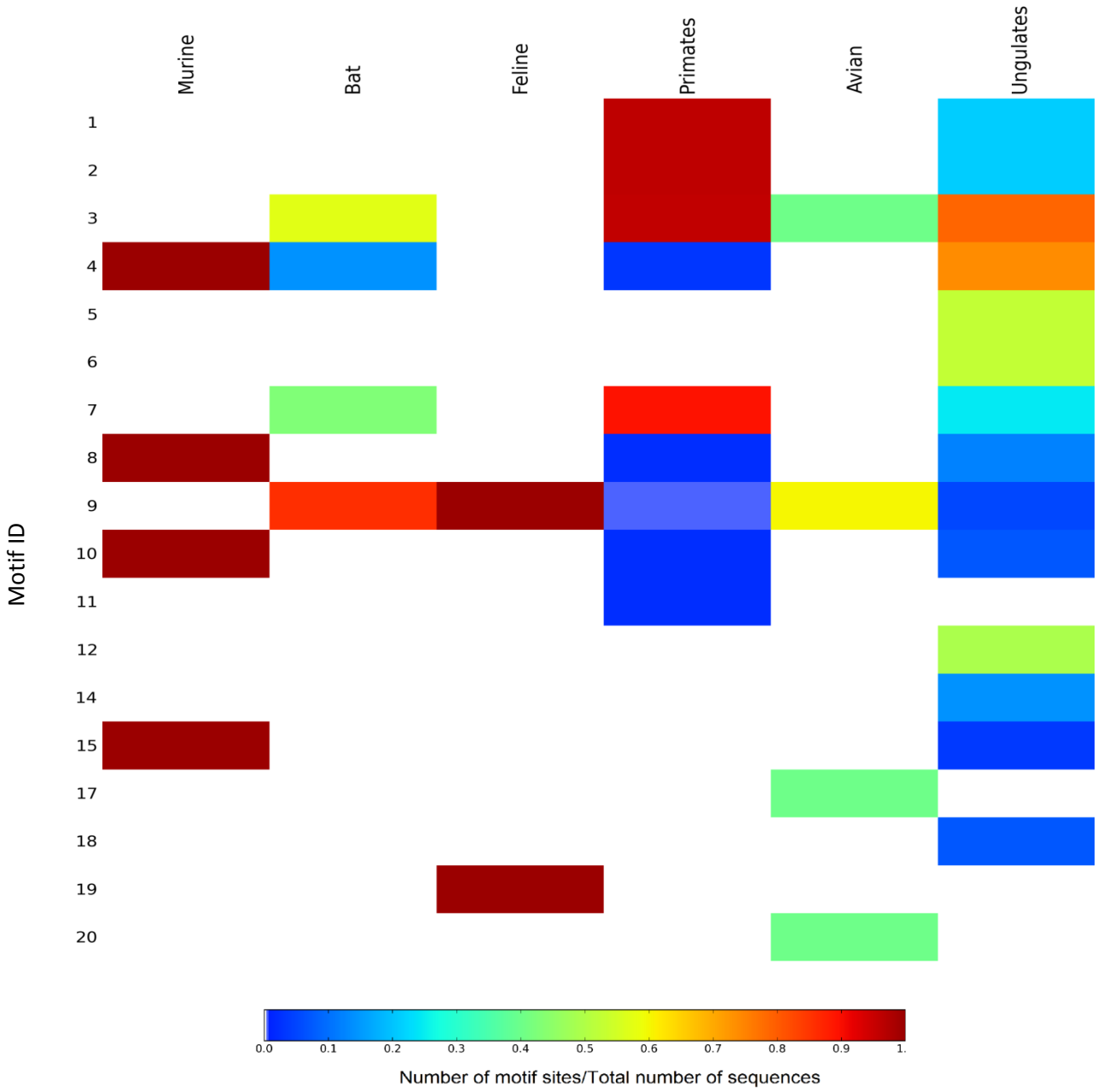


Figure 3.5. Heatmap of VP4 motif conservation across host species of respective picornaviruses. All available picornavirus VP4 protein sequences were subjected to MEME for motif discovery. The conservation of each motif was calculated with respect to sub-groups comprising of sequences from virus of the same species. The Heatmap was generated using the Matplotlib and Python scripting.

3.3.1.2 Motif analysis of the VP2 dataset

The VP2 dataset was comprised of 972 unaligned protein sequences, corresponding to the VP2 capsid protein of 26 individual picornaviruses, isolated from a range of eight different host species. The MEME parameters were set to identify a maximum of 100 motifs ranging in width from 3 to 20 amino acids. Scripting identified 76 significant motifs (E-value <=0.05) of which motifs 39, 65, 67, 70 and 73 were excluded due to unsuitability as predicted by

MAST analysis. Motif conservation analysis was thus performed on the remaining 71 motifs, subsequent to the removal of insignificant sequence specific sites (*p*-value >0.05). Overall motif conservation was determined across 21 virus specific sub-groups, with EV-D, EV-F, EV-G, EV-H, EV-J and the porcine enteroviruses grouped as "Other EV". Similarly conservation was determined across seven host specific sub-groups, with sequences from viruses which infected rare host species grouped as "Uncommon Hosts". This group consisted of tiger and tick species. The particulars of each sub-group are presented in Table 3.2, while the respective heat maps are presented in Figures 3.6 and 3.7. Further analysis involved the mapping of conservation with respect to each individual protein sequence within each sub-group (Appendix 5.2). It was observed that motif conservation was once again virus dependent. However as indicated by Figure 3.7, there was a proportion of motifs which were highly conserved across host species. This result was also found to be bias, as the conserved motifs matched those motifs which were conserved across the virus family (Figure 3.6). As indicated by the low conservation of the remaining motifs, no correlation between proteins from virus of the same host species was observed. Consequent to this observation motif conservation was analysed comprehensively with respect to individual virus species.

Conservation analysis with respect to the 21 viral sub-groups revealed significant conservation of 13 motifs across the *Picornaviridae* family. The VP2 protein forms an integral part of the external virus capsid, playing a role in both antigenicity and host cell receptor binding. As this is highly specific to individual viruses, as well as individual virus strains, motifs involved in these processes should not be conserved across the *Picornaviridae* family. This provides evidence that motifs conserved across the family may be of significant functional importance other than host cell receptor binding and antigenicity and may rather facilitate protein-protein interactions within the subunit interface. As shown in Figure 3.6, the highly conserved motifs included motifs 1-11 and 14-15, each of which is discussed in detail in the following section. Motif conservation was also analysed in comparison to the phylogenetic groupings observed in the previous chapter. However as motif conservation was not virus specific, few explicit correlations could be observed. Although not all unique to the enteroviruses, motif 1-11 were 100% conserved across all EV and RV serotypes, with motif 10 being unique to sequences of the enterovirus genus. Motif 8 was unique to the EV and RV serotypes as well as the SV and bat and feline picornaviruses (Appendix 5.2.1). This correlated directly to the phylogenetic super-grouping of these viruses. With respect to phylogenetic out-group, the monophyletic relationship between HAV and EMV was

supported by motifs 29, 32, 33, 34 and 51, all of which were unique to these viruses. Further support of the VP2 phylogeny was observed in the unique conservation of motif 37 across cardioviruses EMCV and ThV.

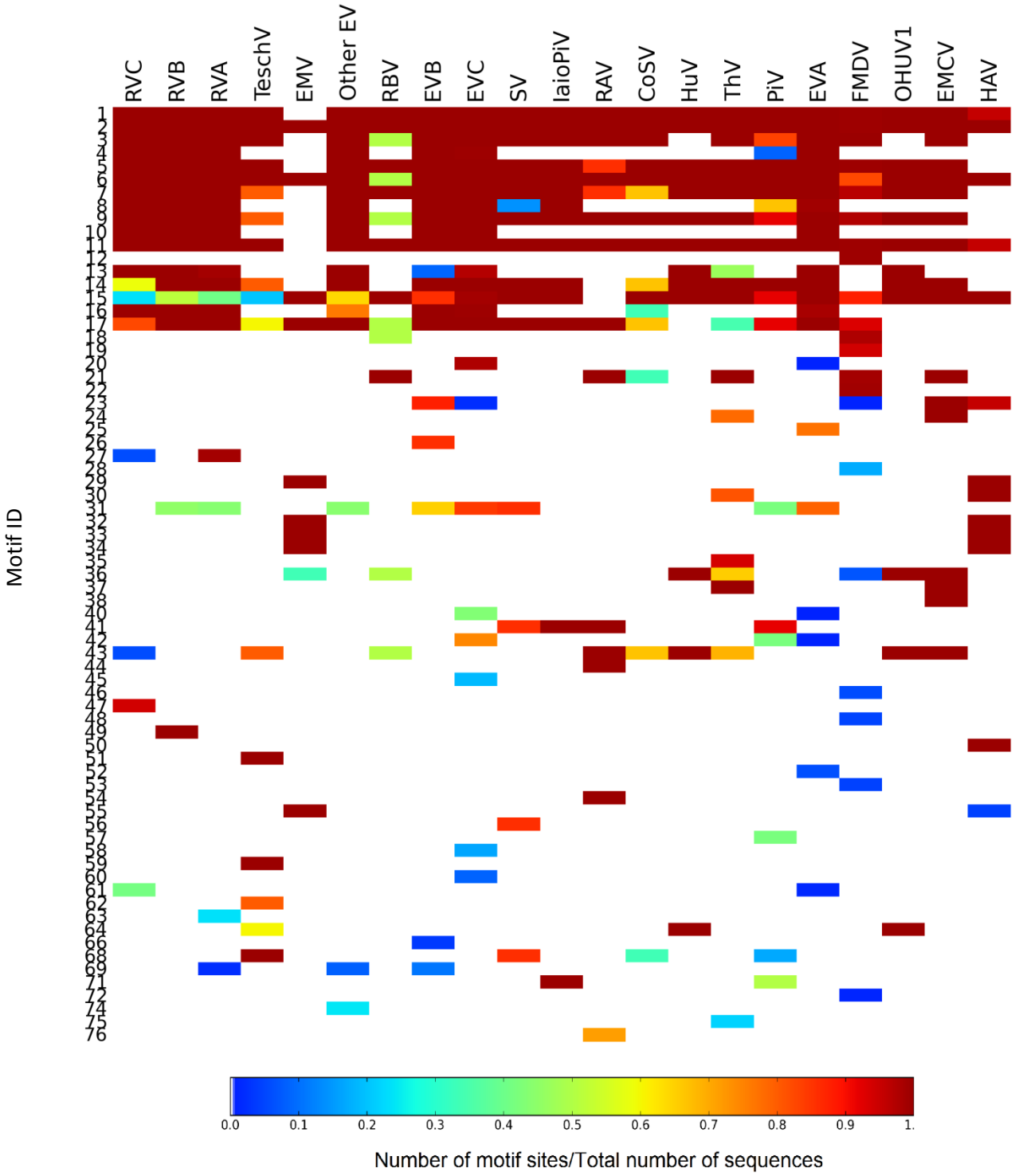


Figure 3.6. Heatmap of VP2 motif conservation across picornavirus species. All available picornavirus VP2 protein sequences were subjected to MEME for motif discovery. The conservation of each motif was calculated with respect to sub-groups comprising of sequences from the same virus species. The Heatmap was generated using the Matplotlib and Python scripting.

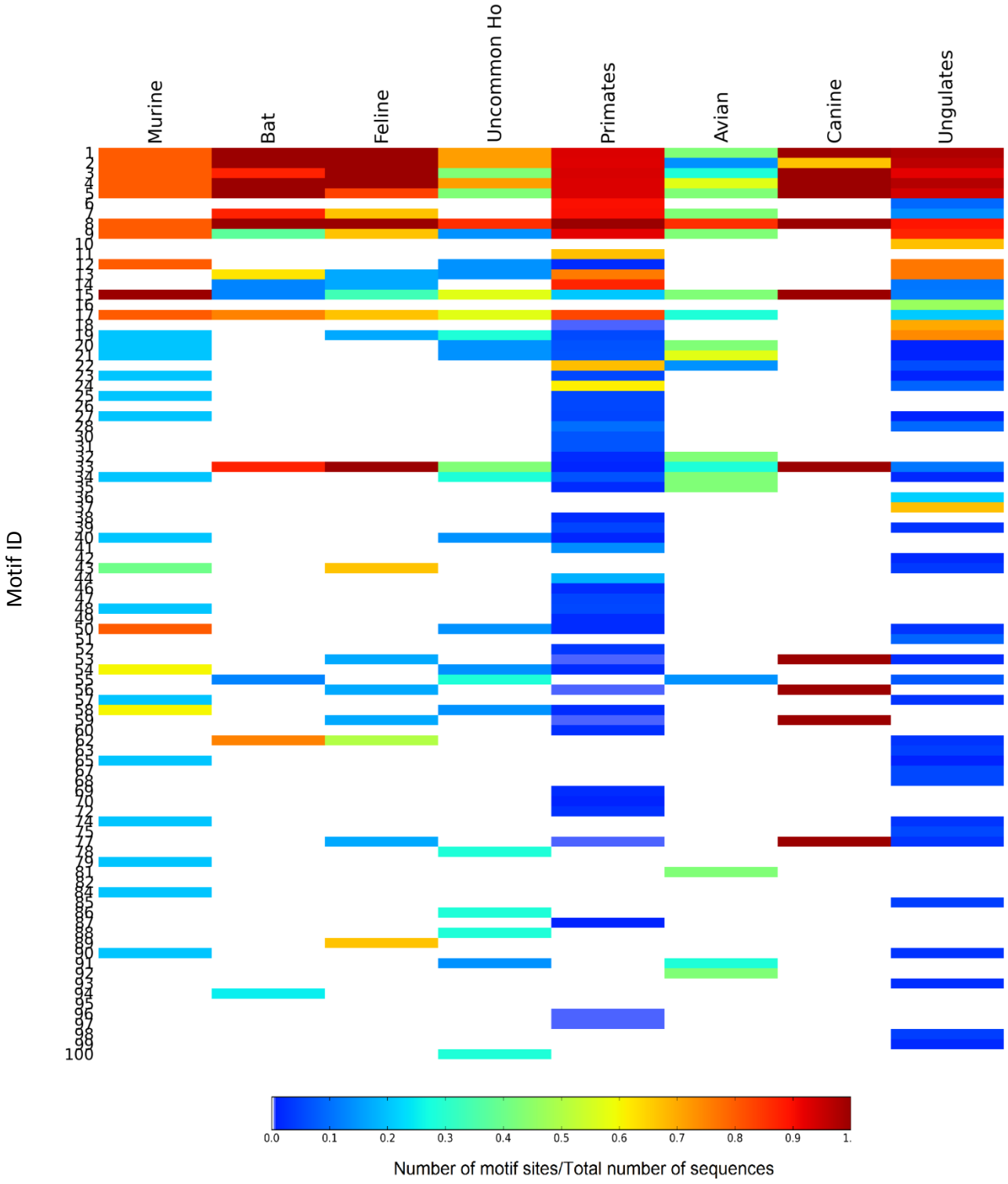


Figure 3.7. Heatmap of VP2 motif conservation across host species of respective picornaviruses. All available picornavirus VP2 protein sequences were subjected to MEME for motif discovery. The conservation of each motif was calculated with respect to sub-groups comprising of sequences from virus of the same species. The Heatmap was generated using the Matplotlib and Python scripting

3.3.1.3 Motif analysis of the VP3 dataset

MEME analysis was performed on a total of 998 VP3 amino acid sequences, which were representative of 38 individual picornavirus species isolated from eight major groups of host species. MEME identified a total of 100 significant motifs (E-value<=0.05) ranging in width from 3 to 20 amino acids. Following subjection to MAST analysis, unsuitable motifs 29, 45, 61, 64, 66, 71, 73, 76, 80 and 83 were excluded from the results set. Conservation of the remaining 90 motifs was determined respectively across 33 viral sub-groups and eight host sub-groups (Figures 3.8 and 3.9) as well as across individual sequences within each sub-group (Appendix 5.3). It was observed that motif conservation was virus dependent, with no obvious correlation between different viruses which infect the same host species. Figure 3.9 indicates a bias conservation of motifs across the feline, bat and canine species as these sub-groups only contained sequences corresponding to bat picornavirus, feline picornavirus and AV respectively. Therefore the apparent conservation is a result of conservation across these individual viral sub-groups. Furthermore the conservation of motifs 1-5 across primate species was due to the large number of EV and RV sequences. This was evident by an analysis of Figure 3.8 which indicates the complete conservation of these motifs across the enterovirus genus. Consequently motif conservation was analysed comprehensively with respect to individual virus species.

The conservation analysis was performed with respect to 33 viral sub-groups. Due to the limited number of sequences available the EV-D, EV-F, EV-G, EV-H, EV-I and the porcine enteroviruses were collectively grouped as "Other EV". It was observed that of the 90 motifs analysed, no motif was completely conserved across the virus family. However, motifs 1-5, 8-9, 15 and 17 had significantly higher conservation. Motif 1 had highest conservation with a total of 925 sites and was significantly conserved across 27 of the 33 viral sub-groups. The motif was identified in sequences from Salivirus, EV serotypes, HuV, FMDV, RV serotypes, EMCV, TV, LV, PaV, TeschV, AV, RBV, RAV, SaKV, RfV, ThV, SV and bat, feline and pigeon picornaviruses. As these viruses infect a wide variety of different hosts including humans, simians, avian species, ungulate species, canine, feline, bat, murine and tortoise there was no explicit correlation between motif conservation and host species observed. Similarly motifs 2, 3, 4, 5 and 8 were all present in over 20 of the 33 virus sub-groups with no specificity to host species. Motifs 9, 15 and 17 were identified in sequences in a total of 16, 18, and 20 sub-groups respectively.

Motif conservation was also comparatively analysed with respect to the phylogenetic clustering determined in the previous chapter. The most obvious correlation is the complete conservation of motifs 1-9 throughout all EV and RV sequences. However it must be noted that motif 6 was the only motif which was unique to sequences of the enterovirus genus. The conservation of motifs 7 and 14 directly correlated to the close evolutionary relationship between the enteroviruses, sapeloviruses and the unclassified bat, feline and pigeon picornaviruses. Motif 7 was completely conserved across and unique to all EV and RV serotypes as well as all sapelovirus sequences and all bat and feline picornavirus sequences (Appendix 5.3.1). Similarly motif 14 was highly conserved across the enteroviruses and the bat picornavirus sequences. The phylogenetic relationships observed within the out-group were supported by the conservation of motifs 20, 21, 23, 25, 27, 32, 53 and 79. More specifically motifs 20 and 21 were conserved across the EMV, PeV, LV, Sebokele virus, HAV, PaV, seal picornavirus and FMiPV. Motifs 23 and 27 were unique to PeV, LV, PaV and Sebokele virus sequences, thus supporting the sub-clustering of these viruses within the out-group. Furthermore motif 25 was unique to PeV, LV and Sebokele virus, while motifs 79 and 84 were unique to LV and sebokele virus sequences. Motif 32 was uniquely conserved across the monophyletic sequences of EMV and HAV, while motif 53 uniquely supported the close relationship between AV and SaKV sequences.

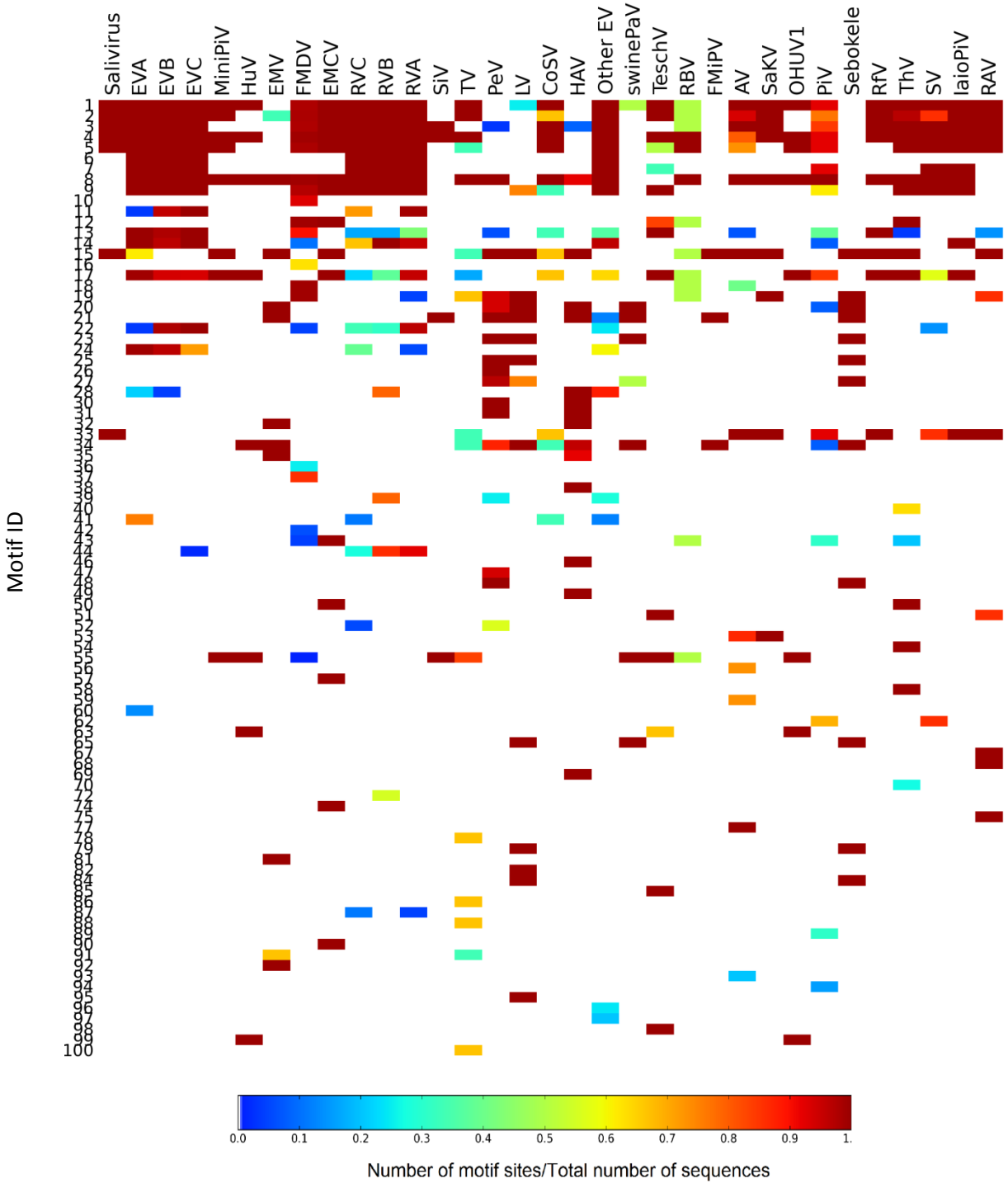


Figure 3.8. Heatmap of VP3 motif conservation across picornavirus species. All available picornavirus VP3 protein sequences were subjected to MEME for motif discovery. The conservation of each motif was calculated with respect to sub-groups comprising of sequences from the same virus species. The Heatmap was generated using the Matplotlib and Python scripting.

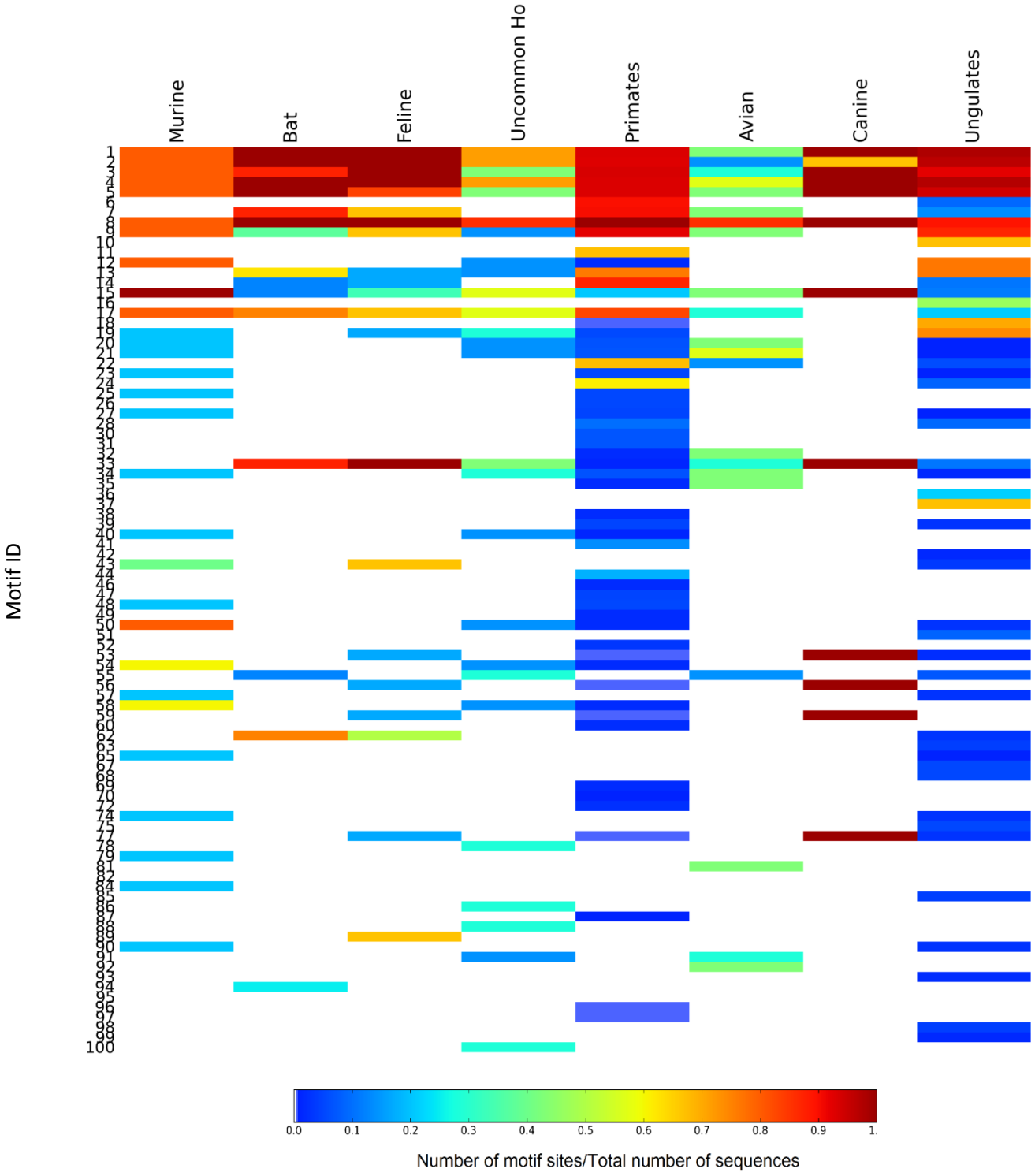


Figure 3.9. Heatmap of VP3 motif conservation across host species of respective picornaviruses. All available picornavirus VP3 protein sequences were subjected to MEME for motif discovery. The conservation of each motif was calculated with respect to sub-groups comprising of sequences from virus of the same species. The Heatmap was generated using the Matplotlib and Python scripting

3.3.1.4 Motif analysis of VP1 dataset

MEME analysis was performed on a total of 1289 VP1 amino acid sequences which were representative of 38 individual picornaviruses. A total of 100 significant motifs (E-value

<=0.05), ranging in width of 3 to 20 amino acids, were identified. The MEME results were subsequently subjected to a MAST analysis which identified motifs 67, 83 and 89 as unsuitable. Thus these motifs were excluded and conservation analysis was performed with respect to the remaining 97 motifs. For conservation analysis, the dataset was sub-divided according to 1) sequences representative of the same virus type and 2) sequences representative of viruses which infect the same host species. A total of 33 viral sub-groups and eight host sub-groups were respectively generated. Due to the limited number of sequences from EV-D, EV-F, EV-G, EV-H, EV-J and the porcine enteroviruses, these sequences were grouped as a single sub-group of Other EV. Likewise the bat picornavirus, feline picornavirus, pigeon picornavirus and seal picornavirus sequences were grouped as PiV. Sequences unique to viruses which infected uncommon hosts were also grouped as a single host sub-group. These hosts included tiger, tick, pale thrush, tortoise and the marine species: seal and fathead minnow. The conservation of each of the 97 motifs was determined across each of the viral and host sub-groups (Figures 3.10 and 3.11), as well as across each sequence within each individual sub-group (Appendix 5.3). It was observed that motif conservation was virus dependent, with no obvious correlation between different viruses which infect the same host species. Although Figure 3.11 indicates a degree of bias conservation across the feline, bat and canine species as these sub-groups only contained sequences corresponding to bat picornavirus, feline picornavirus and AV respectively. Furthermore the conservation of motifs 1-8 across primate species was due to the large number of EV and RV sequences, as Figure 3.10 indicates the complete conservation of these motifs across the enterovirus genus. Consequently motif conservation was analysed comprehensively with respect to individual virus species.

The conservation analysis across the 33 viral sub-groups (Figure 3.10) revealed that none of the 97 motifs were conserved across the *Picornaviridae* family. Motif 6, was observed to be highly significant (E-value = 1.1e-9340) and most conserved across the family, appearing in 22 of the 33 virus groups with a total of 1076 sites. More specifically the motif was highly conserved in representative sequences from Salivirus, EV serotypes, RV serotypes, bat, feline and pigeon picornaviruses, HuV, FMDV serotypes, EMCV serotypes, TeschV, SaKV and SV serotypes. The motif was less conserved amongst TV, LV, RBV and ThV sequences. Motif 28 (E-value = 8.8e-1541) was observed in 21 of the 33 virus groups with a total of 198 sites. Although this is significantly lower than the number of sites of motif 6, motif 28 did not appear in the EV or RV serotypes which were the two largest sequence datasets. Thus the

number of sites of motif 6 is positively skewed and the conservation across the number of virus sub-groups should rather be considered as an indication of significance. Motif 28 was found to be highly conserved across Salivirus, HuV, bat, feline and pigeon picornaviruses, EMV, EMCV, TV, PeV, HAV, CoSV FMiPV, RfV, ThV, SV and RAV serotypes, while less conserved but present in sequences from LV and RBV serotypes. Motif 1 (E-value 8.6e-9627) was also highly conserved across the *Picornaviridae* family. The motif appeared in 18 of the 33 viral sub-groups with a total of 1155 sites. However this motif was highly conserved across the EV and RV serotypes. It was also highly conserved across the FMDV, EMCV, TeschV, AV, HuV, ThV. Although less conserved across the PiV group, (Appendix 5.4.1) revealed 100% conservation across bat picornaviruses. Also highly significant (E-value = 1.9e-4801), was Motif 9. Although the motif only 357 sites it was highly conserved across Salivirus, HuV, FMDV, EMCV, SiV, RBV, RAV, SaKV and RfV virus sequences. The comparatively low number of sites can be explained by the absence of this motif from the EV and RV serotypes. Although the conservation analysis involved 97 significant motifs, only four were found to have substantial conservation across the *Picornaviridae* family. Furthermore, as illustrated by the heat map in Figure 3.10, the motif conservation was observed to be highly specific to individual viruses. In comparison to motif conservation across the other capsid proteins (Figures 3.4, 3.6, 3.8), motifs within VP1 sequences appeared to be substantially more virus specific with minimum conservation across the family. This result was to be expected as VP1 serves as both the antigenic determinant and the protagonist in host cell receptor binding. Therefore it is exposed to greater selectivity pressure.

Although conservation appeared to be highly specific to individual viruses, there was some evidence which correlated to the phylogenetic groupings discussed in the previous chapter. The most obvious correlation was the conservation of motifs 1-8, motif 10, motifs 14-15 and motif 17 which were highly conserved across the EV and RV serotypes. In this instance, motifs 2 and 3 were the most significant as they were unique to and highly conserved across all EV and RV serotypes. Motifs 4,7 and 8 were found have complete conservation across all EV and RV serotypes with high conservation across the bat, feline and pigeon picornaviruses and the SV serotypes (Appendix 5.4.1). This correlated directly to the phylogenetic super-group comprised of the enteroviruses, sapeloviruses and the bat, feline and pigeon picornaviruses, with motifs 4 and 7 being unique to this super-group. Motif 12 supports the clustering of the aphthoviruses FMDV and RAV with the erbovirus RBV strain ec11. This motif was unique to these virus types with complete conservation across all FMDV and RAV

sequences (Appendix 5.4.1). Furthermore the motif was absent from RBV strain p143 which clustered distantly with the CoSV sequences (Appendix 5.4.1). This clustering distinction was also supported by complete conservation of motif 28 amongst the CoSV sequences and RBV p143 but its absences from RBV ec11. The discrepancy observed between the monophyletic relationships within the capsid protein sequences of the CoSV sequences was supported by the conservation pattern of motifs 8 and 35. Motif 8 was also observed in CoSV-D1, while motif 35 was only observed in CoSV-B1 and CoSV-E1 (Appendix 5.4.1). This correlated directly to the unique monophyletic relationships observed for the VP1 sequences where CoSV-B1 and CoSV-E1 were found to be monophyletic with paraphyletic grouping to CoSV-D1. The clustering of the cardioviruses ThV and EMCV were supported by motifs 44 and 45 which were completely conserved and unique to all cardiovirus sequences. Moreover motif 47 was completely conserved amongst all cardiovirus sequences, but was also observed with significantly low conservation (<10%) in SV and RVB sequences. The clustering of PeV, LV, Sebokele and PaV was supported by motifs 48, 50 and 85, all of which are completely conserved and unique to these virus sequences. The sub-clustering of PeV and LV serotypes was further supported by the complete conservation and uniqueness of motif 49. The monophyletic relationship between HAV and EMV was uniquely supported by complete conservation of motifs 38 and 42. While motifs 28, 31, 35, 39 and 47 were common to both virus groups they were not unique to this cluster.

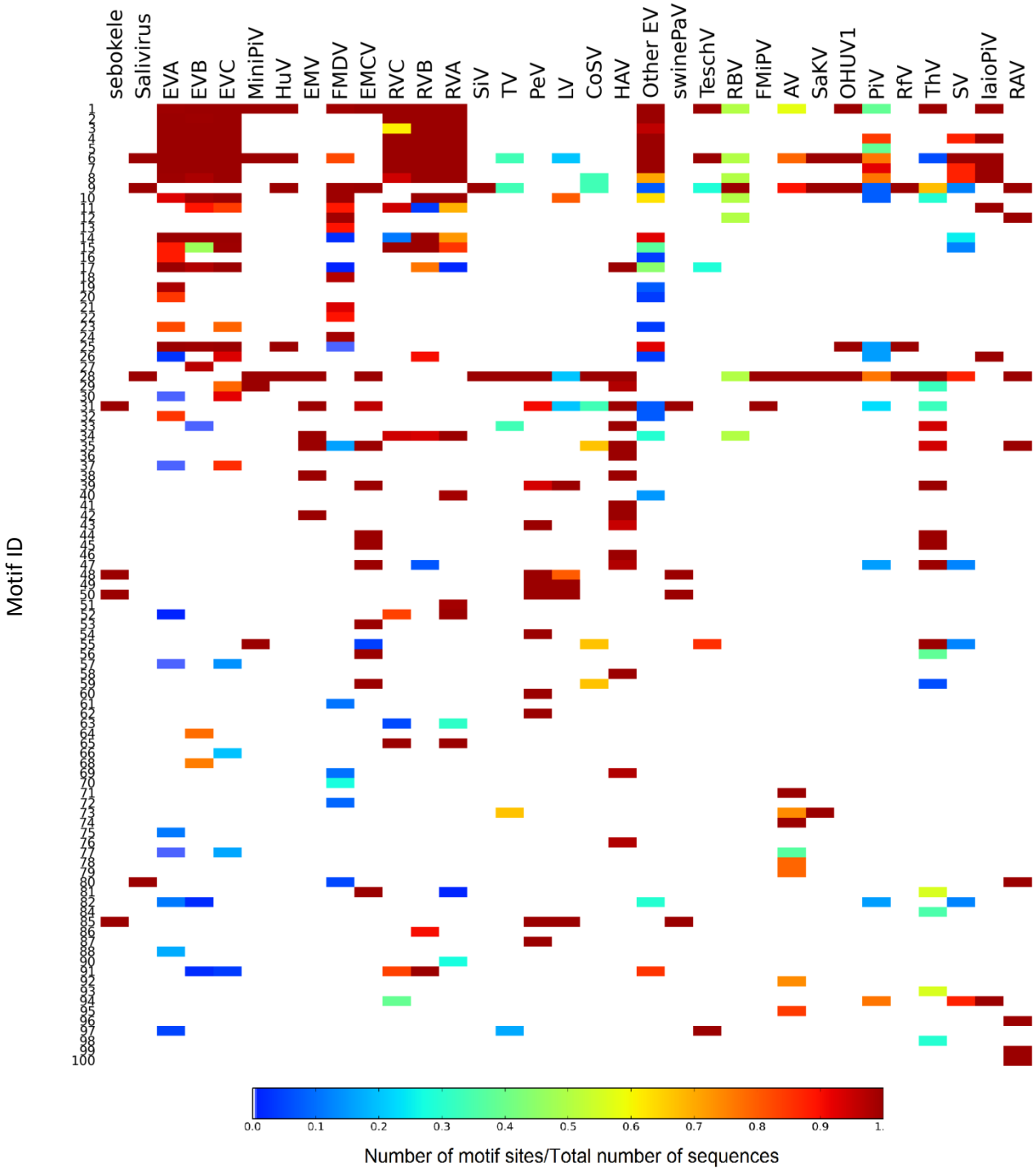


Figure 3.10. Heatmap of VP1 motif conservation across picornavirus species. All available picornavirus VP1 protein sequences were subjected to MEME for motif discovery. The conservation of each motif was calculated with respect to sub-groups comprising of sequences from the same virus species. The Heatmap was generated using the Matplotlib and Python scripting.

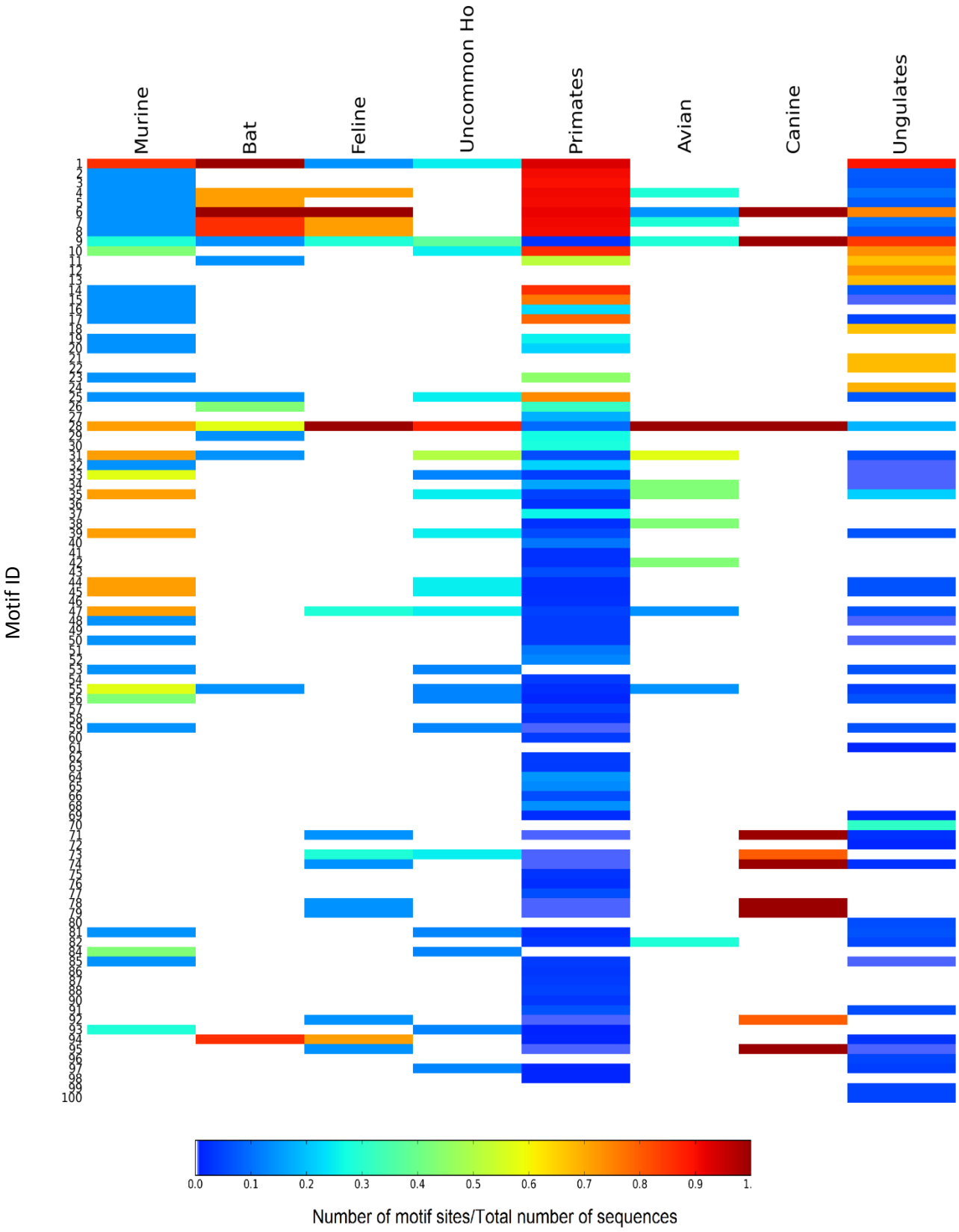
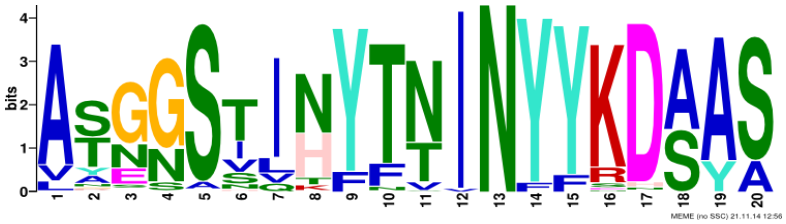
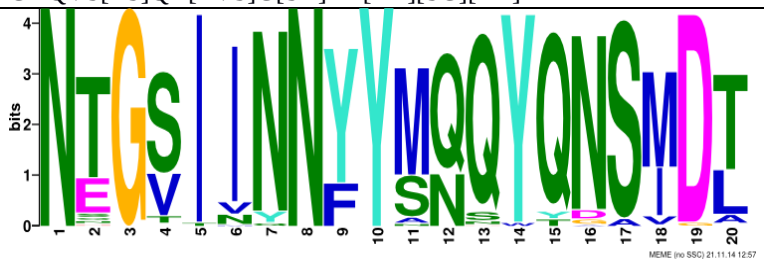
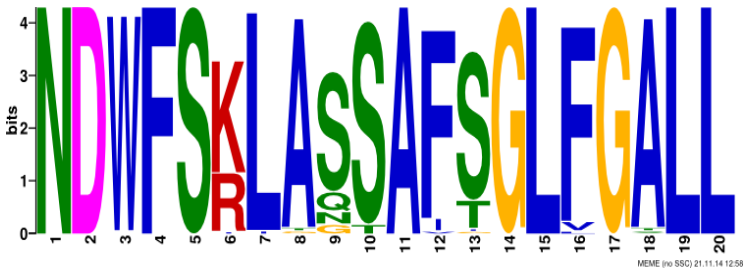


Figure 3.11. Heatmap of VP1 motif conservation across host species of respective picornaviruses. All available picornavirus VP1 protein sequences were subjected to MEME for motif discovery. The conservation of each motif was calculated with respect to sub-groups comprising of sequences from virus of the same species. The Heatmap was generated using the Matplotlib and Python scripting

Table 3.3a. Conserved motifs in picornavirus VP4 proteins. The IDs, logos, corresponding regular expressions and E-values, as depicted by MEME, of motifs which were identified to have significant conservation across the relative structural proteins of picornaviruses are presented. The motif logo depicts the consensus amino acid sequence as of a stack of letters at each position. The relative sizes of the letters indicate the respective frequencies in the sequences. The letters are colored by the physicochemical properties of the relative amino acids: Red) positively charged. Blue) hydrophobic. Green) Polar, non-charged, non-aliphatic residues. Magenta) acidic. Yellow) proline. Pink) histidine. Orange) glycine. Turquoise) tyrosine.

ID	Logo and Regular Expression	E-value
1	 [RK]QDF[ST]QDPSKFTEPVKD[VI][ML]I	1.1e-4930
2	 A[ST][GN][GN]STI[NH]YT[NT]INYYKD[AS][AY][SA]	5.4e-4059
3	 GAQVS[TS]QK[TVS]G[SA]HE[TN][SG][NL]	4.3e-3304
4	 N[TE]G[SV]IINN[YF]Y[MS][QN]QYQNS[MI]D[TL]	5.1e-1454
5	 NDWFS[KR]LASSAFSGLFGALL	9.3e-1086

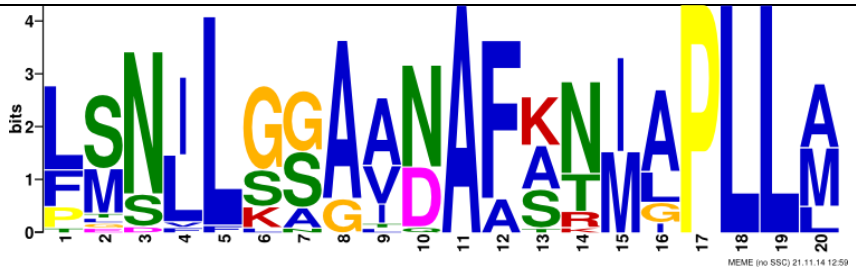
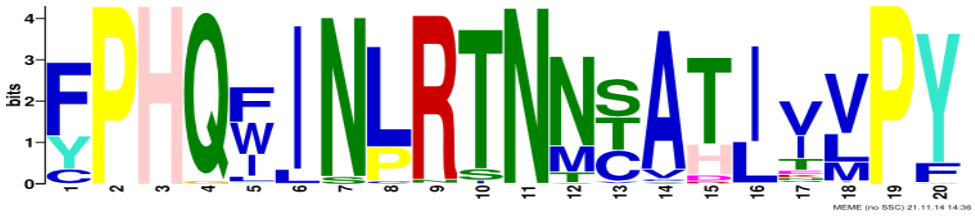
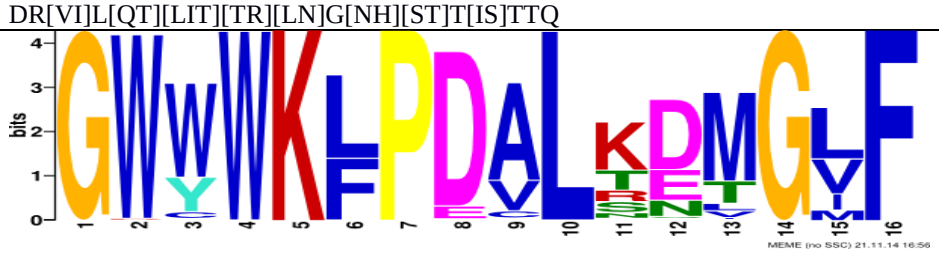
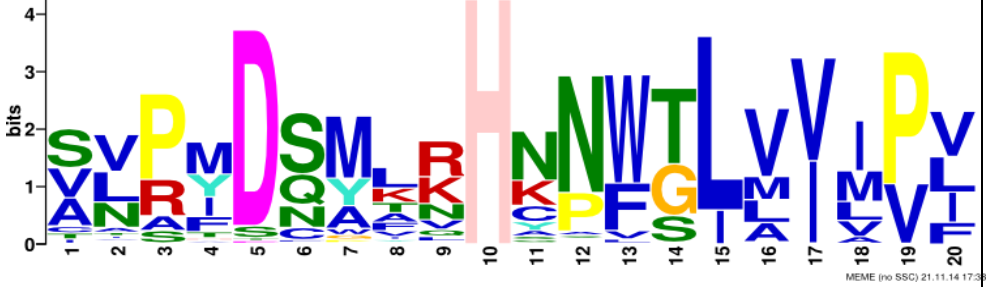
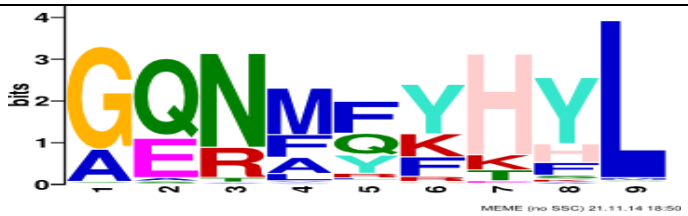
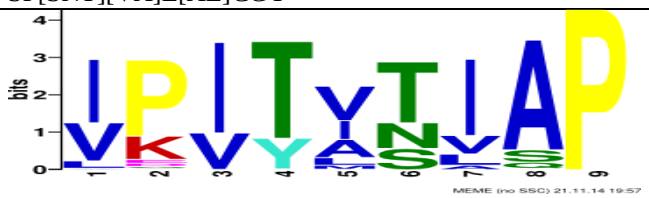
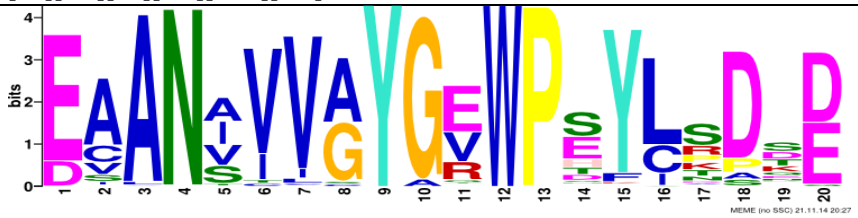
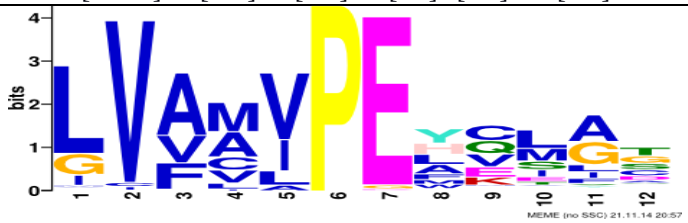
6	 GDNAISGGSNEGSTDTTSTH	1.9e-951
8	 [LF][SM]N[IL]L[GS][GS]A[AV][ND]AF[KAS][NT][IM][AL]PLL[AM]	2.4e-243
10	 [TNS]SSDK[NS]NS[QS]S[SE]G	4.7e-080

Table 3.3b. Conserved motifs in picornavirus VP2 proteins. The IDs, logos, corresponding regular expressions and E-values, as depicted by MEME, of motifs which were identified to have significant conservation across the relative structural proteins of picornaviruses are presented. The motif logo depicts the consensus amino acid sequence as of a stack of letters at each position. The relative sizes of the letters indicate the respective frequencies in the sequences. The letters are colored by the physicochemical properties of the relative amino acids: Red) positively charged. Blue) hydrophobic. Green) Polar, non-charged, non-aliphatic residues. Magenta) acidic. Yellow) proline. Pink) histidine. Orange) glycine. Turquoise) tyrosine.

ID	Logo and Regular Expression	E-value
1	 [FY]PHQ[FWI]IN[LP]RTN[NM][STC]A[TH][IL][VI][VL]PY	5.6e-16217
2	 R[SN]G[YW][TD][VI][HE]VQ[CA][NV][AG][SN][KQ]F[HN][QG]G[CA]L	1.1e-15277
3	 DR[VI]L[QT][LIT][TR][LN]G[NH][ST]T[IS]TTQ	9.3e-10373
4	 GW[WY]WK[LF]PD[AV]L[KT][DE]MG[LV]F	5.6e-9124
5	 [SVA][VLN][PR][MY]D[SQ][MY]L[RK]H[NK][NP][WF][TG]LV[VI][IM][PV][VLI]	5.6e-91246.3e-10804

6	 ATA[VE]D[KQ]P[TS][QE]PD[VT][SA][GV][CN]RFYTL	1.3e-10811
7	 [GA][QE][NR][MF][FQY][YK]HYL	2.3e-4672
8	 SP[SNT][VA]E[AE]CGY	4.2e-4281
9	 [IV][PK][IV][TY][VIA][TN]IAP	1.2e-4333
10	 EAAN[AIVS]VV[AG]YG[EV]WP[SE]Y[LC]SDS[DE]	1.3e-8976
11	 LV[AVF][MA][VI]PEY[CQ][LM][AG]	3.0e-5031
14	 E[FY][NS]GLR[NQ]A[TV]	8.6e-2945

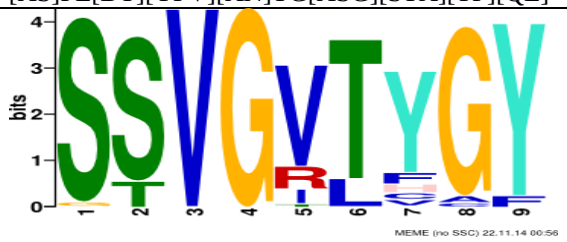
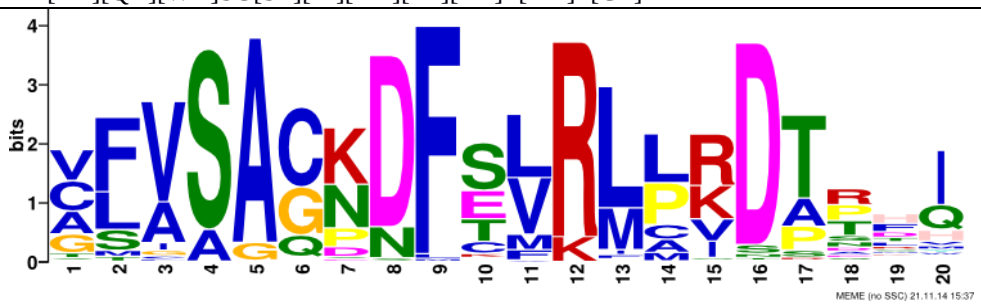
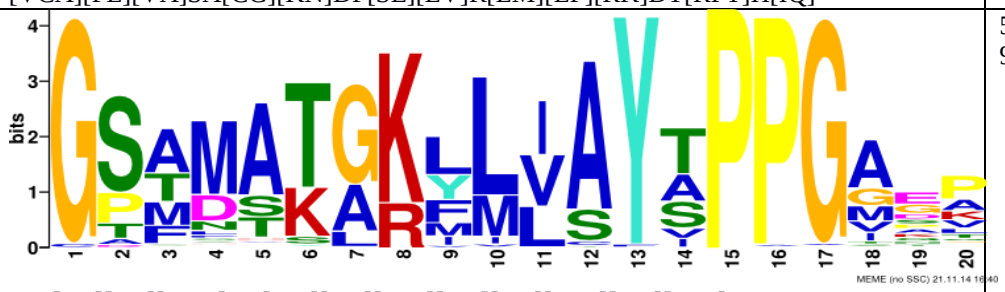
15	 [AS]PL[DT][YFV][AN]TG[ASG][STA][TP][QE]	2.6e-2692
21	 SSVGVTYGY	1.2e-1483
22	 [KT]RE[KL]YQLTL	2.4e-1253

Table 3.3c. Conserved motifs in picornavirus VP3 proteins. The IDs, logos, corresponding regular expressions and E-values, as depicted by MEME, of motifs which were identified to have significant conservation across the relative structural proteins of picornaviruses are presented. The motif logo depicts the consensus amino acid sequence as of a stack of letters at each position. The relative sizes of the letters indicate the respective frequencies in the sequences. The letters are colored by the physicochemical properties of the relative amino acids: Red) positively charged. Blue) hydrophobic. Green) Polar, non-charged, non-aliphatic residues. Magenta) acidic. Yellow) proline. Pink) histidine. Orange) glycine. Turquoise) tyrosine.

ID	Logo and Regular Expression	E-value
1	 [RP][KE][ED]A[MA][LH][GC][TI]H[VA][IVE]WD[VT]GL[QN]S[SK][CF]	7.0e-13961
2	 YY[TA][QH][WY]SG[ST][LI][KN][LF][TH]F[ML]F[CT]	7.6e-10372
3	 [VCA][FL][VA]SA[CG][KN]DF[SE][LV]R[LM][LP][RK]DT[RPT]H[IQ]	1.0e-10200
4	 IPG[ER][VF][TKR]NL[LM][ED][LVI][AC][EQ][VIA][DC][TSP]	5.1e-9091

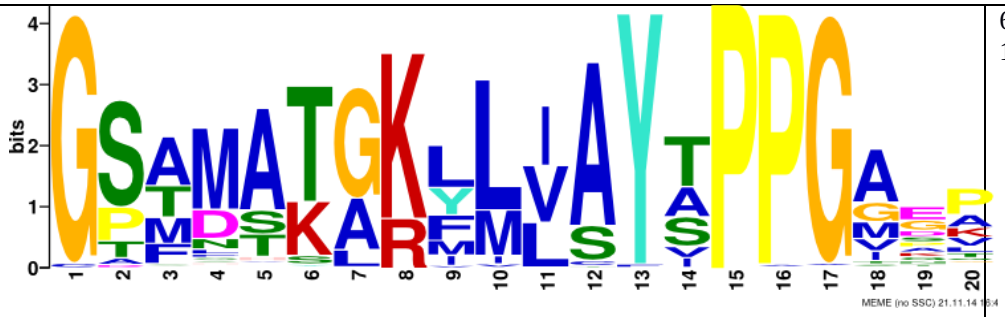
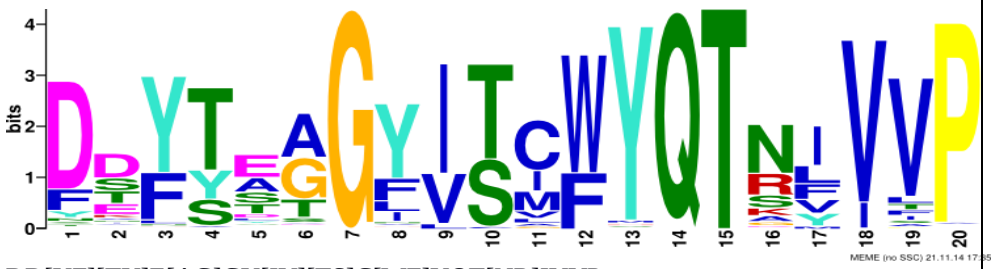
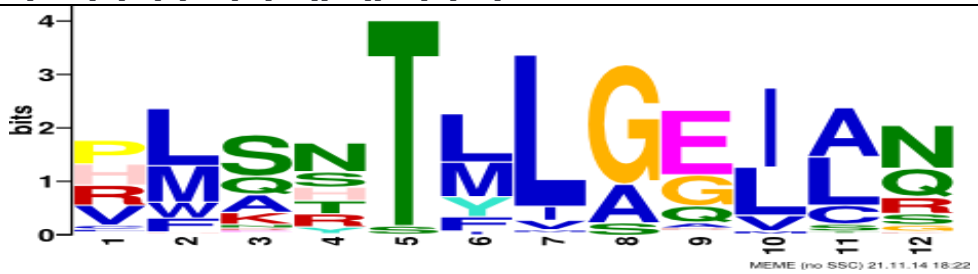
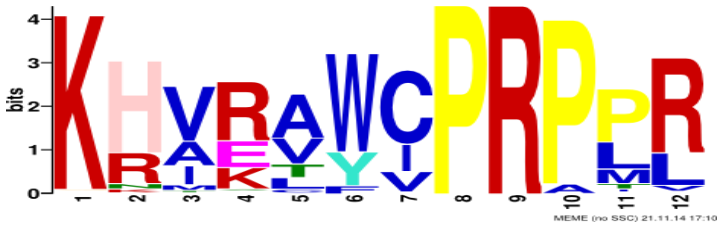
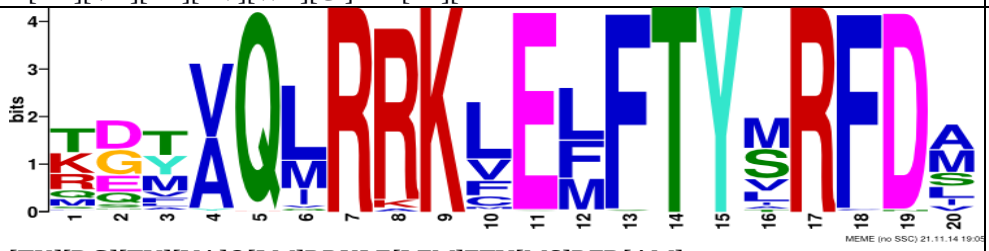
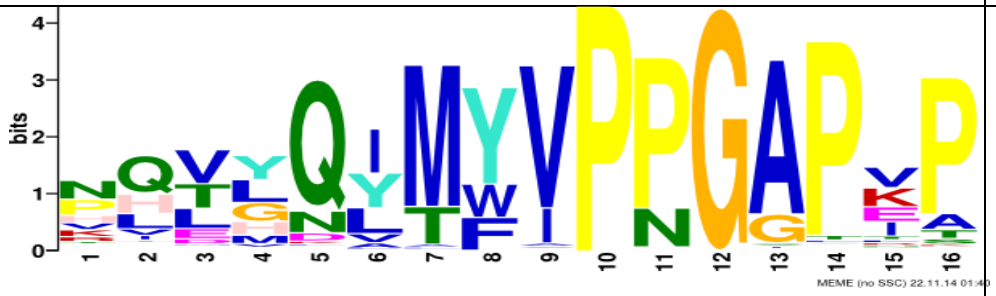
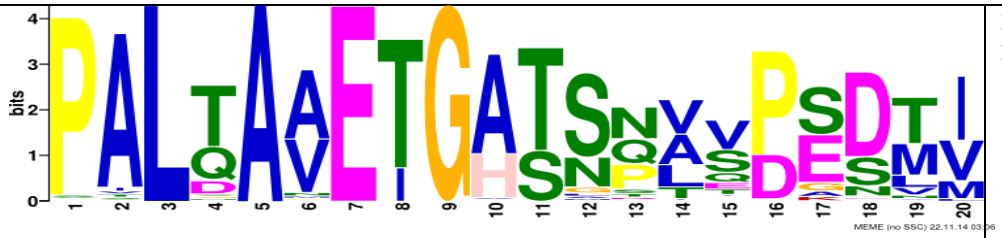
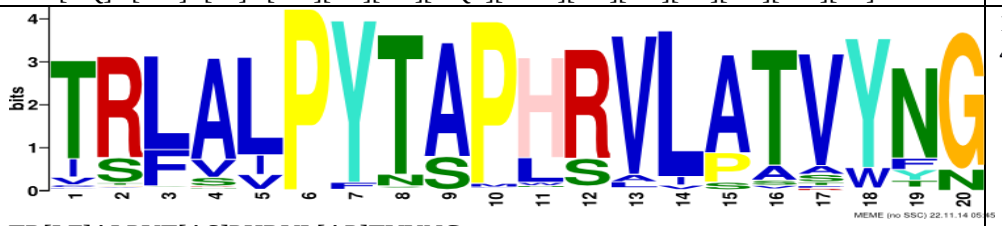
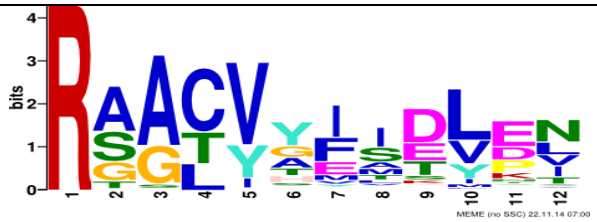
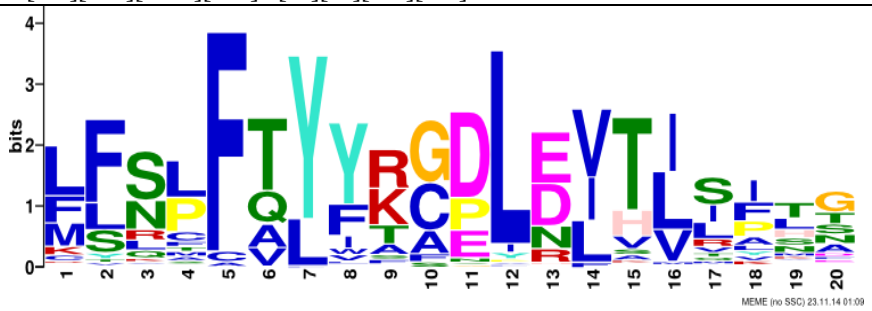
5	 G[SP][AT][MD]A[TK][GA][KR][LYF][LM][IVL]AY[TAS]PPGAEP	6.6e-10532
6	 TPGSNQ[FY]LT[TS]D[DN]FQSP[CS][AI][LM]P	3.3e-8719
7	 DD[YF][TY]E[AG]GY[IV][TS]C[WF]YQT[NR]IVVP	7.6e-7971
8	 T[LFM]V[VI]P[WY]IS[AN][TA][HD]Y[RA]YTA	8.2e-7852
9	 [PHRV][LM]SNT[LM]L[GA][EG][IL][AL][NQ]	1.4e-4401

Table 3.3d. Conserved motifs in picornavirus VP1 proteins. The IDs, logos, corresponding regular expressions and E-values, as depicted by MEME, of motifs which were identified to have significant conservation across the relative structural proteins of picornaviruses are presented. The motif logo depicts the consensus amino acid sequence as of a stack of letters at each position. The relative sizes of the letters indicate the respective frequencies in the sequences. The letters are colored by the physicochemical properties of the relative amino acids: Red) positively charged. Blue) hydrophobic. Green) Polar, non-charged, non-aliphatic residues. Magenta) acidic. Yellow) proline. Pink) histidine. Orange) glycine. Turquoise) tyrosine.

ID	Logo and Regular Expression	E-value
1	 K[HR][VA][RE][AV][WY][CI]PRP[PL][RL	8.6e-9627
2	 [TK][DG][TY][VA]Q[LM]RRKLE[LFM]FTY[MS]RFD[AM]	2.5e-12191
4	 P[FY][MV][SG][IP][AG][NS]AY[SQ][HW]FYDG[YF]	9.2e-9804
5	 [EQ]TR[HC]V[IL][NQ]S[HR][ST][RT][SA]E[ST][TS][IVL]E[SN]F[FL]	8.5e-10140
6	 [NP][QH][VT][YL]Q[IYL][MT][YWF]VP[PN]GAP[VK]P	1.1e-9340

7	 PAL[TQ]A[AV]E[TI]G[AH][TS][SN][NQP][VAL][VS][PD][SE][DS][TM][IV]	1.4e-10005
9	 TR[LF]ALPYT[AS]PHRVL[AP]TVYNG	1.9e-4801
10	 R[AS][AG][CTL][VY]Y[IF][IS][DE][LV]EN	8.5e-4461
28	 [LF]F[SN][LP]F[TQ][YL][YF][RK][GC][DP]L[ED][VIL]T[ILV]S[IF]TG	8.8e-1541

The assessment of the conservation of each motif across the respective virus and host sub-groups allowed for the identification of motifs with significant conservation across the *Picornaviridae* family or with significant conservation across individual sub-groups. The results indicated that motifs had higher specificity to individual virus groups, with no obvious conservation across proteins which were isolated from different viruses which infect the same host species. Therefore further analysis was performed with respect to motifs which were completely conserved across the respective structural proteins of individual virus groups which contained a significant number of sequences. These virus groups included: EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV. Although the RV-C groups also contained a significant number of sequences, they were excluded from further analysis as no representative crystal structure was available. Moreover, the appearance of each motif in

different virus groups was also considered. This was to exclude motifs which were completely unique to a single virus type, as these motifs were expected to be involved in specific host-cell receptor binding or antigenicity rather than capsid subunit interactions. The Tables 3.3.a- 3.3.d lists all motifs which were selected for additional analysis. The motifs were analysed with respect to the representative virus groups in which they were sited (Table 3.4). The further analysis involved the mapping of each of these motifs to respective crystal structures which were representative of the seven virus sub-groups (Table 3.5). The, the crystal structures were subject to the Protein Interaction Calculator (PIC) web-server for the prediction of amino acids involved in inter-protein interactions. This allowed for the identification of motifs which were predicted to facilitate protein-protein interactions between the capsid subunits of the respective viruses. Subsequent to this, these motifs were subjected to an analysis of residue conservation across all sequences of their respective virus groups. Moreover the total number of predicted interactions for each motif specific residue was calculated with respect to the specific structural sequence of each representative virus.

Table 3.4. Subunit specific motifs selected for structural analysis. Further analysis was performed with respect to virus sub-groups which contained a significant number of protein sequences. Individual motif analysis was pertained to representative viruses in which they were sited.

Motif ID	Sited Representative Viruses						
Motifs in VP4 Proteins							
1	*EV-A*	*EV-B*	*EV-C*	*RV-A*	*RV-B*		
2	EV-A	EV-B	EV-C	RV-A	RV-B		
3	EV-A	EV-B	EV-C	RV-A	RV-B		
4	FMDV			ThV			
5	FMDV						
6	FMDV						
8	ThV						
10	ThV						
Motifs in VP2 Proteins							
1	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
2	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
3	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
4	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
5	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
6	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
7	EV-A	EV-B	EV-C	RV-A	RV-B	ThV	
8	EV-A	EV-B	EV-C	RV-A	RV-B		
9	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
10	EV-A	EV-B	EV-C	RV-A	RV-B		
11	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
14	ThV						
15	ThV						
21	FMDV			ThV			
22	FMDV						
Motifs in VP3 Proteins							
1	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
2	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
3	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
4	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
5	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
6	EV-A	EV-B	EV-C	RV-A	RV-B		
7	EV-A	EV-B	EV-C	RV-A	RV-B		
8	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
9	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
Motifs in VP1 Proteins							
1	EV-A	EV-B	EV-C	RV-A	RV-B	FMDV	ThV
2	EV-A	EV-B	EV-C	RV-A	RV-B		
3	EV-A	EV-B	EV-C	RV-A	RV-B		
4	EV-A	EV-B	EV-C	RV-A	RV-B		
5	EV-A	EV-B	EV-C	RV-A	RV-B		
6	EV-A	EV-B	EV-C	RV-A	RV-B		
7	EV-A	EV-B	EV-C	RV-A	RV-B		
9	FMDV						
10	FMDV						
28	ThV						

3.3.2. Crystal structure selection

Representative crystal structures of virus capsid were obtained for the EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV. Structures with the highest quality, as determined by lowest resolution and R-value, were selected and downloaded from the Protein Databank (PDB). The following representative structures were selected: Coxsackievirus A24 from the EV-A group, Coxsackievirus B3 from the EV-B group, Poliovirus Mahoney from the EV-C group, Rhinovirus A16 from the RV-A group, Rhinovirus B3 from the RV-B group, Theilovirus DA from the ThV group and FMDV Strain A22-Iraq from the FDMV group. The respective PDB IDs, resolutions and R-values are shown in Table 3.5. Each of the PDB files a single copy of each VP1, VP2, VP3 and VP4 protein. Thus the structures were representative of a single protomer and not the complete viral capsids.

Table 3.5. Experimental details of the representative crystal structures. Individual crystal structures were selected to represent the picornavirus virus groups of: EV-A, EV-B, EV-C, RV-A, RV-B, ThV and FMDV. The PDB files were obtained from the Protein Databank

PDB ID	Virus Name	Strain	Classified Species	Resolution (A)	R-Value	Reference
4Q4W	Coxsackievirus	A24	EV-A	1.40	0.150	Zocher *et al*., 2014
4GB3	Coxsackievirus	B3	EV-B	2.74	0.330	Yoder *et al*., 2012
1HXS	Poliovirus	Mahoney	EV-C	2.20	0.268	Miller *et al*., 2001
1AYM	Rhinovirus	A16	RV-A	2.15	0.230	Hadfield *et al*., 1997
1RHI	Rhinovirus	B3	RV-B	3.00	0.284	Zhao *et al*., 1996
4IV1	Foot-Mouth-Disease-Virus	A22- Iraq	FMDV	2.10	0.190	Porta *et al*., 2013
1TME	Theilovirus	DA	ThV	2.80	0.300	Grant *et al*., 1992

3.3.3 Protein Interaction Calculator (PIC) Predictions

The PDB files of each of the representative virus protomers (Table 3.5) were individually submitted to the PIC webserver for the identification of specific amino acids predicted to participate in protein-protein interactions within a multichain protein structure. The default settings of the PIC webserver were implemented for the prediction of: 1) hydrophobic interactions within 5A, 2) main chain-main chain hydrogen bonds, 3) main chain-side chain hydrogen bonds, 4) side chain-side chain hydrogen bonds and 5) ionic interactions within 6A. The PIC results of each representative virus, presented as a text file in Appendix, was analysed by Python scripting which incorporated the MEME results of each structural protein dataset. The script mapped each interaction to the motifs identified in each structural protein of the respective representative structural sequence. Thus the script allowed for the identification of motifs which contained amino acids predicted to be interacting, with further

identification of particular motif-motif interactions between the subunits of each representative virus protomers (Table 3.6). Consequent to the substantial amount of data, the study focused on the analysis of motif-motif interactions, rather than the analysis of all motifs which contained interacting amino acids. Figure 3.12, depicts the motifs in each capsid subunit which were predicted to contain residues involved in motif-motif interactions within each of the representative PDB files. Motif 1 in the VP1 subunits, motifs 1, 5 and 11 in the VP2 subunits and motifs 1-3 in the VP3 subunits were predicted to contain interacting residues within all of the representative virus protomers. This may be highly significant as the PDB files were representative of different virus species, from three separate genera. The PDB files representative of the *Enterovirus* genus included only human viruses, while FMDV (genus: *Aphthovirus*) uniquely infects ungulate species and ThV DA (genus: *Cardiovirus*) infects murine species. Thus these motifs may serve as a drug target for a diverse range of picornaviruses, which vary in pathogenicity and host range specificity. Amongst these findings, other significant interactions were also observed. The VP1 motifs 2, 4, 5 and 7 were predicted to facilitate protein-protein interactions within the protomers of all the represented virus structures of the *Enterovirus* genus. Alternatively, motif 6 was predicted to be involved in subunit-subunit interactions throughout all the representative enteroviruses with the exception of EV-B. Motifs 9 and 10 were predicted to facilitate subunit interactions in FMDV (strain: A22-Iraq), while motif 28 was predicted to be interacting within the protomer of ThV (strain: DA). With respect to the VP2 subunits, motifs 7, 8, 9 and 14 were found to not be involved in interactions with identified motifs within the other capsid subunits, while motifs 6 and 22 and motif 15 were predicted to be interacting only within the FMDV and ThV protomer respectively. In contrast, motif 21 was found to facilitate subunit interactions within the representatives protomers of both FMDV and ThV. Motifs 2, 4 and 10 were predicted to be interacting in all enteroviruses, with additionally interaction of motif 2 in ThV. Motif 3 was only predicted facilitate interactions within the representative viruses of EV-A, EV-B and RV-B.

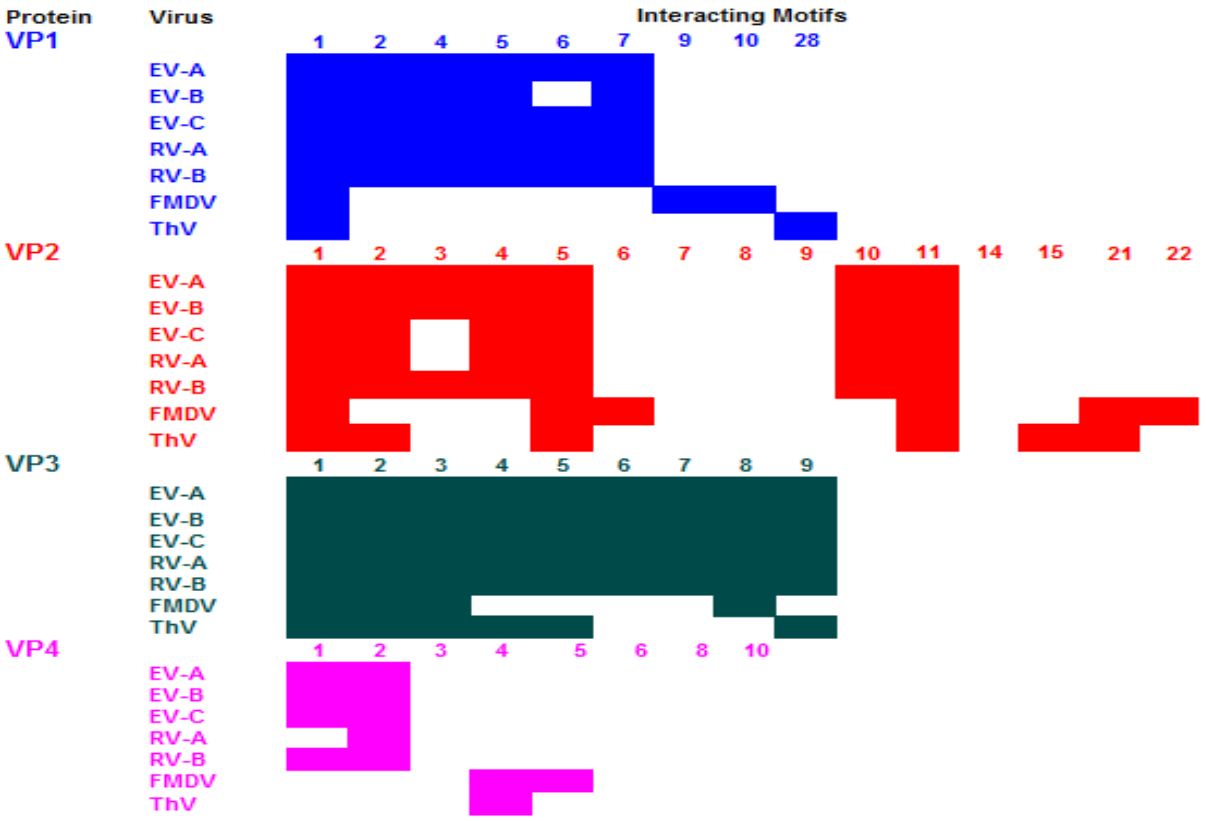
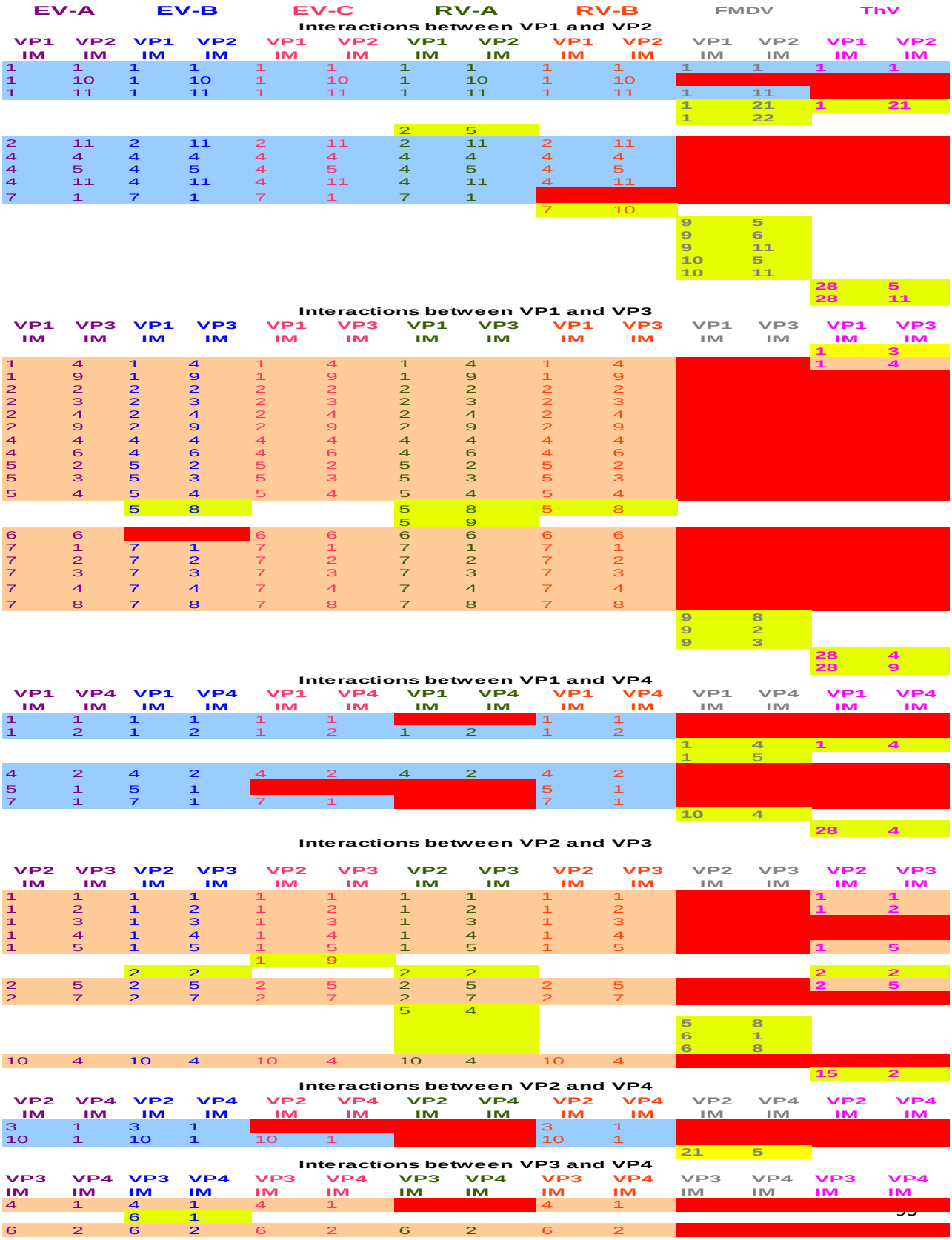


Figure 3.12. Predicted interacting motifs (IMs). Crystal structures of the viral capsid protomer of representative picornaviruses were submitted to the Protein Interaction Calculator (PIC) for the prediction of residues involved in protein-protein interactions. Predicting interacting residues were matched against residues of virus specific motif block diagrams for the identification of IM within the subunits of each representative virus. Blue) IM within subunitVP1, Red) IM within subunit VP2, Grey) IM within subunit VP3 and Pink) IM within subunit VP4.

The motifs 4-9 of the VP3 proteins were predicted to be interacting within all enteroviruses, with additional interaction of motifs 4, 5, and 9 in FMDV and motif 8 in ThV. Pertaining to the VP4 subunits, the motifs 3, 6, 8 and 10 were predicted to be non-interacting. Motif 2 was predicted to facilitate motif-motif interactions between the subunits of all representative enterovirus protomers, while motif 4 was predicted as interacting within the FMDV and ThV protomers. Motif-motif interactions involving motif 1 was limited to the enteroviruses with the exclusion of RV-A, while the interactions of motif 5 was limited to FMDV. In summary, the study identified seven motifs which were highly conserved and predicted to be interacting within the protomers of all seven representative sub-groups. Moreover five motifs were predicted to be interacting within the protomers of six of the seven representative viruses, while VP2.21 and VP4.4 were predicted to be interacting in both the FMDV and ThV viruses. Furthermore nine conserved motifs were predicted to facilitate subunit interactions across the protomers of the enteroviruses, with a total of four motifs and two motifs predicted

to be interacting uniquely within FMDV and ThV respectively. The conservation of interacting motifs across the closely related enteroviruses may further support the evolutionary mechanism of genetic recombination between closely related species and virus subtypes. This was further supported by the motifs which were predicted to be uniquely interactive within the FMDV and ThV viruses respectively. Although the results are based on an analysis of a single representative structure of each species, each of these motifs was identified to be completely conserved across all available strains of the respective viruses. Thus the predicted interactions within the protomers of these structures may be applicable to all strains of the respective representative viruses. For complete analysis of subunit-subunit interactions, the predicted interacting motifs were mapped to the relative chains of the respective PDB files and visualised in PyMOL (Figures 3.15 and 3.16). The results are discussed in the next section. Subsequent to this, a comprehensive analysis of the type of interactions between motif specific residues was performed (Section 3.3.5).

Table 3.6. Predicted subunit motif-subunit motif interactions in representative picornaviruses. Crystal structures of the viral capsid protomer of representative picornaviruses were submitted to the Protein Interaction Calculator (PIC) for the prediction of residues involved in protein-protein interactions. Predicting interacting residues were matched against residues of virus specific motif block diagrams for the identification of IM within the subunits of each representative virus. Pale Blue and Pale Orange highlighter depict conserved subunit motif-subunit motif interactions. Red highlighter depicts missing interactions. Yellow highlighter depicts additional predicted interactions



3.3.4 Structural Mapping of Interacting Motifs

Python scripting, with the incorporation of PyMOL commands, facilitated the mapping of the predicted interacting motifs to the respective PDB files. Captured images of the mapped protomers for each representative virus are presented in Figures 3.15 and 3.16. The interacting motifs are coloured by subunit with VP1 motifs depicted as the reds and browns, VP2 motifs as the yellows and greens, VP3 motifs as the purples and blues and VP4 motifs as the oranges. A standardised colour key was used for the motifs conserved across all the representative viruses, as well as the viruses of the *Enterovirus* genus (Figure 3.15). However discrepancies in the predictions of interacting motifs were observed in the FMDV and ThV viruses (Table 3.6) and thus there are variations within the colour key of the mapping of these structures (Figures 3.16 and 3.17). As previously stated, the analysis of the PIC results allowed for the identification of specific subunit motif-subunit motif interactions. As shown in Table 3.6, significant conservation of these specific interactions was observed, particularly with respect to viruses of the *Enterovirus* genus. For simplicity the motifs will be reference in the format of: *Subunit.Motif ID*. The results indicated that the assembly of the virus protomers may be facilitated through a network of multiple subunit motif-subunit motif interactions. The networks of conserved multiple interactions were mapped with respect to all seven representative viruses (Figure 3.13), the representative enteroviruses (Figure 3.14) and the FMDV and ThV representative viruses (Figure 3.17). As indicated by Figure 3.13 the interaction between VP1.1 and VP2.1 was completely conserved across all representative viruses, while the following interactions were conserved across the enteroviruses and ThV DA: VP1.1 with VP3.4; VP2.1 with VP3.1, VP3.2 and VP3.5; VP2.2 with VP3.1. The interaction between VP1.1 and VP2.11 was conserved across the enteroviruses and FMDV A22-Iraq. These observations were further supported by PyMOL results (Figures 3.15 and 3.16). As indicated VP1.1 (Red) was observed to be in close proximity to VP2.1 (Yellow) in all seven of the PDB files, while motifs VP1.1 (Red), VP3.4 (Marine), VP3.1 (Blue),VP3.2 (Lightblue) and VP3.5 (Deeppurple) were all observed to be in close proximity in the enterovirus PDB files and the ThV DA PDB file. Moreover the structural mapping of VP1.1 (Red) in proximity to VP2.11 (Limon) supported the interaction of these motifs within the enterovirus and FMDV PDB files. The conservation of these interactions provided evidence that the motifs may be critical for the assembly of the protomers of many picornaviruses. Furthermore VP1.1 was identified as the protagonist for interactions with subunits VP2 and VP3, while VP2.1 appeared as the protagonist for the interaction between VP2 and VP3.

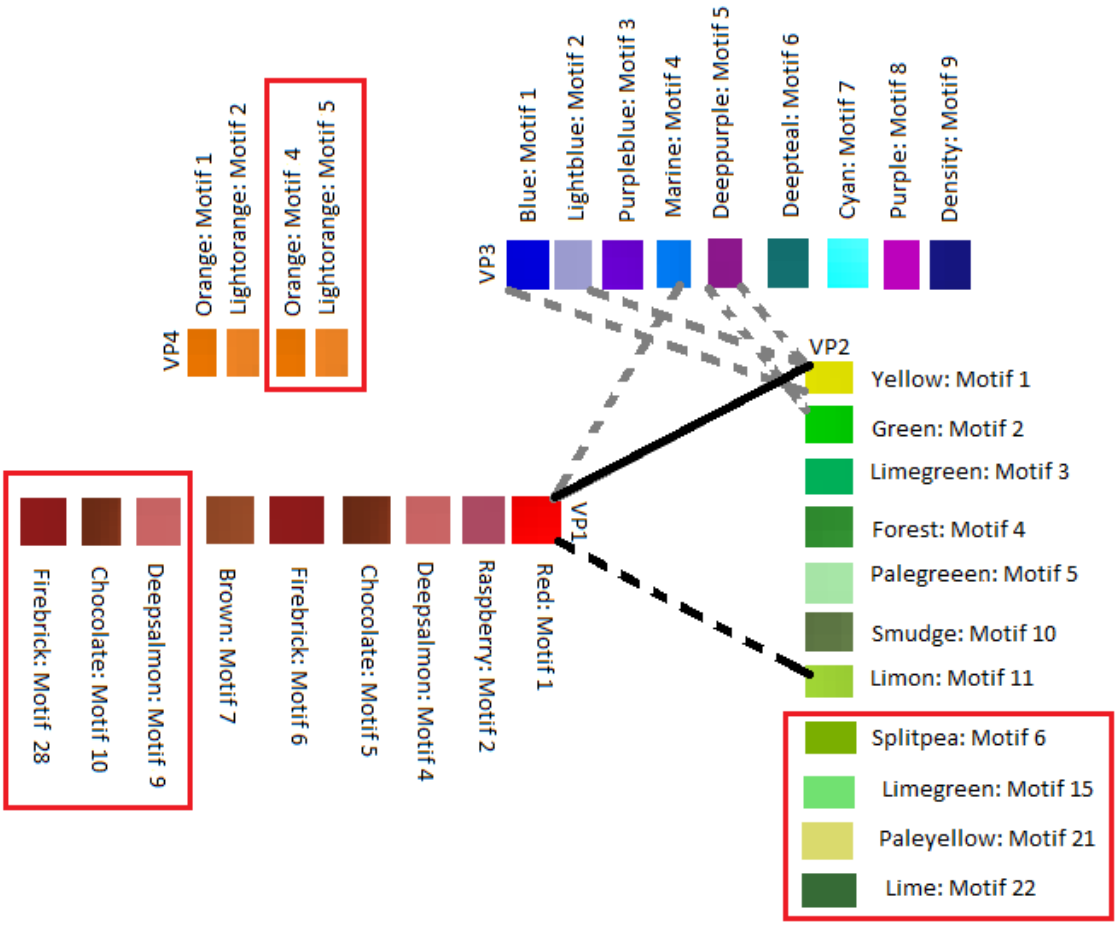


Figure 3.13. Network of conserved motif-motif interactions between subunits of picornavirus capsid protomers. Crystal structures of the viral capsid protomer of representative picornaviruses were submitted to the Protein Interaction Calculator (PIC) for the prediction of residues involved in protein-protein interactions. Predicting interacting residues were matched against residues of virus specific motif block diagrams for the identification of IM within the subunits of each representative virus. Representative viruses of the species: EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV were included. Solid Black Lines) Interactions conserved across all representative picornaviruses. Dotted Black Lines) Interactions conserved across all enterovirus and FMDV representative viruses. Dotted Grey Lines) Interactions conserved across all enterovirus and ThV representative viruses. Red Boxes) Alternative colour key used for the FMDV and ThV viruses

As previously stated significant conservation of particular interactions was observed amongst the subunits of the representative enteroviruses. These interactions were identified and mapped as a conserved network with respect to their representative structures (Figure 3.14). The results indicated that VP1.7 and VP1.4 may play a crucial role in the assembly of enteroviruses, as they were predicted to interact with a total of five and four VP3 motifs respectively. Furthermore VP1.7 was also predicted to interact with VP2 and additional interactions between VP1.4 and three of the VP2 motifs were predicted. The assembly of the subunits VP1 and VP2 was predicted to be further supported by interactions of VP2.11 with

multiple VP1 motifs. The results also indicated the fundamental interactions between VP2.1 with five motifs of the VP3 subunit. Pertinence to the VP3 subunits, the motifs 2-4 all were predicted to have multiple interactions with subunit VP1. The observed protagonist VP3.4 was predicted to interact with five motifs of the VP1 subunit. The structural visualisation of the motifs provided additional evidence that the subunits may be assembled through a network of multiple interactions. As shown in Figure 3.15, the interacting protagonists: VP1.1 (red), VP1.2 (raspberry), VP1.4 (deepsalmon), VP1.7 (brown), VP2.1 (yellow), VP2.2 (green), VP2.11 (limon), VP3.1 (blue), VP3.2 (lightblue), VP3.3 (purpleblue), VP3.4 (marine) and VP3.5 (deeppurple) were conserved, throughout the five representative enteroviruses, with close proximity to facilitate a multiple interaction network. The conservation of these specific subunit motif-subunit motif interactions across the representative enteroviruses may provide additional evidence that capsid proteins are evolving through genetic recombination between closely related species.

Although a significant proportion of the interactions were observed to be conserved across the representative enteroviruses, species-specific discrepancies were also observed. As shown in Table 3.6, the interaction between VP1.7 and VP2.1 was replaced by VP2.10 in the representative RV-B3, while the interactions of VP1.1 with VP4.1; VP1.5 with VP4.1; VP1.7 and VP4.1; VP2.3 with VP4.1; VP2.10 with VP4.1 and VP3.4 and VP4.1 were absent in the RV-A16 virus. Moreover the interactions of VP1.6 with VP3.6; VP1.5 with VP4.1 and VP2.3 with VP4.1 were absent in EV-C virus Polio Mahoney. Additionally interactions were observed between VP1.5 and VP3.8; VP2.2 and VP3.2; VP3.6 and VP4.1 in the EV-B virus (CBV-B3). Auxiliary interactions within the interface of RV-A16 were also predicted between VP1.5 and VP3.8; VP1.5 and VP3.9; VP2.2 and VP3.2; VP2.5 and VP3.4. The additional interactions were predicted to supplement the interactions between motifs which were also predicted to be involved in conserved interactions. Thus the findings provided further support of the importance of these motifs, rather than the identification of alternative interactions for protomer assembly.

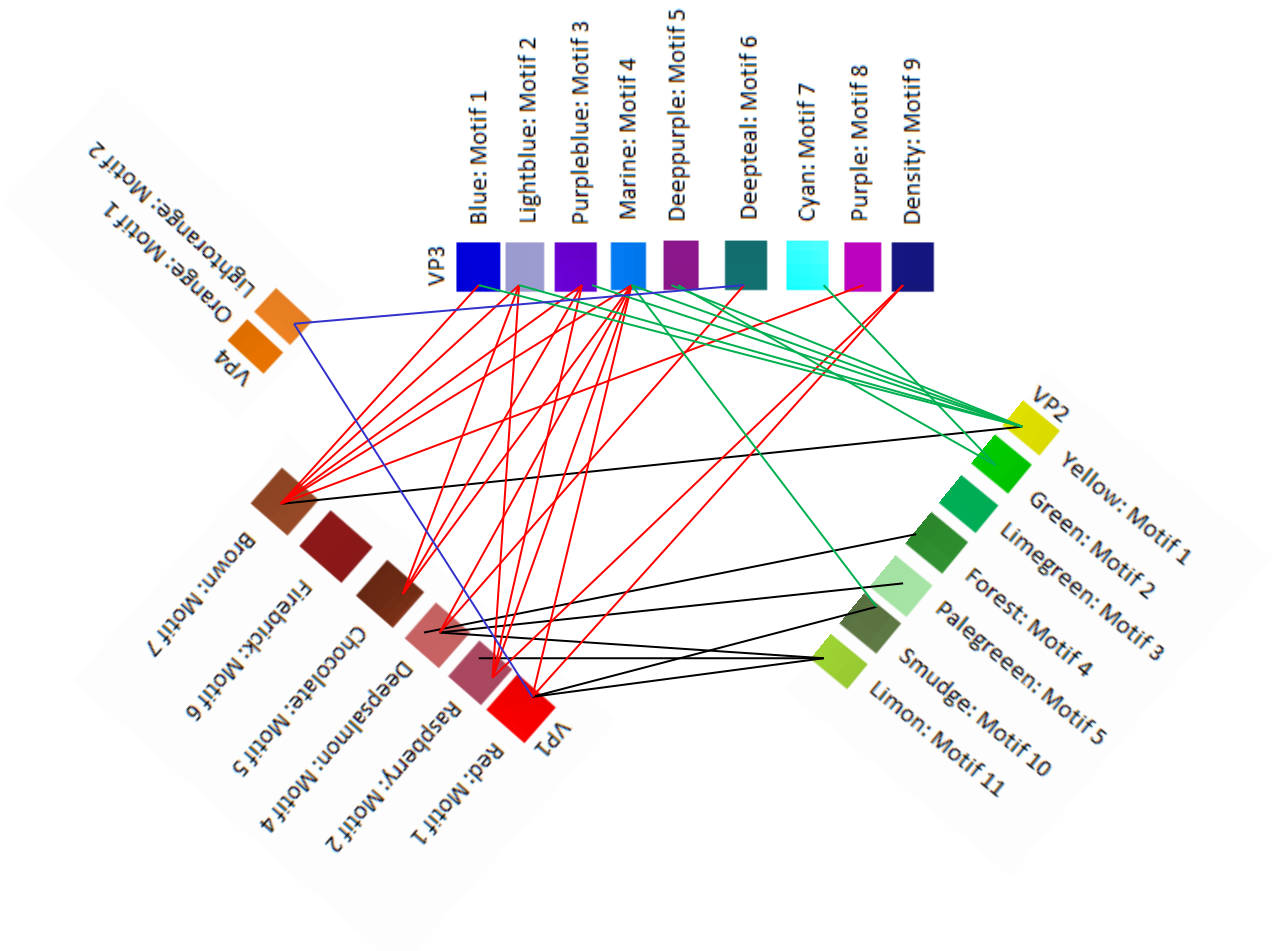


Figure 3.14. Network of conserved motif-motif interactions between subunits of representative enterovirus capsid protomers. Crystal structures of the viral capsid protomer of representative picornaviruses were submitted to the Protein Interaction Calculator (PIC) for the prediction of residues involved in protein-protein interactions. Predicting interacting residues were matched against residues of virus specific motif block diagrams for the identification of IM within the subunits of each representative virus. Representative viruses of the enterovirus species: EV-A, EV-B, EV-C, RV-A, RV-B, were included. Solid Black Lines) Conserved interactions between enterovirus VP1 and VP2 subunits. Solid Red Lines) Conserved interactions between enterovirus VP1 and VP3 subunits. Solid Green Lines) Conserved interactions between enterovirus VP2 and VP3 subunits. Solid Blue Lines) Conserved interactions between enterovirus VP4 and other subunits.

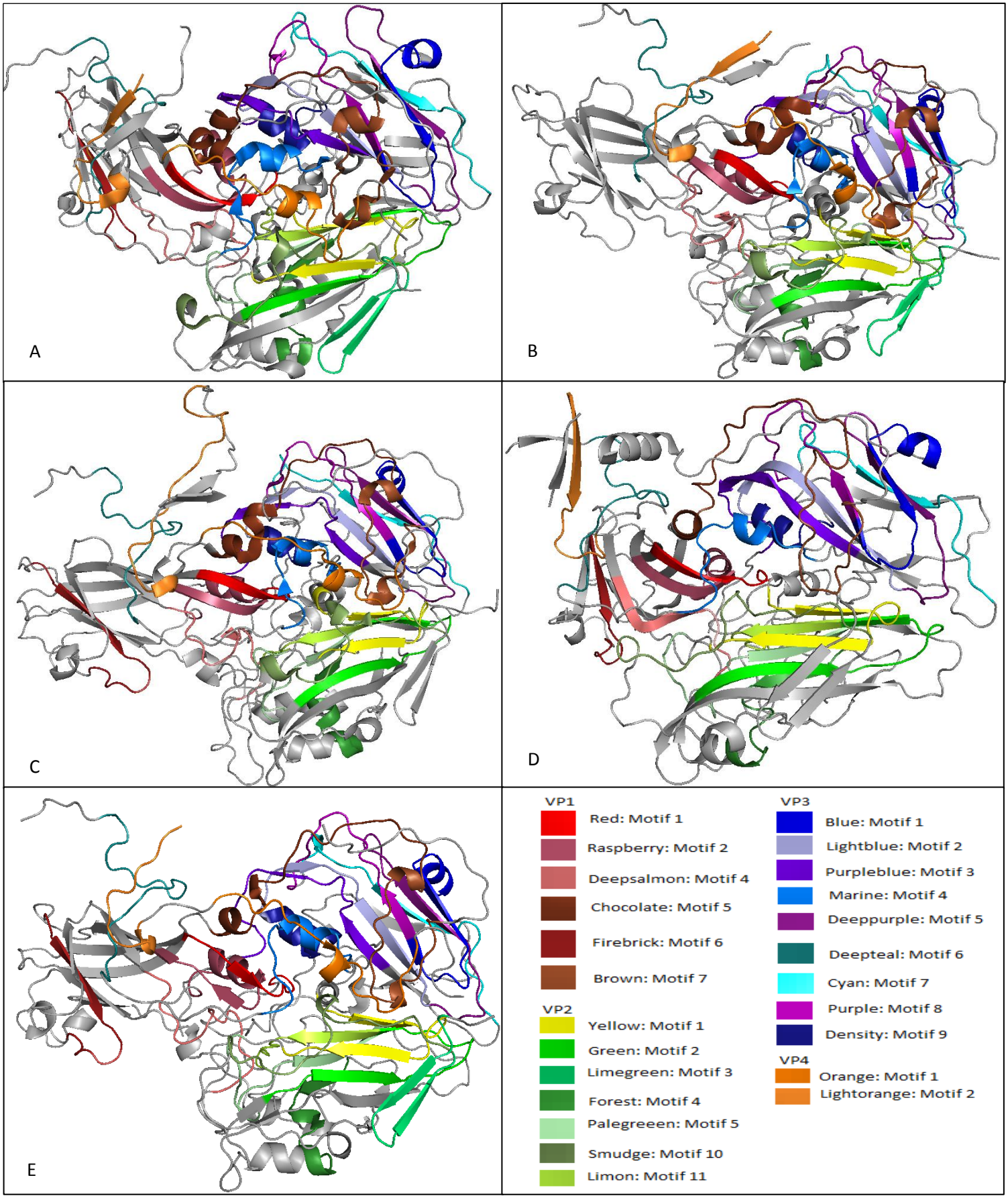


Figure 3.15. Structural mapping of predicted interacting motifs within representative enterovirus capsid protomers. Crystal structures of the viral capsid protomer of representative picornaviruses were submitted to the Protein Interaction Calculator (PIC) for the prediction of residues involved in protein-protein interactions. Predicting interacting residues were matched against residues of virus specific motif block diagrams for the identification of IM within the subunits of each representative virus. Motifs are coloured according to the bottom right key. A) EV-A CV-A24. B) EV-B CV-B3. C) EV-C PV Mahoney. D) RV-A16. E) RV-B3

The analysis of motif-motif interactions within the subunit interface of FMDV and ThV indicated significant variations from observations in the enteroviruses. Moreover significant differences were also observed between the interactions of the FMDV interface and ThV interface. The network of multiple interactions between the subunits within the protomers of these viruses is depicted in Figure 3.17. As indicated the interaction between VP1.1 and VP2.1 was conserved across both viruses. As previous stated this interaction was also conserved across the representative enteroviruses. Conservation was also observed in the interactions of VP1.1 with VP4.4 and VP1.1 with VP2.21. As VP1.1 was also predicted to be highly interactive within the enteroviruses, this result provided additional evidence of the protagonist role that VP1.1 may play in the assembly of many picornaviruses. As indicated by the dotted green lines, a significant proportion of the interactions were observed to be unique to subunits of the FMDV protomer. The figure indicated that VP1.9 played a significant role in the interaction of VP1 with both the VP2 and VP3 subunits. This was further supported by the PyMOL results, which showed a possible interactive loop in close proximity to the multiple VP2 motifs (Figure 3.16). Similarly the dotted red lines indicated a total of eight interactions which were uniquely predicted within the ThV protomer. VP1.28 (depicted as a loop region in Figure 3.16) was predicted to interact with VP2, VP3 and VP4 subunits. As in the enteroviruses, the interaction between ThV VP2 and VP3 was predicted to be maintained by VP2.1. Conversely, in FMDV the interaction appeared to be maintained by multiple interactions of VP3.8 and VP2.6. Moreover PyMOL visualisation indicated the close proximity of these motifs (Figure 3.16), with VP3.8 depicted as a purple loop interacting with the splitpea coloured helix of VP2.6. Although the predicted interactions are only based on a single representative PDB file of each virus, the identified motifs were found to be completely conserved across all available sequences of the respective virus strains. Thus it is likely that the predicted interactions may be conserved across all virus strains. Therefore the observation of virus dependent interactions may provide further support of genetic recombination between closely related species.

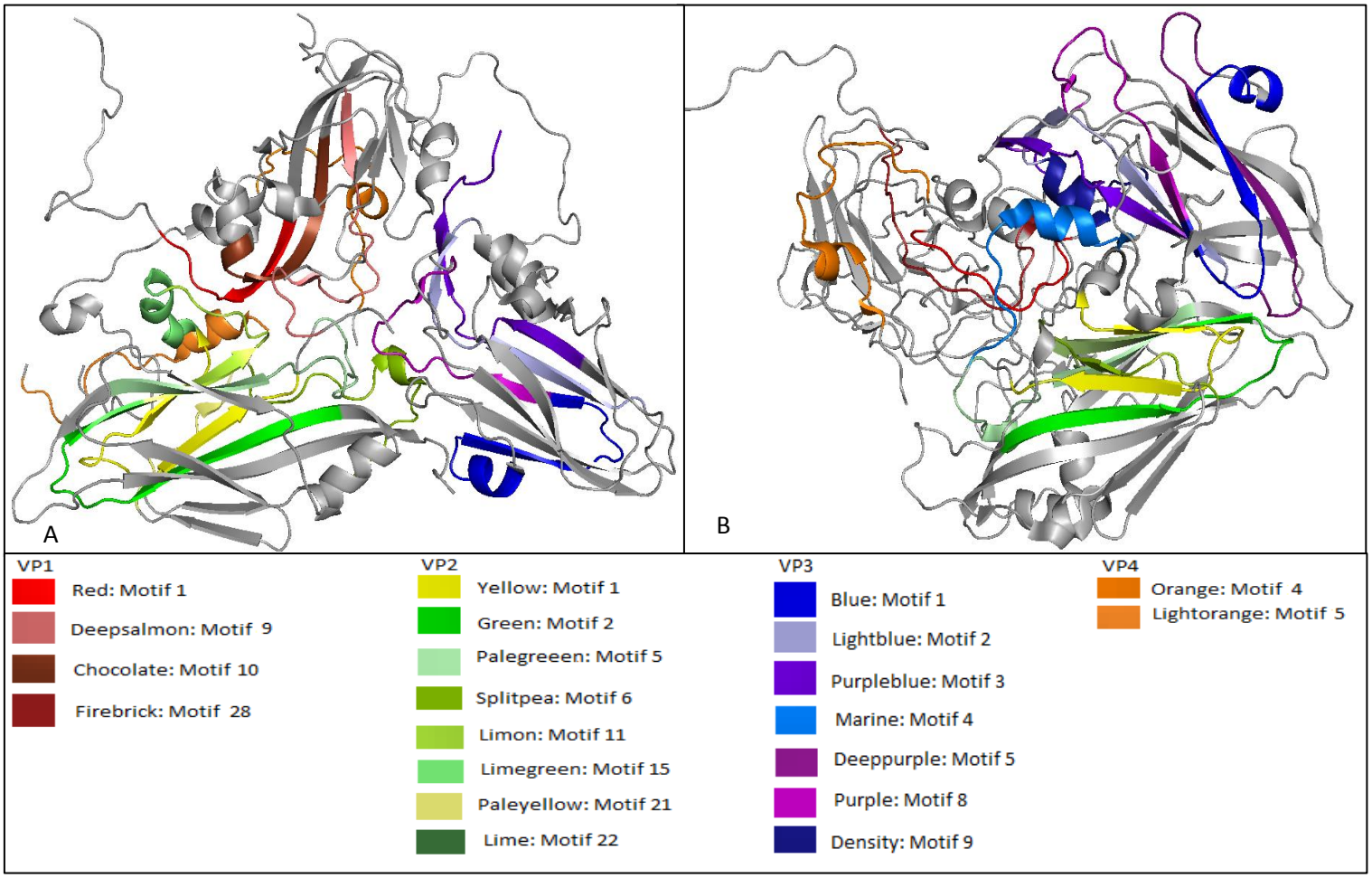


Figure 3.16. Structural mapping of predicted interacting motifs within representative FMDV and ThV capsid protomers. Crystal structures of the viral capsid protomer of representative picornaviruses were submitted to the Protein Interaction Calculator (PIC) for the prediction of residues involved in protein-protein interactions. Predicting interacting residues were matched against residues of virus specific motif block diagrams for the identification of IM within the subunits of each representative virus. Motifs are coloured according to the bottom right key. A) FMDV A22-Iraq. B) ThV DA.

Although the study identified a network of motif-motif interactions which may be fundamental in the assembly of enteroviruses, FMDV and ThV virus strains, time constraints only allowed for further analysis of motifs which were predicted to play a protagonist role in capsid subunit interactions. The motifs selected for further analysis included VP4.2, VP4.4, VP2.1, VP2.11, VP3.1 and VP1.1, while the details of all predicted interactions are presented in Table 3.7.

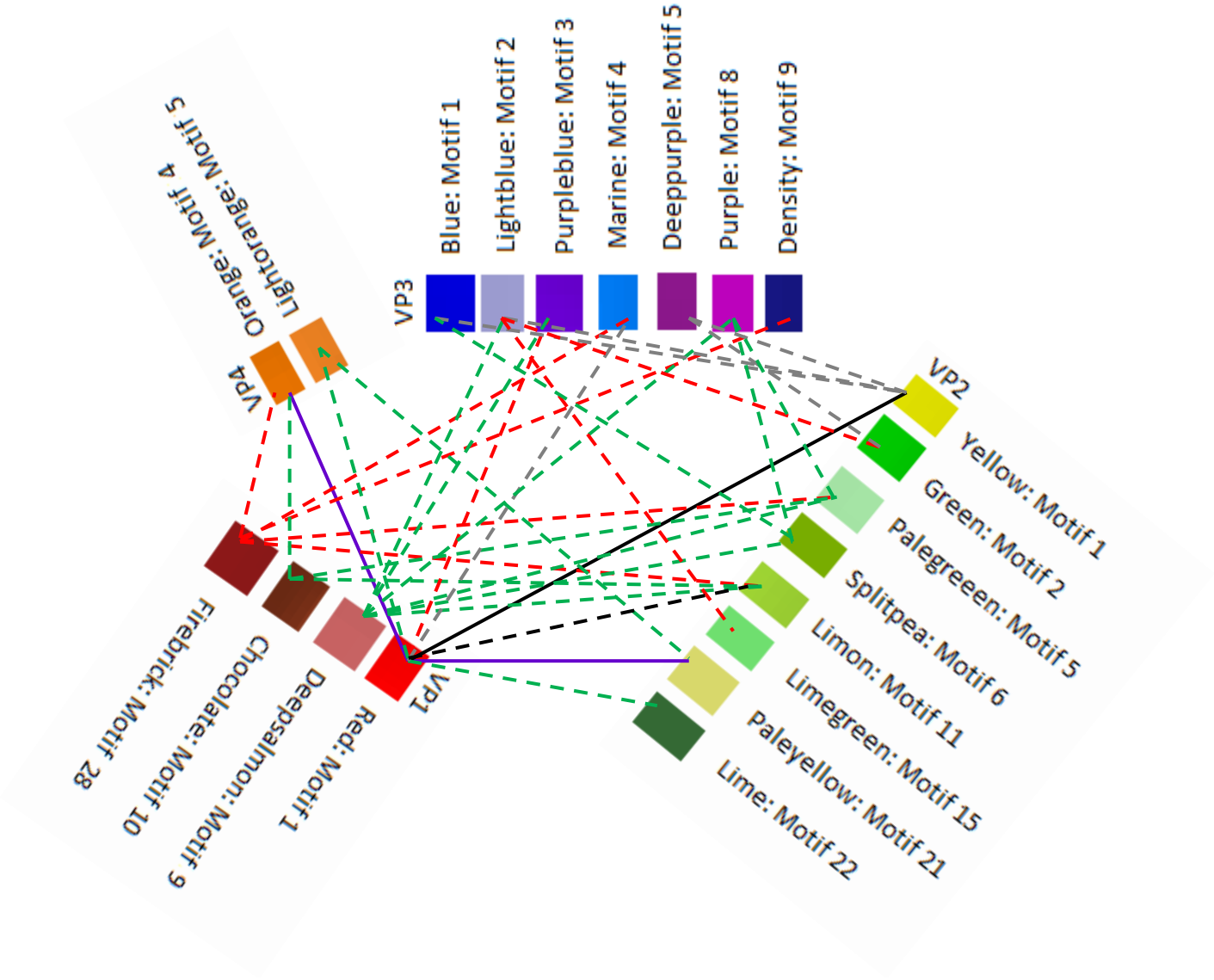


Figure 3.17. Network of conserved motif-motif interactions between subunits of representative FMDV and ThV capsid protomers. Crystal structures of the viral capsid protomer of representative picornaviruses were submitted to the Protein Interaction Calculator (PIC) for the prediction of residues involved in protein-protein interactions. Predicting interacting residues were matched against residues of virus specific motif block diagrams for the identification of IM within the subunits of each representative virus. Solid Black Lines) Interactions conserved across all representative picornaviruses. Solid Purple Lines) Interactions conserved across both FMDV and ThV representative viruses. Dotted Grey Lines) Interactions conserved across all enterovirus and ThV representative viruses. Dotted Black Lines) Interactions conserved across all enterovirus and FMDV representative viruses. Dotted Red Lines) Interactions unique to the ThV representative viruses. Dotted Green Lines) Interactions unique to the FMDV representative viruses.

Additional analysis involved the calculation of the total number of predicted interactions per residue within the relative block sequence of the respective representative viruses. The principle aim was to identify residues which were predicted to play a principal role in motif-motif interactions in each of the virus structures. This was supplemented by a comprehensive analysis of residue conservation across all available strains of the respective virus sub-groups: EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV. Residue conservation was calculated individually for each motif as a fraction of the number of residues sites at each position per total number of sequences within each sub-group. Moreover the regular expression of each motif was also analysed and any residue deviations were identified. Matplotlib was used to construct histogram plots of residue conservation within each sub-group for each motif. Upper-case letters were used to indicate residues of the regular expression, while lower-case letters indicated the presence of residues which were not present in the regular expression of the motif. The next section discusses the prediction of the principle interacting residues, as well as the conservation of these residues across all available strains of the respective viruses.

3.3.5 Analysis of motif-specific interacting residues

The results indicated that a change in residue conservation at specific positions within each motif was largely dependent on individual viral species. Moreover substitutions commonly involved similar amino acids, thus motif function appeared to positively selected and maintained across different picornavirus species, thus furthering the evidence that these motifs may be functionally important. Frequent substitutions between phenylalanine (F) and tyrosine (Y) were observed. This is a well-known substitution as both amino acids present hydrophobic properties with aromatic side chains. Furthermore they differ only by the presence of a hydroxyl group on the benzene ring of the tyrosine side chain (Voet and Voet, 2011). Moreover frequent substitutions were also observed between the positively charged lysine (K) and arginine (R) residues, thus retaining the possibility of ionic protein-protein interactions. Frequent substitutions were also observed amongst the aliphatic hydrophobic residues: alanine (A), valine (V), isoleucine (I), leucine (L) and methionine (M), while the polar amino acids serine (S), threonine (T), glutamine (Q) and asparagine (N) were commonly substituted for each other. A detailed analysis of residue substitution and conservation, respect to each motif, is discussed in the following sections.

3.3.5.1 Analysis of VP4 motifs

Previous analysis of motifs within the VP4 protein dataset, predicted that motifs 1, 2, 4 and 5 may play a role in subunit-subunit interactions within the protomers of various picornaviruses. Specifically motifs 1 and 2 were predicted to play a role in enteroviruses; motif 5 in FMDV and motif 4 in both FMDV and ThV. Consequent to the conservation of interactions, motifs 2 and 4 were selected for a comprehensive residue analysis.

3.3.5.1.1 Analysis of motif 2

Motif 2 was completely conserved and unique to all sequences of the enterovirus genus (Figure 3.4). The motif logo (Table 3.3a) indicated high conservation of the Ser5 (polar), Ile12 (hydrophobic), Asn13 (polar) and Asp17 (negatively charged) residues, thus indicating that these residues may be functionally important. This result was further supported by the PIC analysis (Figure 3.18), where Ile12, Asn13 and Asp17 were predicted to play a role in protein-protein interaction across several of the representative viruses. Residue conservation was calculated with respect to the following specific virus sub-groups: EV-A (61 sequences), EV-B (121 sequences), EV-C (55 sequences), RV-A (35 sequences), RV-B (22 sequences). Figure 3.18 indicates virus specific residue substitution. Furthermore substitution was observed to maintain physicochemical properties. This was indicated by the substitutions between: 1) positively charged residues Lys16 and Arg16 (EV-C serotypes), 2) aliphatic hydrophobic residues Ile7, Leu7, Val7 and 3) aromatic residues Tyr and Phe at positions 9, 14 and 15, 4). Table 3.7 indicates that the hydrophobic interaction of Tyr15 with Pro16 of VP3.3 was conserved across the representative viruses, with the exception of RV-A in which Phe15 was predicted to interact with Pro16. Greatest deviation from the regular expression was observed at positions 2 and 3 in EV-A and RV-C serotypes, position 5 in EV-B and RV-B serotypes and positions 2 and 16 in EV-C serotypes. Figure 3.18 indicates that no interactions were predicted at these relative sites, thus providing further evidence that these residues may not be functionally important for capsid assembly. With respect to EV-A, the residues Ile12, Phe14, Phe15 and Asn17 were predicted to play a critical role in protein-protein interaction and were highly conserved within motif 2 of the EV-A VP4 sequences. Contrast to this finding, Ala19 and Ser20 were predicted as principle interacting residues but were not well conserved across other EV-A sequences. However these residues were also predicted to be the principle interactors within the other representative enteroviruses, with significant conservation across the respective sub-groups. Analysis of the EV-B virus (CV-B3) indicated that all interacting residues: Asn13, Phe15, Arg16, Asp17, Ser18 and Ser20 were all highly conserved across the EV-B dataset. This correlation was also observed in the

other enterovirus datasets, specifically with respect to the conservation of Ile12, Asn13, Tyr14, Try/Phe15, Asp17, Ala19 and Ser20.

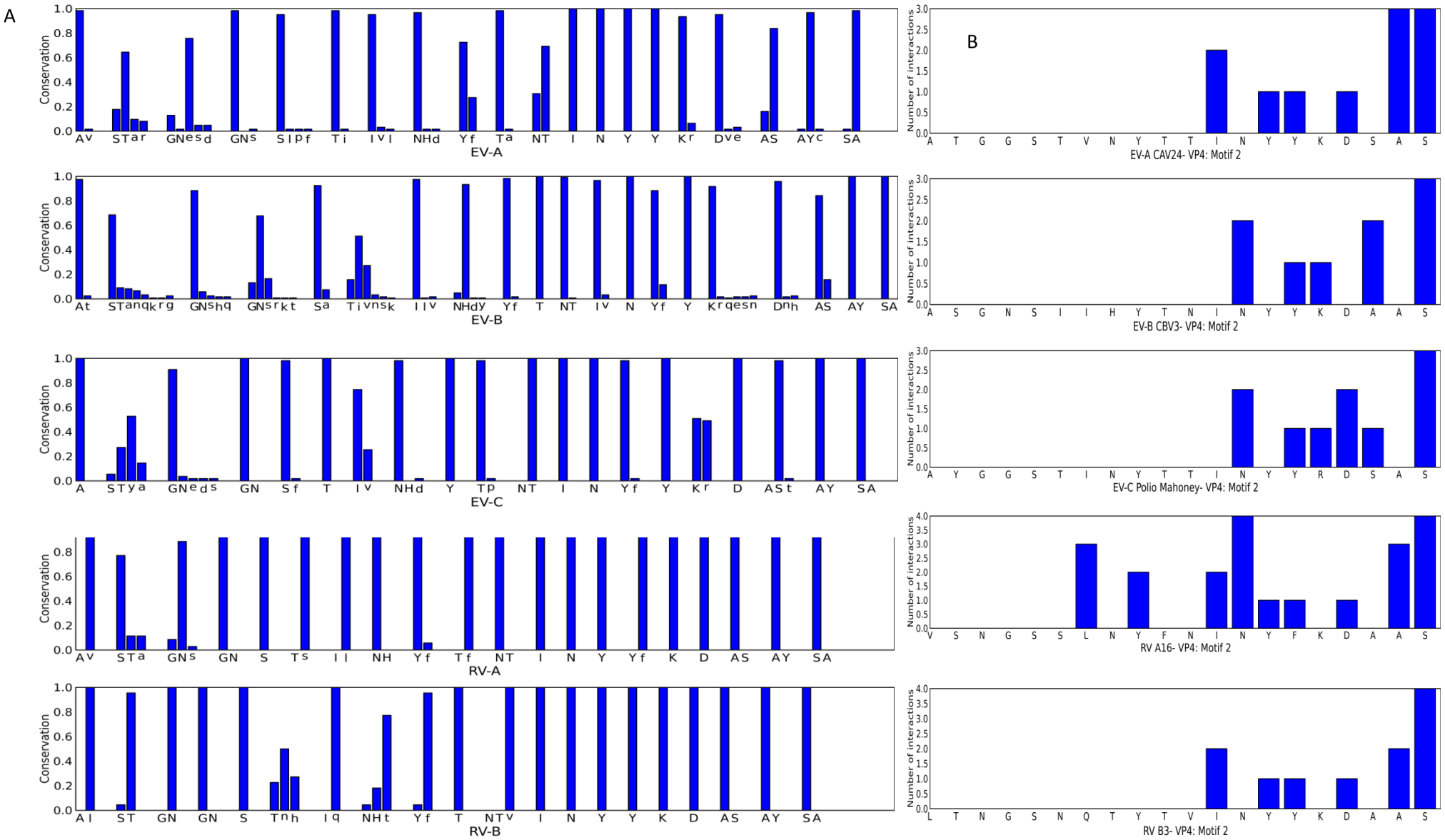


Figure 3.18. Histogram plots of virus specific residue analysis of VP4 Motif 2. A) Residue conservation determined across all available VP4 sequences of EV-A, EV-B, EV-C, RV-A and RV-B virus species. Uppercase letters indicate residues of the regular motif expression. Lowercase letters indicate deviation from the regular motif expression. Conservation calculated as a fraction of 1. B) Total number of predicted interacting residues within the motif block sequence of specific representative virus sequences. Interacting residues were predicted by PIC and mapped to corresponding motifs by Python scripting.

3.3.5.1.2 Analysis of motif 4

Motif 4 was observed in nine different virus sub-groups, as shown in Figure 3.4, with complete conservation across the 65 FMDV sequences and the 16 ThV sequences. Residue conservation was analysed with respect to these two virus groups. As evident by the motif logo (Table 3.3a) residues Asn1, Gly3, Ile5, Asn8, Tyr10, Try14 and Asp19 were highly conserved across all FMDV and ThV sequences. The functionally importance of Tyr14 and Asp19 was further supported by the PIC interacting prediction in both the FMDV and ThV representative viruses (Figure 3.19). Moreover Tyr15 was also predicted to play an interacting role in both viruses. Residue preference was observed to be virus dependent with the following conserved substitutions: Thr2 in FMDV to Glu2 in ThV, Ser4 in FMDV to Val4 in ThV, Met11 in FMDV to Ser11 in ThV, Met18 in FMDV to Ile18 in ThV and Thr20 in FMDV to Leu20 in ThV, all of which appeared to facilitate additional interactions within the ThV protomer. The study identified five possible interacting residues within motif 2 of FMDV A22-Iraq and 10 within ThV DA, all of which were completely conserved across the 65 FMDV VP4 sequences and 16 ThV sequences respectively. Principle interacting residues included Asn1, Ser17 and Asp19 in FMDV and Glu2, Ile18 and Asp19 in ThV.

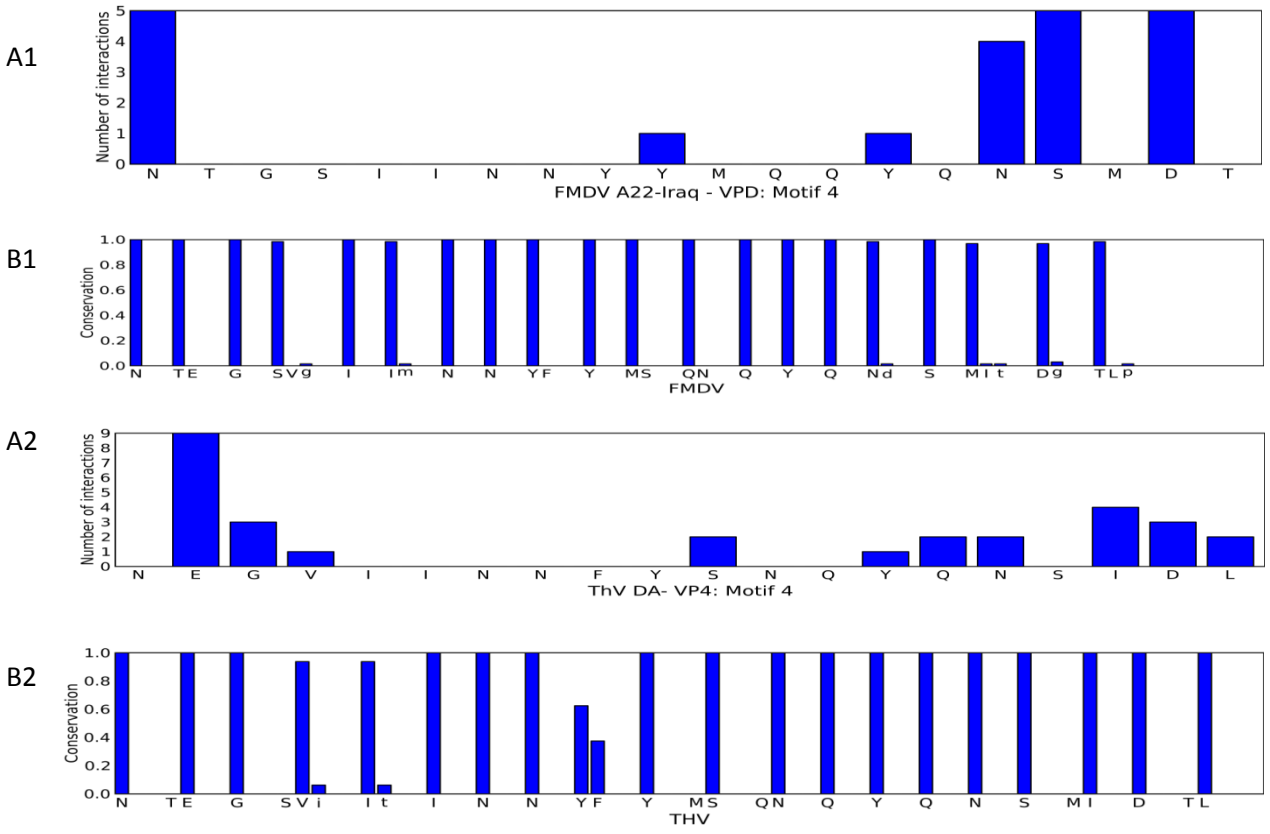


Figure 3.19. Histogram plots of virus specific residue analysis of VP4 Motif 4. A.1 and A.2) Total number of predicted interacting residues within the motif block sequence of specific representative virus sequences. Interacting residues were predicted by PIC and mapped to corresponding motifs by Python scripting. B.1 and B.2) Residue conservation determined across all available VP4 sequences of FMDV and ThV species respectively. Uppercase letters indicate residues of the regular motif expression. Lowercase letters indicate deviation from the regular motif expression. Conservation calculated as a fraction of 1

3.3.5.2 Analysis of VP2 motifs

Previous analysis of predicted interaction residues within the VP2 dataset revealed that motifs 1, 5 and 11 were predicted to facilitate motif-motif interactions within the subunit interface of the representative enteroviruses, FMDV A22-Iraq and ThV DA. Additional mapping of the subunit-subunit interactions (Figures 3.13-3.14 and 3.17), identified motifs 1 and 11 as protagonist IMs for all representative viruses. Thus these motifs were further analysed with respect to interactions per residue and virus-specific residue conservation. Specific residue conservation was calculated over the following virus sub-groups: EV-A (147 sequences), EV-B (147 sequences), EV-C (176 sequences), RV-A (89 sequences), RV-B (31 sequences), FMDV (229 sequences) and ThV (32 sequences).

3.3.5.2.1 Analysis of motif 1

Motif 1 was highly conserved across 32 of the 33 virus sub-groups, with absence only from the EMV sequences. As indicated by the logo (Table 3.3b) the residues Pro2, His3, Asn11 and Pro19 were highly conserved across the *Picornaviridae* family. The motif appeared to be the key player in the interaction with subunit VP3 in enteroviruses and ThV D. While it also facilitated interaction with VP1.1 through a conserved interaction in all representative viruses (Figure 3.13). The residues Phe1/Tyr1, Gln4, Asn7, Arg9 and Pro19 were predicted to be involved in multiple interactions within the subunit interface of all enteroviruses and ThV DA (Figure 3.20). With respect to the representative FMDV strain, Phe1, Phe5 and Thr9 were predicted as the principle interacting residues. Analysis of Table 3.7, indicated the conservation of hydrophobic interactions between Phe1 and VP1.1, with the exception of EV-A where Phe1 is substituted with interacting residue Tyr1. Furthermore Table 3.7 also indicated that the hydrogen bond interaction between Arg9 and VP3.1.Gly15 was predicted to be completely conserved across the representative enteroviruses and ThV DA. Additionally the hydrophobic interaction between Pro10 and VP3.4.Pro2 was conserved across all representative enteroviruses. The functional importance of all residues mentioned above, was further supported by their significant conservation across all strains of the relative respective viruses (Figure 3.20).

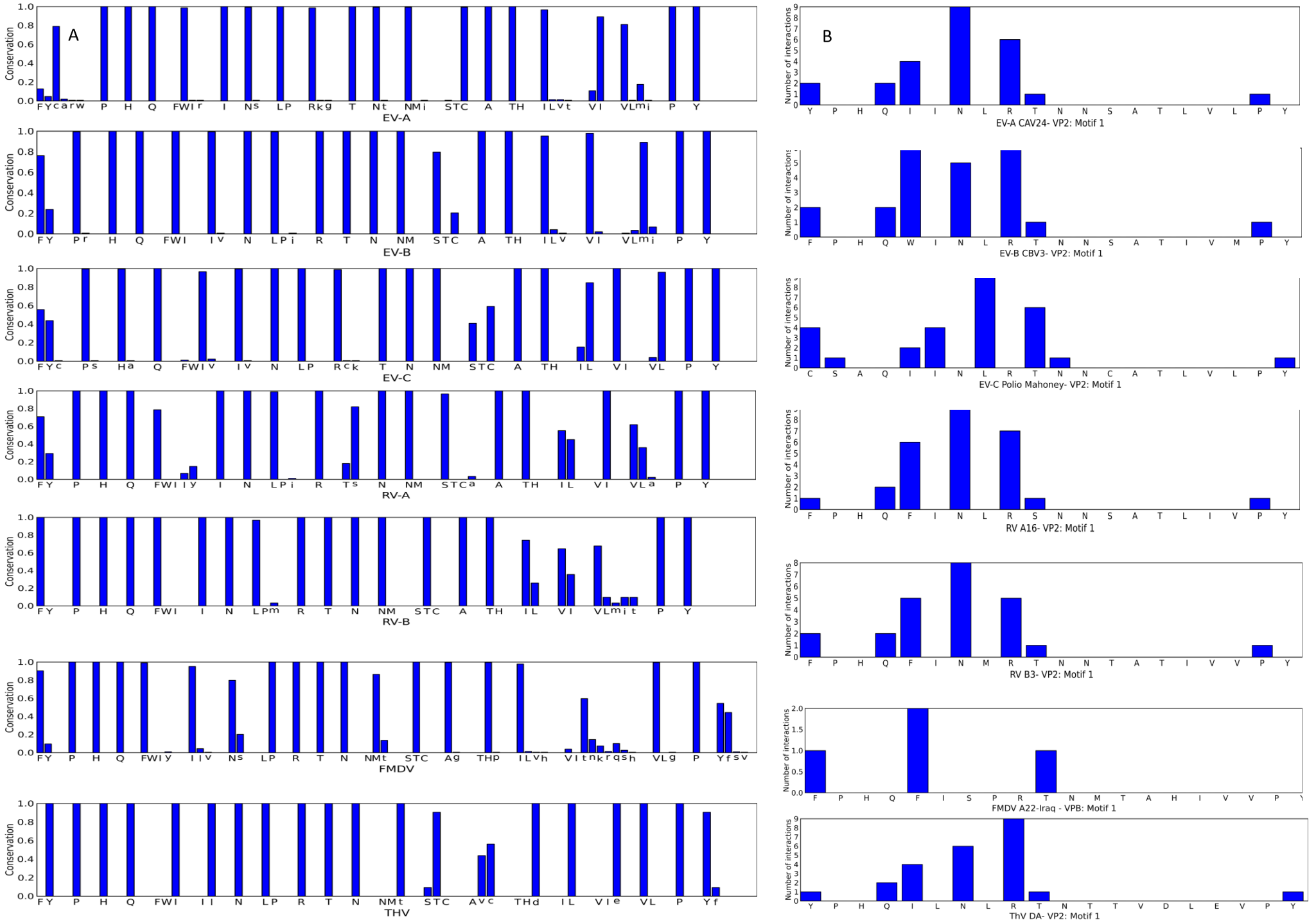


Figure 3.20. Histogram plots of virus specific residue analysis of VP2 Motif 1. A) Residue conservation determined across all available VP2 sequences of EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV virus species. Uppercase letters indicate residues of the regular motif expression. Lowercase letters indicate deviation from the regular motif expression. Conservation calculated as a fraction of 1. B) Total number of predicted interacting residues within the motif block sequence of specific representative virus sequences. Interacting residues were predicted by PIC and mapped to corresponding motifs by Python scripting.

3.3.5.2.2 Analysis of motif 11

Motif 11 was also highly conserved across the *Picornaviridae* family, with complete conservation across all enterovirus, FMDV and ThV sequences. Residues Val2, Pro6 and Glu7 were conserved throughout all seven sub-groups (Figure 3.21). Moreover Pro6 and Glu7 were predicted to be involved in multiple interactions across all seven representative viruses. Table 3.7, indicates the conservation of the side chain-side chain hydrogen bond between Glu7 with VP1.2.Tyr15 across the five representative enteroviruses. While in FMDV the residue was predicted to interact with VP1.10.Tyr5 and VP1.28.Tyr7 in ThV DA. Significant variations from the regular expression were observed at positions 8-12 throughout all virus sub-groups. However the calculation of number of interactions per residue for each representative virus indicates that these sites were predicted to play principle roles in protein-protein interaction. As motif 11 was primarily involved with the interaction of the highly variable VP1 subunit, this observation could be a result of virus-specific selectivity towards the respective VP1 subunits. Conservation of residues at these sites did appear to be conserved across their specific virus sub-groups, particularly the residues Phe8 (ThV), His8 (RV-A, RV-B), Gln9 (RV-A, RV-B), Thr10 (ThV), Met10 (EV-A), Ala11 (EV-C, RV-A, RV-B) and Ser12 (RV-B).

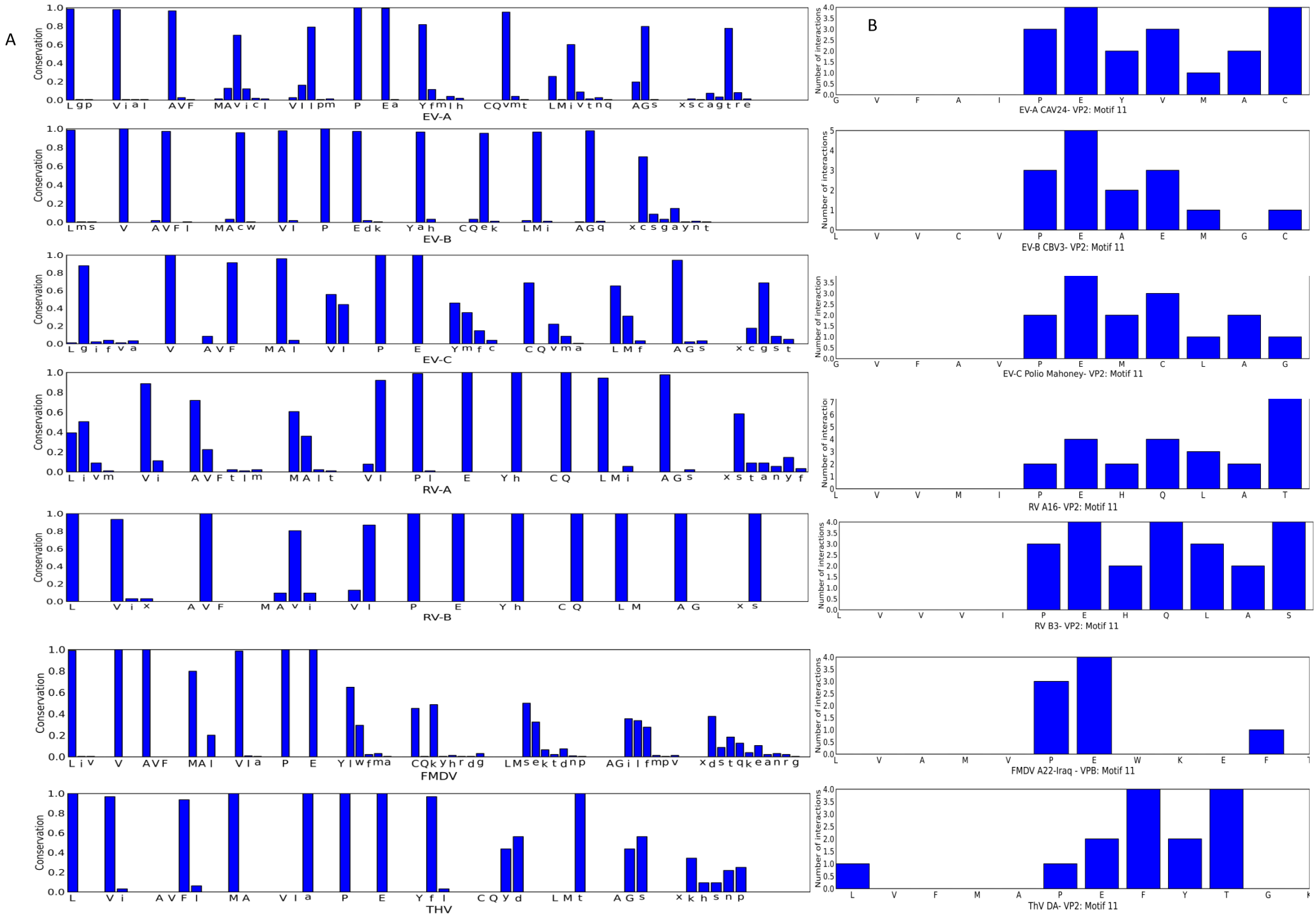


Figure 3.21. Histogram plots of virus specific residue analysis of VP2 Motif 11. A) Residue conservation determined across all available VP2 sequences of EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV virus species. Uppercase letters indicate residues of the regular motif expression. Lowercase letters indicate deviation from the regular motif expression. Conservation calculated as a fraction of 1. B) Total number of predicted interacting residues within the motif block sequence of specific representative virus sequences. Interacting residues were predicted by PIC and mapped to corresponding motifs by Python scripting.

3.3.5.3 Analysis of VP3 motifs

Previous analysis of predicted interaction residues within the VP3 dataset, revealed that motifs 1, 2 and 3 were predicted to facilitate motif-motif interactions within the subunit interface of the representative enteroviruses, FMDV A22-Iraq and ThV DA. Additional mapping of the subunit-subunit interactions (Figure 3.13) identified motifs 1 and 2 as protagonist IMs for all representative viruses. Due to time constraints only motif 1 was further analysed with respect to interactions per residue and virus-specific residue conservation. Specific residue conservation was calculated over the following virus sub-groups: EV-A (111 sequences), EV-B (147 sequences), EV-C (181 sequences), RV-A (89 sequences), RV-B (29 sequences), RV-C (18 sequences), FMDV (229 sequences) and ThV (30 sequences).

3.3.5.3.1 Analysis of motif 1

Motif 1 was highly conserved across the VP3 proteins of the *Picornaviridae* family. As shown in Figure 3.8, the motif had significant conservation across sequences of the larger sub-groups of the enteroviruses as well as the FMDV and ThV sub-groups. As indicated by the motif logo (Table 3.3.c), residues Ala4, Trp12, Asp13, Gly15, Leu16 and Ser18 had significant conservation across the *Picornaviridae* family. The functional importance of these motifs was further supported by the PIC results which predicted Trp12 as interacting within the five enteroviruses as well as Gly15 and Ser18 in the five enteroviruses and ThV DA (Figure 3.22). Moreover Table 3.7 indicated the conservation of the side chain-side chain hydrogen bond between Ser18 and VP1.6.Glu7 across the five enteroviruses, while Table 3.7 shows the conservation of the main chain-side chain hydrogen bond between Gly15 with VP2.1.Arg9 in the enteroviruses as well as ThV DA. Virus specific residue analysis revealed significant conservation of Arg1, Met5, Leu6 and Gln17 across all sequences of the *Enterovirus* genus, while Pro1, Ala5, His6 and Asn17 were completely conserved across all FMDV sequences and Gln6, Try9 and Asn17 were conserved across all ThV sequences. The PIC results predicted Gln17 and Asn17 as principle IMs across the enteroviruses and ThV DA respectively (Figure 3.22).

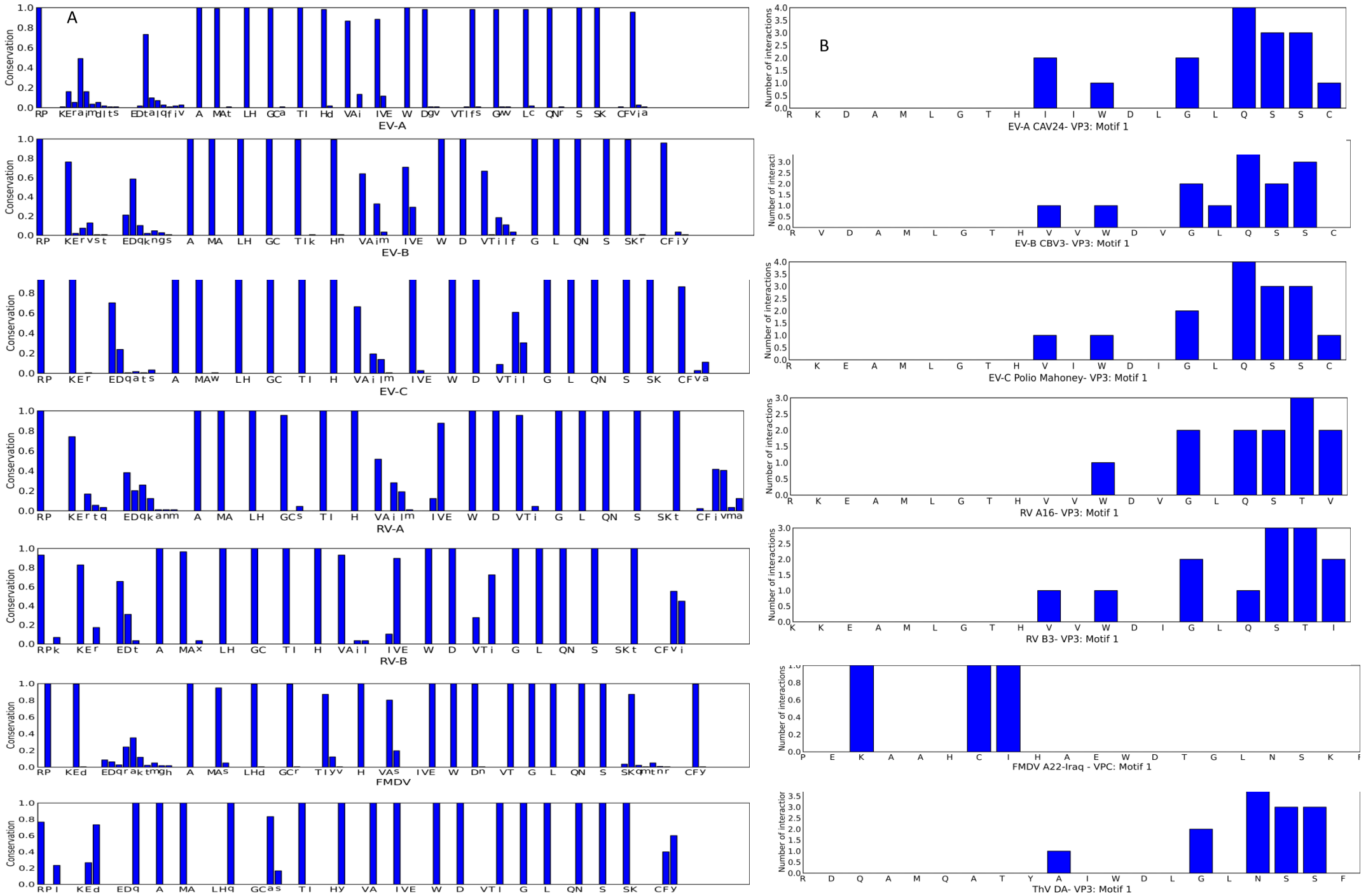


Figure 3.22. Histogram plots of virus specific residue analysis of VP3 Motif 1. A) Residue conservation determined across all available VP3 sequences of EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV virus species. Uppercase letters indicate residues of the regular motif expression. Lowercase letters indicate deviation from the regular motif expression. Conservation calculated as a fraction of 1. B) Total number of predicted interacting residues within the motif block sequence of specific representative virus sequences. Interacting residues were predicted by PIC and mapped to corresponding motifs by Python scripting.

3.3.5.4 Analysis of VP1 motifs

Previous analysis of predicted interaction residues with respect to the VP1 dataset revealed that motifs 1 may participate in motif-motif interactions within the subunit interface of the all representative enteroviruses, FMDV A22-Iraq and ThV DA. Additional mapping of the subunit-subunit interactions (Figure 3.13- 3.14 and 3.17), identified motifs 1 as protagonist IMs for all representative viruses as well as: motifs 2, 4 and 7 as principle IMs in the enteroviruses, motif 9 in FMDV A22-Iraq and Motif 28 in ThV. Consequent to time constraints the study focused on additional analysis of motif 1 with respect to the following groups: EV-A (111 sequences), EV-B (147 sequences), EV-C (181 sequences), RV-A (89 sequences), RV-B (29 sequences), RV-C (18 sequences), FMDV (229 sequences) and ThV (30 sequences). The total number of interactions per residue site was calculated respectively for each representative virus. This was followed by an analysis of the motif specific residues throughout all available strains of each virus.

3.3.5.4.1 Analysis of motif 1

Motif 1 was predicted to participate in conserved interactions with VP2, VP3 and V4 in all the representative enteroviruses as well as FMDV and ThV. Moreover, the direct interaction between motif 1 with VP2.1 was completely conserved (Figure 3.13). As shown in Figure 3.10, motif 1 was one of more conserved VP1 motifs, with sites in 15 of the 33 virus sub-groups. The motif was completely conserved across the larger sub-groups of the enterovirus genus as well as the FMDV and ThV sequences. As indicated by the motif logo (Table 3.3d), the residues Lys1, Pro8, Arg9 and Pro10 were significantly conserved across the *Picornaviridae* family and completely conserved across all sequences of the seven sub-groups shown in Figure 3.23. The PIC results further supported the possible functional importance of these residues with Lys1 predicted to interact within the protomer of all representative enteroviruses and ThV DA, while Pro8, Arg9 and Pro10 were predicted as interacting residues throughout all the representative structures (Figure 3.23). The analysis also revealed that His2 was also completely conserved across the sub-groups of the enterovirus genus, while FMDV sequences favoured Arg2 and ThV favoured Lys2, with all of these residues having positively charged side chains which were predicted to have multiple interactions (Figure 3.23). This observation provides evidence for the conservation of similar amino acids which may be of functional importance. This was further supported by the significant conservation of the interacting motifs Trp5 in the enteroviruses and Phe5 in the ThV sequences.

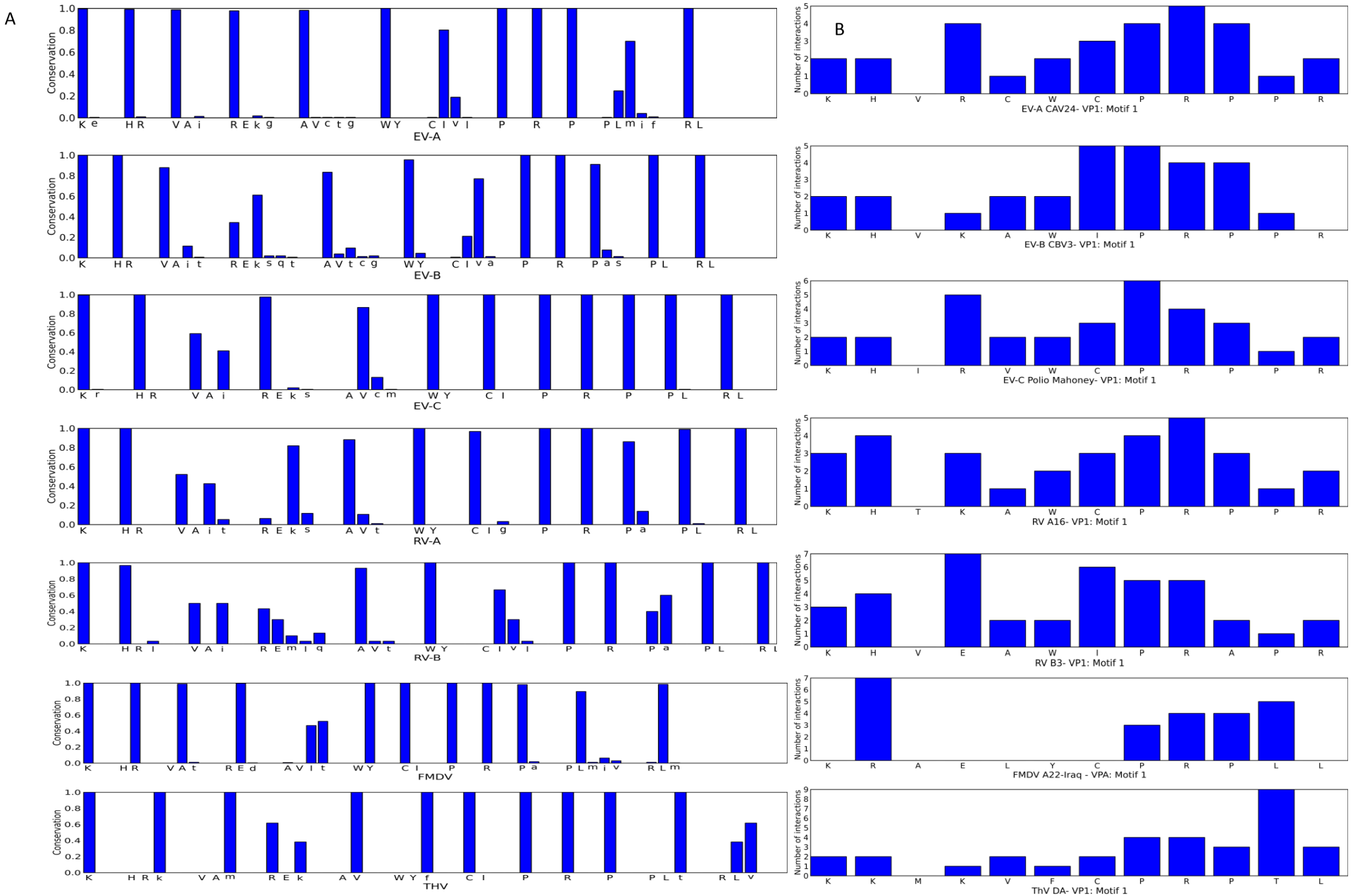


Figure 3.23. Histogram plots of virus specific residue analysis of VP1 Motif 1. A) Residue conservation determined across all available VP1 sequences of EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV virus species. Uppercase letters indicate residues of the regular motif expression. Lowercase letters indicate deviation from the regular motif expression. Conservation calculated as a fraction of 1. B) Total number of predicted interacting residues within the motif block sequence of specific representative virus sequences. Interacting residues were predicted by PIC and mapped to corresponding motifs by Python scripting.

Table 3.7. Details of predicted subunit motif-subunit motif interactions in representative picornaviruses. Crystal structures of the viral capsid protomer of representative picornaviruses were submitted to the Protein Interaction Calculator (PIC) for the prediction of residues involved in protein-protein interactions. Predicting interacting residues were matched against residues of virus specific motif block diagrams for the identification of IM within the subunits of each representative virus. RES: Residue. I: Type of Interactions. HY: Hydrophobic. SSH: Side chain-side chain hydrogen bond. MSH: Main chain-side chain hydrogen bond. MMH: Main chain-main chain hydrogen bond.

EV-A			EV-B				EV-C			RV-A			RV-B			FMDV			ThV		
RES	RES	I	RES	RES	I	RES		I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I
Interaction between VP1 -Motif 1 and VP2-Motif 1																				
ARG9	TYR1	SSH	ILE7	PHE1	HY	PRO8	VAL1	HY				ALA10	PHE1	HY						
PRO10	TYR1	HY	PRO10	PHE1	HY	PRO10	PHE2	HY	PRO10	PHE1	HY	ILE7	PHE1	HY	PRO10	PHE1	HY	PRO10	TYR1	HY
						PRO10	VAL1	HY												
Interaction between VP1 -Motif 1 and VP2-Motif 10																				
CYS7	TYR9	SSH	ILE7	TYR9	HY	CYS7	TYR9	SSH	CYS7	TYR9	SSH	ILE7	TYR9	HY						
PRO8	TYR9	HY	PRO8	TYR9	HY	PRO8	TYR9	HY	PRO8	TYR9	HY	PRO8	TYR9	HY						
Interaction between VP1 -Motif 1 and VP2-Motif 11																				
ARG9	GLU7	MSH	ARG9	GLU7	MSH	ARG9	GLU7	MSH	ARG9	GLU7	MSH	ARG9	GLU7	MSH						
ARG9	GLU7	MSH	ARG9	GLU7	MSH	ARG9	GLU7	MSH	ARG9	GLU7	MSH	ARG9	GLU7	MSH						
ARG9	PRO6	MSH	ARG9	PRO6	MSH	ARG9	PRO6	MSH	ARG9	PRO6	MSH	ARG9	PRO6	HY						
ARG9	PRO6	MSH	ARG9	PRO6	MSH	ARG9	PRO6	MSH	ARG9	PRO6	MSH	ARG9	PRO6	MSH						
			ILE7	PRO6	HY							ILE7	PRO6	MSH						
Interaction between VP1 -Motif 1 and VP2-Motif 21																				
															PRO8	TYR9	HY	PRO8	TYR9`	HY
Interaction between VP1 -Motif 1 and VP2-Motif 22																				
															PRO10	LEU9	HY			
															PRO10	LEU7	HY			
															PRO10	PHE1	HY			
															LEU11	GLN6	MSH			
															LEU11	GLN6	MSH			
															LEU11	GLN6	MSH			
															LEU11	GLN6	MSH			
															LEU11	GLN6	MSH			
Interaction between VP1 -Motif 2 and VP2-Motif 5																				
									TYR15	ALA1	HY									
Interaction between VP1 -Motif 2 and VP2-Motif 11																				
TYR15	GLU7	SSH	TYR15	GLU7	SSH	TYR15	GLU7	SSH	TYR15	GLU7	SSH	TYR15	GLU7	SSH						
TYR15	PRO6	HY																		
Interaction between VP1 -Motif 4 and VP2-Motif 4																				
ASP14	LYS5	Ionic	ASP14	LYS5	SSH	ASP14	LYS5	Ionic	ASP14	LYS5	Ionic	ASP14	LYS5	HY						
ASP14	LYS5	SSH				ASP14	LYS5	SSH	ASP14	LYS5	SSH	ASP14	LYS5	Ionic						
ASP14	LYS5	SSH				ASP14	LYS5	SSH												
Interaction between VP1 -Motif 4 and VP2-Motif 5																				
ALA6	LEU2	HY	ASN7	SER1	MMH	ASN7	SER2	MMH	ALA6	ALA1	HY	ALA6	VAL2	MMH						
ASN7	SER1	MMH	ASP14	ASN11	MMH	ASN7	SER2	MSH	ALA6	VAL2	HY	ASN10	SER1	MMH						
ASN7	SER1	MSH	ASP14	ASN11	MSH	ASN7	SER2	MSH	ALA8	ALA1	HY	ASN10	SER1	MMH						
ASN7	SER1	MSH	ASP14	ASN11	MSH	ASP14	ASN12	MMH	ASP14	ASN11	MMH	ASN10	SER1	SSH						
ASP14	ASN11	MMH	ASP14	THR14	SSH	PHE16	LYS10	MMH	ASP14	HIS10	MSH	ASP14	ASN11	SSH						
PHE16	LYS9	MMH	TRP16	ARG9	MMH	SER10	SER2	SSH	SER7	ALA1	MMH	SER7	SER1	SSH						
SER10	SER1	SSH				SER10	SER2	SSH	TYR10	ALA1	HY	TYR16	ARG9	SSH						
SER10	SER1	SSH							TYR10	VAL2	HY									
									TYR16	ARG9	MMH									
Interaction between VP1 -Motif 4 and VP2-Motif 11																				
ASP14	GLU7	MSH	ASP14	ALA8	MSH	ASP14	CYS9	MSH	ASP14	GLU7	MSH	ASP14	GLU7	MSH						
ASP14	TYR8	MSH	ASP14	ALA8	MSH	ASP14	GLU7	MSH	ASP14	HIS8	MSH	ASP14	HIS8	MSH						
ASP14	TYR8	MSH	PHE12	GLU7	MSH	ASP14	MET8	MSH	ASP14	HIS8	MSH	ASP14	HIS8	MSH						
PHE12	VAL9	HY	PHE12	GLU7	MSH	ASP14	MET8	MSH	TYR13	GLN9	MSH	TYR13	GLN9	MSH						
			TYR13	GLU9	MSH	TYR13	CYS9	MSH												
						TYR13	CYS9	MSH												
Interaction between VP1 -Motif 7 and VP2-Motif 1																				
GLU7	ASN7	SSH	ALA6	TRP5	HY	GLU7	ASN8	SSH	ALA6	PHE5	HY	GLU7	ARG3	MSH						
GLU7	ASN7	SSH	GLU7	THR10	SSH	GLU7	ASN8	SSH	GLU7	ASN7	SSH	GLU7	ARG3	MSH						
GLU7	ASN7	SSH	GLU7	TRP5	MMH	GLU7	ASN8	SSH	GLU7	ASN7	SSH	GLU7	ASN7	MMH						
GLU7	ASN7	SSH	THR8	GLN4	MSH	GLU7	ASN8	SSH	GLU7	ASN7	SSH	GLU7	ASN7	SSH						
GLU7	ILE5	MMH	THR8	GLN4	MSH	GLU7	ILE6	MMH	GLU7	ASN7	SSH	GLU7	ASN7	SSH						
GLU7	THR10	SSH				GLU7	THR11	SSH	GLU7	PHE5	MMH	GLU7	ASN7	SSH						
THR8	GLN4	MSH				THR8	GLN5	MSH	GLU7	SER10	SSH	GLU7	PHE5	Ionic						
THR8	GLN4	MSH				THR8	GLN5	MSH	THR8	GLN4	MSH	GLU7	THR10	SSH						
VAL6	ILE5	HY				VAL6	ILE6	HY	THR8	GLN4	MSH	THR8	GLN4	SSH						
Interaction between VP1 -Motif 9 and VP2-Motif 5																				
															VAL13	VAL1	HY			
															ALA15	VAL1	HY			
															ASN19	LYS11	MMH			
															HIS11	ASN2	SSH			
															HIS11	ASN2	SSH			
															HIS11	ASN2	SSH			
															ASN19	LYS11	MSH			
															ASN19	LYS11	MSH			
Interaction between VP1 -Motif 9 and VP2-Motif 6																				
															ARG12	ASP5	SSH			
															ARG12	ASP5	SSH			
															ARG12	ASP5	SSH			
															ARG12	ASP5	SSH			
Interaction between VP1 -Motif 9 and VP2-Motif 11																				
															ASN19	GLU7	MSH			
Interaction between VP1 -Motif 10 and VP2-Motif 5																				
															TYR5	VAL1	HY			
Interaction between VP1 -Motif 10 and VP2-Motif 11																				
															TYR5	PRO6	HY			
															TYR5	GLU7	SSH			
Interaction between VP1 -Motif 28 and VP2-Motif 5																				
																		TYR7	ILE1	HY
Interaction between VP1 -Motif 28 and VP2-Motif 11																				
																		TYR7	PRO6	HY
																		TYR7	GLU7	SSH

EV-A			EV-B			EV-C			RV-A			RV-B			FMDV			ThV		
RES	RES	I	RES	RES	I	RES		I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I
Interaction between VP1 -Motif 1 and VP3-Motif 3																				
																		LEU12	TRP20	HY
Interaction between VP1 -Motif 1 and VP3-Motif 4																				
ARG4	GLU4	Ionic	ALA5	VAL5	MMH	ARG4	GLU4	Ionic	ALA5	VAL5	MMH	ALA5	VAL5	HY				CYS7	GLY3	MMH
ARG4	GLU4	SSH	ILE7	ILE1	HY	ARG4	GLU4	SSH	CYS7	GLY3	MMH	ALA5	VAL5	MMH				CYS7	CYS2	MSH
ARG4	GLU4	SSH	ILE7	GLY3	MMH	ARG4	GLU4	SSH	CYS7	PRO2	MSH	GLU4	LYS4	Ionic				LEU12	TRP20	HY
ARG4	GLU4	SSH	LYS4	GLU4	SSH	ARG4	GLU4	SSH	LYS4	GLU4	Ionic	GLU4	LYS4	SSH				LYS4	GLU4	Ionic
CYS5	VAL5	MMH	PRO8	VAL5	HY	CYS7	GLY3	MMH	PRO8	VAL5	HY	ILE7	ILE1	HY				PHE6	MET1	HY
CYS7	GLY3	MMH	PRO8	ILE11	HY	CYS7	PRO2	MSH	PRO8	MET11	HY	ILE7	GLY3	MMH				PRO8	PHE5	HY
CYS7	PRO2	MSH	TRP6	ILE1	MSH	PRO8	VAL5	HY	TRP6	ILE1	MSH	PRO8	VAL5	HY				PRO8	LEU11	HY
PRO8	LEU11	HY				PRO8	LEU11	HY				PRO8	ILE11	HY				PRO8	LEU14	HY
TRP6	ILE1	MSH				TRP6	ILE1	MSH				TRP6	ILE1	HY				VAL5	PHE5	HY
						VAL5	VAL5	MMH				TRP6	ILE1	MSH				VAL5	PHE5	MMH
Interaction between VP1 -Motif 1 and VP3-Motif 9																				
PRO11	MET6	HY	PRO11	LEU6	HY	PRO11	MET6	HY	PRO11	LEU6	HY	PRO11	LEU6	HY						
						PRO8	MET6	HY												
Interaction between VP1 -Motif 2 and VP3-Motif 2																				
ARG8	TYR1	SSH	ARG8	TYR1	SSH	ARG8	TYR1	SSH	ARG8	TYR1	SSH	LEU12	TYR1	HY						
ARG8	TYR1	SSH	PHE12	TYR1	HY	ARG8	TYR1	SSH	ARG8	TYR1	SSH									
PHE12	TYR1	HY				PHE12	TYR1	HY	MET12	TYR1	HY									
Interaction between VP1 -Motif 2 and VP3-Motif 3																				
GLN5	ASP16	SSH	ALA4	ILE20	HY	GLN5	ASP16	SSH	LYS9	ASP16	Ionic	GLN5	ASP16	SSH						
GLN5	ASP16	SSH	GLN5	ASP16	SSH	GLN5	ASP16	SSH				GLN5	ASP16	SSH						
GLN5	ASP16	SSH	GLN5	ASP16	SSH	GLN5	ASP16	SSH				GLN5	ASP16	SSH						
GLN5	ASP16	SSH	GLN5	ASP16	SSH	GLN5	ASP16	SSH				GLN5	ASP16	SSH						
LYS9	ASP16	Ionic	GLN5	ASP16	SSH	GLN5	ASP16	SSH				GLN5	ASP16	SSH						
VAL4	ILE20	HY	GLN5	ASP16	SSH	GLN5	ASP16	SSH				GLN5	ASP16	SSH						
			GLN5	ASP16	SSH	GLN5	ASP16	SSH				GLN5	ASP16	SSH						
			GLN5	ASP16	SSH	GLN5	ASP16	SSH				GLN5	ASP16	SSH						
			GLN5	ASP16	SSH	VAL4	ILE20	HY				VAL4	ILE20	HY						
			GLN5	ASP16	SSH															
			GLN5	ASP16	SSH															
			GLN5	ASP16	SSH															
			GLN5	ASP16	SSH															
Interaction between VP1 -Motif 2 and VP3-Motif 4																				
PHE12	MET8	HY	PHE12	LEU8	HY	PHE12	MET8	HY	MET12	MET11	HY	LEU12	LEU8	HY						
PHE12	LEU11	HY	PHE12	ILE11	HY	PHE12	LEU11	HY	PHE13	VAL5	HY	PHE13	VAL5	HY						
PHE13	VAL5	HY	PHE13	VAL5	HY	PHE13	VAL5	HY	PHE13	LEU8	HY	PHE13	LEU8	HY						
PHE13	MET8	HY	PHE13	LEU8	HY	PHE13	MET8	HY	PHE13	MET11	HY	TYR15	ILE1	HY						
PHE13	LEU11	HY	PHE13	ILE11	HY	PHE13	LEU11	HY	TYR15	ILE1	HY									
TYR15	ILE1	HY	TYR15	ILE1	HY	TYR15	ILE1	HY												
Interaction between VP1 -Motif 2 and VP3-Motif 9																				
ARG8	GLU9	Ionic	ARG8	GLU9	SSH	ARG8	GLU9	Ionic	ARG8	GLU9	Ionic	LEU12	ILE10	HY						
ARG8	GLU9	SSH	ARG8	GLU9	SSH	ARG8	GLU9	SSH	ARG8	GLU9	SSH	LYS8	GLU9	Ionic						
ARG8	GLU9	SSH	ARG8	GLU9	SSH	ARG8	GLU9	SSH	ARG8	GLU9	SSH	LYS8	GLU9	SSH						
PHE12	ILE10	HY	PHE12	ILE10	HY	PHE12	ILE10	HY	MET12	ILE10	HY									
PHE12	MET6	HY				PHE12	MET6	HY	MET12	LEU6	HY									
Interaction between VP1 -Motif 4 and VP3-Motif 6																				
VAL3	LEU19	HY	PRO1	MET19	HY	VAL3	LEU19	HY	PRO1	LEU19	HY	VAL3	LEU19	HY						
Interaction between VP1 -Motif 5 and VP3-Motif 2																				
GLU13	TYR2	MSH	GLU13	TYR2	MSH	GLU13	TYR2	MSH	MET14	PHE2	HY	GLU13	TYR2	MSH						
PHE19	TYR1	HY	PHE19	TYR1	HY	PHE19	TYR1	HY	PHE19	TYR1	HY	PHE19	TYR1	HY						
PHE19	TYR2	HY	PHE19	TYR2	HY	PHE19	TYR2	HY	PHE19	PHE2	HY	PHE19	TYR2	HY						
Interaction between VP1 -Motif 5 and VP3-Motif 3																				
ARG11	CYS6	MSH	ARG11	CYS6	MSH	ARG11	CYS6	MSH	GLU13	MET13	MSH	GLU13	LEU13	MSH						
ARG11	CYS6	MSH	ARG11	CYS6	MSH	ARG11	CYS6	MSH	GLU13	MET13	MSH	GLU13	MET14	MSH						
ARG11	PHE9	MSH	ARG11	PHE9	MSH	ARG11	PHE9	MSH	GLU13	ALA14	MSH	PHE19	MET14	HY						
ARG11	PHE9	MSH	ARG11	PHE9	MSH	ARG11	PHE9	MSH	MET14	LEU11	HY									
ARG11	PHE9	MSH	ARG11	PHE9	MSH	ARG11	PHE9	MSH	MET14	LEU11	MSH									
GLU13	LEU13	MSH	GLU13	LEU13	MSH	ARG9	ASP8	Ionic	THR10	CYS10	MSH									
GLU13	LEU13	MSH	GLU13	LEU14	MSH	ARG9	ASP8	MSH												
GLU13	LEU14	MSH	PHE19	LEU14	HY	ARG9	ASP8	MSH												
PHE19	LEU14	HY				ARG9	ASP8	MSH												
						ARG9	ASP8	MSH												
						GLU13	LEU13	MSH												
						GLU13	LEU13	MSH												
						GLU13	LEU14	MSH												
						PHE19	LEU14	HY												
Interaction between VP1 -Motif 5 and VP3-Motif 4																				
ARG11	GLU13	Ionic	ARG11	GLU13	SSH	ARG11	GLU13	Ionic	LEU11	ILE9	HY	GLY11	ASN7	MSH						
ARG11	GLU13	SSH	ARG11	GLU13	SSH	ARG11	GLU13	SSH	LEU11	ASN7	MSH	GLY11	ASN7	MSH						
ARG11	GLU13	SSH	ARG11	GLU13	SSH	ARG11	GLU13	SSH	LEU11	ASN7	MSH	PHE19	LEU8	HY						
ARG11	GLU13	SSH	ARG11	GLU13	SSH	ARG11	GLU13	SSH	MET14	LEU8	HY	PHE9	LEU9	HY						
ARG11	GLU13	SSH	ARG11	ASN7	MSH	ARG11	GLU13	SSH	MET14	ILE9	HY	THR14	LEU8	MMH						
ARG11	ASN7	MSH	ARG11	ASN7	MSH	ARG11	ASN7	MSH	MET14	LEU8	MMH	THR14	ASN7	SSH						
ARG11	ASN7	MSH	ILE16	VAL5	HY	ARG11	ASN7	MSH	MET14	ASN7	MSH	THR14	ASN7	SSH						
LEU16	VAL5	HY	ILE16	LEU8	HY	ILE16	VAL5	HY	MET14	ASN7	MSH	THR14	ASN7	SSH						
LEU16	MET8	HY	ILE16	LYS6	MMH	ILE16	MET8	HY	PHE19	LEU8	HY	THR14	ASN7	MSH						
LEU16	PHE6	MMH	PHE19	LEU8	HY	ILE16	LYS6	MMH	SER15	ASN7	MSH	THR14	ASN7	MSH						
PHE19	MET8	HY	SER14	LEU8	MMH	PHE19	MET8	HY	VAL16	VAL5	HY	VAL16	VAL5	HY						
SER14	MET8	MMH	SER14	ASN7	MSH	SER14	MET8	MMH	VAL16	LYS6	MMH	VAL16	ARG6	MMH						
SER14	ASN7	SSH	SER14	ASN7	MSH	SER14	ASN7	SSH												
SER14	ASN7	SSH	THR15	ASN7	MSH	SER14	ASN7	SSH												
SER14	ASN7	SSH				SER14	ASN7	SSH												
SER14	ASN7	MSH				SER14	ASN7	MSH												
SER14	ASN7	MSH				SER14	ASN7	MSH												
THR15	ASN7	MSH				SER14	ASN7	MSH												
						SER14	ASN7	MSH												
						SER15	ASN7	MSH												

EV-A			EV-B			EV-C			RV-A			RV-B			FMDV			ThV		
RES	RES	I	RES	RES	I	RES		I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I
Interaction between VP1 -Motif 5 and VP3-Motif 8																				
			HIS9	HIS11	SSH				GLN9	GLN11	SSH	PHE9	PHE12	HY						
			HIS9	HIS11	SSH				GLN9	GLN11	SSH									
									GLN9	GLN11	SSH									
									GLN9	GLN11	SSH									
Interaction between VP1 -Motif 5 and VP3-Motif 9																				
									MET14	ILE10	HY									
Interaction between VP1 -Motif 6 and VP3-Motif 6																				
TYR8	LEU19	HY				TYR8	LEU19	HY	TYR8	LEU19	HY	TYR8	LEU19	HY						
Interaction between VP1 -Motif 7 and VP3-Motif 1																				
ALA2	ILE10	HY	ALA2	VAL10	HY	ALA2	VAL10	HY	ALA6	THR19	MSH	ALA2	VAL10	HY						
ALA2	CYS20	MSH	ALA6	SER19	MSH	ALA2	CYS20	MSH	ASP4	THR19	MMH	ALA2	ILE20	HY						
GLU7	SER18	SSH	GLU7	SER18	SSH	GLU7	SER18	SSH	ASP4	THR19	MMH	ASN6	THR19	MSH						
GLU7	SER18	SSH	GLU7	SER18	SSH	GLU7	SER18	SSH	GLU7	SER18	SSH	GLU7	SER18	SSH						
LEU3	TRP12	HY	LEU3	TRP12	HY	LEU3	TRP12	HY	GLU7	SER18	SSH	GLU7	SER18	SSH						
THR4	SER19	MMH	THR4	SER19	MMH	THR4	SER19	MMH	LEU3	TRP12	HY	LEU3	TRP12	HY						
THR4	SER18	MMH	THR4	SER19	MMH	THR4	SER18	MMH	LEU3	VAL20	HY	LEU3	ILE20	HY						
THR4	SER19	MMH				THR4	SER19	MMH	LEU3	GLN17	MSH	THR4	THR19	MMH						
VAL19	ILE10	HY				VAL6	SER19	MSH	LEU3	GLN17	MSH	THR4	SER18	MMH						
VAL6	SER19	MSH							VAL2	VAL20	HY	THR4	THR19	MMH						
Interaction between VP1 -Motif 7 and VP3-Motif 2																				
ALA14	ARG10	MSH	ALA6	MET14	HY	ASN12	LYS10	MSH	ALA6	MET14	HY	LEU12	ARG10	MSH						
ALA14	ARG10	MSH	GLU7	MET14	SSH	VAL6	LEU14	HY	ASN12	ARG10	MSH	THR14	ARG10	MSH						
ALA14	ARG10	MSH	GLU7	MET14	SSH				ASN12	ARG10	MSH	THR14	ARG10	MSH						
GLN13	ARG10	MSH	SER12	LYS10	MSH				ASN12	ARG10	MSH									
GLN13	ARG10	MSH							GLU7	MET14	SSH									
GLY12	ARG10	MSH							GLU7	MET14	SSH									
GLY12	ARG10	MSH							GLU7	MET14	SSH									
GLY12	ARG10	MSH							GLU7	MET14	SSH									
VAL6	LEU14	HY							ILE14	ARG10	MSH									
									ILE14	ARG10	MSH									
Interaction between VP1 -Motif 7 and VP3-Motif 3																				
VAL15	PRO7	HY	ALA6	PHE2	HY	LEU14	CYS6	MSH	ALA6	PHE2	HY	PRO13	PRO7	HY						
VAL6	PHE2	HY	VAL14	CYS6	MSH	VAL6	PHE2	HY	GLN15	CYS6	SSH	THR14	CYS6	MSH						
									GLN15	CYS6	SSH									
									GLN15	CYS6	SSH									
									GLN15	LYS7	MSH									
									GLU17	LYS7	Ionic									
									GLU17	LYS7	SSH									
									GLU17	LYS7	SSH									
Interaction between VP1 -Motif 7 and VP3-Motif 4																				
GLY12	ASP15	MSH	THR11	ASP15	MSH	THR11	ASP15	MSH	LYS13	GLN13	MSH	THR11	GLY15	MSH						
SER11	ASP15	MSH							THR11	ASP15	MSH									
Interaction between VP1 -Motif 7 and VP3-Motif 8																				
ALA14	LEU3	HY	ALA2	VAL1	HY	ALA2	THR1	MMH	ASN12	SER1	SSH	ALA2	VAL1	MMH						
ALA2	THR1	MMH	ALA2	VAL1	MMH	ASN12	THR1	SSH	ASN12	SER1	SSH	ASN19	THR3	SSH						
ILE20	PRO5	HY	MET20	PRO5	HY	ASN12	THR1	SSH	ASN12	SER1	SSH	ASN19	THR3	SSH						
PRO16	LEU3	HY	PRO16	PRO5	HY	ASN12	THR1	SSH	ILE14	VAL3	HY	ASN19	THR3	SSH						
PRO16	PRO5	HY	VAL14	VAL1	HY	LEU14	VAL3	HY	ILE20	PRO5	HY	ASN19	THR3	MSH						
VAL19	LEU3	HY				PRO16	VAL3	HY	PRO16	VAL3	HY	ASN19	THR3	MSH						
						PRO16	PRO5	HY	PRO16	PRO5	HY	ASN19	THR3	MSH						
						VAL20	PRO5	HY	VAL2	SER1	MMH	PRO16	PRO5	HY						
												VAL20	PRO5	HY						
Interaction between VP1 -Motif 9 and VP3-Motif 2																				
															THR8	GLN4	MSH			
															THR8	GLN4	MSH			
Interaction between VP1 -Motif 9 and VP3-Motif 3																				
															ALA9	ARG12	MSH			
Interaction between VP1 -Motif 9 and VP3-Motif 8																				
															PRO10	ALA10	HY			
															PRO10	TYR12	HY			
															PRO10	TYR14	HY			
															PRO10	ASP11	MSH			
Interaction between VP1 -Motif 28 and VP3-Motif 4																				
																		PHE2	LEU8	HY
																		PRO4	PHE5	HY
																		PRO4	LEU8	HY
																		PRO4	LEU11	HY
																		PHE5	PHE5	HY
																		TYR7	MET1	HY
Interaction between VP1 -Motif 28 and VP3-Motif 9																				
																		PHE2	ALA9	HY
																		PHE2	VAL10	HY
																		PRO4	MET6	HY

EV-A			EV-B			EV-C			RV-A			RV-B			FMDV			ThV		
RES	RES	I	RES	RES	I	RES		I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I
Interaction between VP1 -Motif 1 and VP4-Motif 1																				
PRO8	PHE11	HY	PRO8	PHE12	HY	PRO8	PHE12	HY				PRO8	PHE11	HY						
Interaction between VP1 -Motif 1 and VP4-Motif 2																				
LYS1	ALA19	MSH	LYS1	ALA20	MSH	LYS1	ALA20	MSH	LYS1	ALA19	MSH	LYS1	ALA19	MSH						
									LYS1	SER20	MSH	LYS1	SER20	MSH						
Interaction between VP1 -Motif 1 and VP4-Motif 4																				
															ARG2	ASP19	Ionic	LYS1	GLU3	Ionic
															ARG2	ASP19	SSH	LYS2	ASP20	Ionic
															ARG2	ASP19	SSH	LYS1	GLU3	SSH
															ARG2	ASP19	SSH	LYS2	SER18	MSH
															ARG2	ASP19	SSH			
															ARG2	SER17	MSH			
															ARG2	SER17	MSH			
Interaction between VP1 -Motif 1 and VP4-Motif 5																				
															PRO8	PHE4	HY			
Interaction between VP1 -Motif 4 and VP4-Motif 2																				
PRO1	ALA19	HY	PRO1	ALA20	HY	PRO1	ALA20	HY	PRO1	ALA19	HY	PRO1	ALA19	HY						
Interaction between VP1 -Motif 5 and VP4-Motif 1																				
ARG11	GLN6	SSH	ARG11	GLN7	SSH							ASP15	SER3	SSH						
ARG11	GLN6	SSH	ARG11	GLN7	SSH							ASN10	GLN2	MSH						
			ARG11	GLN7	SSH							ASN10	GLN2	MSH						
			ARG11	GLN7	SSH							ASP15	SER3	MSH						
			ASN18	ARG2	SSH															
			ASN18	ARG2	MSH															
			ASN18	ARG2	MSH															
Interaction between VP1 -Motif 7 and VP4-Motif 1																				
ALA10	VAL15	HY	HIS10	GLU14	Ionic	ALA10	PRO15	HY				ALA10	PRO14	HY						
ALA10	MET19	HY	THR8	MET20	SSH	ALA10	ILE16	HY				ALA10	VAL15	HY						
SER11	THR12	MMH	HIS10	GLU14	SSH	ALA10	LEU20	HY				ALA10	MET19	HY						
THR8	MET19	SSH	HIS10	MET20	SSH	PRO13	LEU20	HY				LEU12	MET19	HY						
GLN13	THR12	MSH	GLN13	THR13	SSH	THR11	THR13	MMH				THR11	THR12	MMH						
GLN13	THR12	MSH	GLN13	THR13	SSH							THR8	MET19	SSH						
			GLN13	THR13	SSH															
			HIS10	GLU14	SSH															
			HIS10	GLU14	SSH															
Interaction between VP1 -Motif 10 and VP4-Motif 4																				
															ASP9	ASN16	SSH			
															ASP9	ASN16	SSH			
															ASP9	ASN16	SSH			
															ASP9	ASN16	SSH			
															ASP9	SER17	SSH			
															ASP9	SER17	SSH			
															ASP9	SER17	MSH			
Interaction between VP1 -Motif 28 and VP4-Motif 4																				
																		LYS9	ASP20	Ionic
																		ASP11	SER18	SSH
																		ASP11	SER18	SSH
																		ASP11	SER18	MSH

EV-A			EV-B			EV-C			RV-A			RV-B			FMDV			ThV		
RES	RES	I	RES	RES	I	RES		I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I
Interaction between VP2 -Motif 1 and VP3-Motif 1																				
ARG9	GLY15	MSH	ARG9	GLY15	MSH	ARG10	GLY15	MSH	ARG9	GLY15	MSH	ARG9	GLY15	MSH				ARG9	GLY15	MSH
ARG9	GLY15	MSH	ARG9	GLY15	MSH	ARG10	GLY15	MSH	ARG9	GLY15	MSH	ARG9	GLY15	MSH				ARG9	GLY15	MSH
Interaction between VP2 -Motif 1 and VP3-Motif 2																				
ASN7	CYS16	SSH	ASN7	CYS16	SSH	ASN8	CYS16	SSH	ASN7	CYS16	SSH	ASN7	MET14	SSH				ASN7	PHE15	MSH
ASN7	CYS16	SSH	ASN7	CYS16	SSH	ASN8	CYS16	SSH	ASN7	CYS16	SSH	ASN7	MET14	SSH				ILE5	VAL14	HY
ASN7	CYS16	SSH	ASN7	CYS16	SSH	ASN8	CYS16	SSH	ASN7	CYS16	SSH	ASN7	TYR15	MSH						
ASN7	PHE15	MSH	ASN7	PHE15	MSH	ASN8	PHE15	MSH	ASN7	PHE15	MSH	ASN7	TYR15	MSH						
ASN7	PHE15	MSH	ASN7	PHE15	MSH	ASN8	PHE15	MSH	ASN7	PHE15	MSH									
ILE5	LEU14	HY	TRP5	MET14	HY	ILE6	LEU14	HY	PHE5	MET14	HY	PHE5	MET14	HY						
Interaction between VP2 -Motif 1 and VP3-Motif 3																				
ILE5	PHE2	HY	TRP5	PHE2	HY	ILE6	PHE2	HY	PHE5	PHE2	HY	PHE5	PHE2	HY						
Interaction between VP2 -Motif 1 and VP3-Motif 4																				
PRO19	PRO2	HY	PRO19	PRO2	HY	PRO20	PRO2	HY	PRO19	PRO2	HY	PRO19	PRO2	HY						
						VAL1	ILE14	HY												
Interaction between VP2 -Motif 1 and VP3-Motif 5																				
ARG9	SER2	MSH	ARG9	SER2	MSH	ARG10	SER2	MSH	ARG9	GLY1	MSH	ARG9	GLY1	MSH				ARG9	GLY1	MSH
ARG9	SER2	MSH	ARG9	SER2	MSH	ARG10	SER2	MSH	ARG9	THR2	MSH	ARG9	PRO2	MSH				ARG9	ALA2	MSH
ARG9	ALA5	MSH	ARG9	ALA3	MSH	ARG10	ALA5	MSH	ARG9	THR2	MSH	ARG9	PRO2	MSH				ARG9	ALA2	MSH
ARG9	ALA5	MSH	ARG9	ALA3	MSH	ARG10	ALA5	MSH	ARG9	THR5	MSH							ARG9	ALA3	MSH
									ARG9	THR5	MSH							ARG9	ALA3	MSH
																		ARG9	VAL5	MSH
																		ARG9	VAL5	MSH
Interaction between VP2 -Motif 1 and VP3-Motif 9																				
						VAL1	MET6	HY												
Interaction between VP2 -Motif 2 and VP3-Motif 2																				
			CYS19	CYS16	SSH				THR19	CYS16	SSH							GLY18	THR16	MMH
			CYS19	CYS16	SSH															
Interaction between VP2 -Motif 2 and VP3-Motif 5																				
GLN17	SER2	MSH	LYS14	ALA3	MMH	GLN17	SER2	MSH	GLN17	THR2	MSH	LYS14	ALA3	MMH				PHE15	MET4	HY
LYS14	MET3	MMH	LYS14	MET4	MMH	LYS14	MET3	MMH	LYS14	ALA3	MMH	LYS14	LEU4	MMH				ALA17	ALA2	HY
LYS14	MET4	MMH	PHE15	MET4	HY	LYS14	MET4	MMH	LYS14	ASN4	MSH	PHE15	LEU4	HY				GLN14	ALA3	MMH
PHE15	MET4	HY				PHE15	MET4	HY	LYS14	ASN4	MSH							GLN14	MET4	MMH
Interaction between VP2 -Motif 2 and VP3-Motif 7																				
PHE15	PRO20	HY	PHE15	PRO20	HY	PHE15	PRO20	HY	PHE15	PRO20	HY	PHE15	PRO20	HY						
Interaction between VP2 -Motif 5 and VP3-Motif 4																				
									ALA1	ILE1	HY									
Interaction between VP2 -Motif 5 and VP3-Motif 8																				
															ARG3	ASP11	Ionic			
															ARG3	ASP11	SSH			
															ARG3	ASP11	SSH			
															ARG3	ASP11	SSH			
															ARG3	ASP11	SSH			
Interaction between VP2 -Motif 6 and VP3-Motif 1																				
															LEU15	ILE15	HY			
Interaction between VP2 -Motif 6 and VP3-Motif 8																				
															LEU15	PRO5	HY			
															ILE15	VAL8	HY			
															ASN11	SER8	MMH			
															ASN11	SER8	MSH			
															ASN11	SER8	MSH			
															ASN11	ALA9	MSH			
															ASN11	ALA9	MSH			
															SER13	TYR6	MSH			
Interaction between VP2 -Motif 10 and VP3-Motif 4																				
TYR9	PRO2	HY	TYR9	PRO2	HY	TYR9	PRO2	HY	TYR9	PRO2	HY	TYR9	PRO2	HY						
			VAL11	PRO2	HY				VAL11	PRO2	HY									
Interaction between VP2 -Motif 15 and VP3-Motif 2																				
																		SER1	THR16	SSH

EV-A			EV-B			EV-C			RV-A			RV-B			FMDV			ThV		
RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I
Interaction between VP2 -Motif 3 and VP4-Motif 1																				
ARG4	ASP17	Ionic	ARG4	ASP18	Ionic							ASP1	ASP17	SSH						
ARG4	ASP17	SSH	ARG4	ASP18	SSH							ASP1	ASP17	SSH						
ARG4	ASP17	SSH	ARG4	ASP18	SSH							ASP1	ASP17	SSH						
ARG4	ASP17	Ionic	ARG4	ASP18	Ionic							ASP1	ASP17	SSH						
ARG4	ASP17	SSH	ARG4	ASP18	SSH							ASP1	ASP17	SSH						
ARG4	ASP17	SSH	ARG4	ASP18	SSH							ASP1	ASP17	SSH						
												ASP1	ASP17	SSH						
												ASP1	ASP17	SSH						
												GLN4	ASP17	SSH						
												GLN4	ASP17	SSH						
												GLN4	ASP17	SSH						
												GLN4	ASP17	SSH						
												GLN4	ASP17	SSH						
												GLN4	ASP17	SSH						
												GLN4	ASP17	SSH						
Interaction between VP2 -Motif 10 and VP4-Motif 1																				
ALA5	LYS16	MMH	ASN4	ASP18	MSH	ASN4	ASP18	MSH				ALA5	LYS16	MMH						
ALA5	LYS16	MMH	ASN4	ASP18	MSH	ASN4	ASP18	MSH				ALA5	LYS16	MMH						
ASN4	MET19	SSH	ASN4	ASP18	MSH	ASN4	ASP18	MSH				ASN4	ASP17	SSH						
ASN4	MET19	SSH	ASN4	ASP18	MSH	ASN4	ASP18	MSH				ASN4	ASP17	SSH						
ASN4	ASP17	MSH	ASN4	ASP18	MSH	ASN4	ASP18	MSH				ASN4	ASP17	SSH						
ASN4	ASP17	MSH	ASN4	ASP18	MSH	ASN4	ASP18	MSH				ASN4	ASP17	SSH						
ASN4	LYS16	MSH	TYR9	PHE12	HY	SER5	LYS17	MMH				ASN4	ASP17	MSH						
ASN4	ASP17	MSH	TYR9	PHE12	HY	SER5	LYS17	MMH				ASN4	ASP17	MSH						
ASN4	MET19	SSH	VAL5	LYS17	MMH	TYR9	PHE12	HY				ASN4	ASP17	MSH						
ASN4	MET19	SSH	VAL5	LYS17	MMH	TYR9	PHE12	HY				ASN4	ASP17	SSH						
ASN4	ASP17	MSH	VAL6	PRO15	HY	VAL6	PRO15	HY				ASN4	ASP17	SSH						
ASN4	ASP17	MSH	VAL6	VAL16	HY	VAL6	ILE16	HY				ASN4	ASP17	SSH						
ASN4	LYS16	MSH	VAL6	PRO15	HY	VAL6	PRO15	HY				ASN4	ASP17	SSH						
ASN4	ASP17	MSH	VAL6	VAL16	HY	VAL6	ILE16	HY				ASN4	ASP17	MSH						
GLU11	LYS10	Ionic	VAL7	PRO15	MMH	VAL7	PRO15	MMH				ASN4	ASP17	MSH						
GLU11	LYS10	Ionic	VAL7	PRO15	MMH	VAL7	PRO15	MMH				ASN4	ASP17	MSH						
TYR9	PHE11	HY										ILE6	PRO14	HY						
TYR9	PHE11	HY										ILE6	VAL15	HY						
VAL6	PRO14	HY										ILE6	PRO14	HY						
VAL6	VAL15	HY										ILE6	VAL15	HY						
VAL6	PRO14	HY										TYR9	PHE11	HY						
VAL6	VAL15	HY										TYR9	PHE11	HY						
VAL7	PRO14	MMH										VAL7	PRO14	MMH						
VAL7	PRO14	MMH										VAL7	PRO14	MMH						
Interaction between VP2 -Motif 21 and VP4-Motif 5																				
															TYR7	TRP3	HY			
															TYR9	TRP3	HY			
															TY9	PHE4	HY			

EV-A			EV-B			EV-C			RV-A			RV-B			FMDV			ThV		
RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I	RES	RES	I
Interaction between VP3 -Motif 4 and VP4-Motif 1																				
GLU10	ASP7	MSH	GLU10	ASP8	MSH	GLU10	GLN7	SSH				GLU10	ASP7	MSH						
GLU4	LYS10	Ionic	GLU10	LYS11	Ionic	GLU10	GLN7	SSH				GLU10	LYS10	Ionic						
GLU4	LYS10	MSH	GLU4	LYS11	MSH	GLU10	GLN7	SSH				ILE11	PHE11	HY						
ILE14	PHE11	HY	LYS6	ASP4	Ionic	GLU10	GLN7	SSH				LEU9	MET6	HY						
LEU11	PHE11	HY	LYS6	ASP4	SSH	GLU4	LYS11	Ionic				VAL14	PHE11	HY						
VAL5	PHE11	HY	VAL14	PHE12	HY	GLU4	GLN7	MSH				VAL5	PHE11	HY						
			VAL5	PHE12	HY	GLU4	GLN7	MSH												
						GLU4	LYS11	MSH												
						ILE14	PHE12	HY												
						LEU11	PHE12	HY												
						VAL5	PHE12	HY												
Interaction between VP3 -Motif 6 and VP4-Motif 1																				
			ASP12	ARG2	Ionic															
			ASP12	ARG2	SSH															
			ASP12	ARG2	SSH															
Interaction between VP3 -Motif 6 and VP4-Motif 2																				
PRO16	TYR15	HY	GLN14	ILE13	MSH	PRO16	TYR16	HY	PRO16	PHE15	HY	PRO16	TYR15	HY						
SER15	SER20	MMH	GLN14	ILE13	MSH	GLN14	ILE13	MSH	SER15	SER20	MMH	SER15	SER20	MMH						
GLN14	ILE12	MSH	GLN14	TYR15	MSH	GLN14	ILE13	MSH	GLN14	ASN13	SSH	GLN14	ILE12	MSH						
GLN14	ILE12	MSH	PRO16	TYR16	HY	GLN14	TYR15	MSH	GLN14	ASN13	SSH	GLN14	ILE12	MSH						
CYS17	SER20	MSH							GLN14	ASN13	SSH	SER17	SER20	MSH						
GLN14	TYR14	MSH							GLN14	ASN13	SSH	GLN14	TYR14	MSH						
SER15	SER20	MSH							GLN14	ILE12	MSH	SER15	SER20	MSH						
									GLN14	ILE12	MSH									
									CYS17	SER20	MSH									
									GLN14	TYR14	MSH									
									SER15	SER20	MSH									

3.4 Conclusions

Firstly, the study aimed to identify possible interacting SLiMs which were conserved throughout the individual structural proteins of the *Picornaviridae* family. Furthermore the study aimed to determine the pattern of motif conservation, with an analysis of motifs which were conserved across individual virus species as well motifs which were conserved across different viruses which infect same host species. The motif analysis was performed individually for each structural protein dataset. The results indicated that motif conservation was largely species specific, with no significant conservation observed across the structural proteins of viruses with the same host species. Specifically no correlation between the motifs conserved across the human enteroviruses, human cardioviruses or human cosaviruses was observed. Similar there appeared to be no relation between the motifs conserved across enteroviruses which infect ungulate species, FMDV or the erboviruses. Virus specific conservation was particularly evident in the VP1 dataset, with significant conservation of motifs across sequences of individual subgroups of virus species rather than across the *Picornaviridae* family. A degree of discrepancy with respect to motif conservation across the other protein datasets was observed. The analysis of the VP2, VP3 and VP4 datasets revealed a small proportion of motifs which were significantly more conserved across the *Picornaviridae* family. This study suggests that this is resultant of the increased selective pressure, imposed on the VP1 capsid protein, for host cell specificity. Moreover, the study also suggests that motifs which were observed to be uniquely conserved across individual virus species are more likely to be involved in host cell receptor binding or serve as virus-specific linear B-cell epitope regions. In contrast motifs which were conserved across the strains of several viral species may be more likely to facilitate protein-protein interactions within the capsid subunit-subunit interface. It was also observed that motif conservation did correlate to the phylogenetic groupings which were predicted in Chapter 2. This correlation was observed with respect to all four protein datasets. This was particularly evident in motifs which were highly conserved across and unique to the *Enterovirus* genus, as well as motifs which were conserved across the *Aphthovirus, Cardiovirus*, *Sapelovirus* and *Parechovirus* genera. Furthermore unique conservation was also observed between species of the corresponding phylogenetic out-groups (Chapter 2). The conservation of motifs across virus-subtypes, strains and closely related species may further support the theory of genetic recombination with functional regions of the capsid proteins being interchangeable between virus sub-types. Thus the results support the implications of the monophyletic groupings observed in the phylogenetic analysis in Chapter 2. Specifically, it was suggested that the

monophyletic topology of the genera was indicative of genetic recombination and horizontal gene transfer, rather than vertical evolution from primitive virus strains. As the study primarily aimed to identify interacting motifs within the protomers of picornavirus capsids, structural analysis was performed only with respect to motifs which were significantly conserved across the viral family or genera. Furthermore, structural analysis was performed with respect to representative capsid structures of the virus sub-groups which contained a significant number of sequences. Thus the significant conservation of these motifs may serve as further evidence of their functional importance. More specifically, the study additionally analysed motifs which were significantly conserved across the EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV sub-groups. Representative crystal structures of each sub-group were subjected to the PIC webserver for the prediction of interacting residues. Iterative Python scripted was incorporated to mapped predicted interacting residues to the corresponding residues of highly conserved motifs.

In summary, the study identified seven motifs which were highly conserved and predicted to be interacting within the protomers of all seven representative sub-groups. Moreover five motifs were predicted to be interacting within the protomers of six of the seven representative viruses, while VP2.21 and VP4.4 were predicted to be interacting in both the FMDV and ThV viruses. Furthermore nine conserved motifs were predicted to facilitate subunit interactions across the protomers of the enteroviruses, with a total of four motifs and two motifs predicted to be interacting uniquely within FMDV and ThV respectively. Moreover the study predicted that the assembly of picornavirus protomers is resultant of a network of multiple motif-motif interactions between the independent structural proteins. The study identified the principle interacting motifs as: VP1.1 across all seven representative viruses, VP1.2, VP1.4 and VP1.7 across all representative enteroviruses, VP1.9 in the representative FMDV virus,VP1.28 in the representative ThV virus, VP2.11 across all seven representative viruses, VP2.1 across the enteroviruses and ThV, VP2.6 in the representative FMDV virus, VP3.1-VP3.5 across the enteroviruses and ThV, VP3.8 in the representative FMDV, VP4.2 in the enterovirus representatives and VP4.4 in the FMDV and ThV representatives. Consequent to time constraints only six of these motifs were further analysed with respect to specific residue interactions and conservation. Thus future work could include comprehensive analysis of the remaining motifs.

The study identified the conservation of several predicted motif specific residue interactions, as well principle interacting residues which were significantly conserved across respective

virus sub-groups or the *Picornaviridae* family. With respect to motif VP4.2, the principle interacting residues were predicted as Ile12, Phe14, Phe.Tyr15, Asn17, Ala19 and Ser20. These motifs were also found to be significantly conserved throughout all available VP4 sequences, corresponding to the respective enterovirus sub-groups. The principle interacting residues of motif VP4.4 were identified as: Asn1, Asn17, Ser18 and Asp19 in FMDV, with all residues Asn1 and Ser18 conserved across all available FDMV sequences. While Glu2, Ile18 and Asp19 were predicted as the principle interacting residues within ThV DA, with all residues completely conserved across the ThV sequences.

Further analysis of VP2 motifs focused on motifs 1 and 11. With respect to motif VP2.1 the residues Phe1/Tyr1, Gln4, Asn7, Arg9 and Pro19 were predicted to be involved in multiple interactions within the subunit interface of all enteroviruses and ThV DA. The hydrophobic interaction between Phe1 and VP1.1, with the exception of EV-A where Phe1 was substituted with interacting residue Tyr1, was conserved across the representative enteroviruses and ThV DA. Furthermore the hydrogen bond interaction between Arg9 and VP3.1.Gly15 was also predicted to be completely conserved across the representative enteroviruses and ThV DA. Additionally the hydrophobic interaction between Pro10 and VP3.4.Pro2 was conserved across all representative enteroviruses. Pertaining to the representative FMDV strain, Phe1, Phe5 and Thr9 were predicted as the principle interacting residues. Analysis of motif VP2.11 revealed that the residues Pro6 and Glu7 were conserved sequences throughout all seven sub-groups, with predicted interactions conserved across all seven representative viruses. The side chain-side chain hydrogen bond between Glu7 with VP1.2.Tyr15 was predicted to be conserved across the five representative enteroviruses. While in FMDV Glu7 was predicted to interact with VP1.10.Tyr5 and VP1.28.Tyr7 in ThV DA.

Further analysis of motifs from the VP3 and VP1 subunits was pertained to VP3.1 and VP1.1. With respect to VP3.1, the PIC results predicted Trp12 as interacting within the five enteroviruses as well as Gly15 and Ser18 in the five enteroviruses and ThV DA. The side chain-side chain hydrogen bond between Ser18 and VP1.6.Glu7 was conserved across the five enteroviruses, while the main chain-side chain hydrogen bond between Gly15 with VP2.1.Arg9 was conserved across the enteroviruses as well as ThV DA. Gln17 and Asn17 were predicted as principle interacting residues, with conservation across the enteroviruses and ThV DA respectively. In contrast, Lys2, Cys6 and Ile7 were predicted as the principle interacting residues within FMDV, with Lys2 and Cys6 conserved across all 299 VP3 FMDV sequences. In the analysis of VP1.1, Lys1 was predicted to interact within the protomer of all

representative enteroviruses and ThV DA, while Pro8, Arg9 and Pro10 were predicted as principle interacting residues throughout all the representative structures. The analysis also revealed that His2 was also completely conserved across the sub-groups of the enterovirus genus, while FMDV sequences favoured Arg2 and ThV favoured Lys2, with all of these residues having positively charged side chains which were predicted to have multiple interactions.

As previously stated time constraints did not allow for a comprehensive analysis of all predicted interacting motifs and residues. However plots of residue specific interactions and residue conservation were generated for all highly conserved motifs (Appendices 6 and 7). Thus the study has collected a substantial amount of data for future analysis.

4. Conclusions and Future Work

The overall aim of the study was to broaden the understanding of the evolution and function of the structural proteins across the *Picornaviridae* family. The study had three principle objectives. Firstly a comprehensive analysis of the phylogenetic relationships amongst the individual structural proteins was performed to identify evolutionary patterns across sub-types of individual picornaviruses as well as determine co-host phylogenetic relationships. Correlations and discrepancies in the phylogeny of the independent structural proteins were also identified. Secondly the functions of the structural proteins were further investigated by an exhaustive MEME motif analysis. The conservation of motifs across the viral family, with specific identification of conserved short linear motifs (SLiMs) which may facilitate protein subunit-subunit interactions within the protomer of picornavirus capsids was determined. Motif conservation was assessed across the individual structural proteins of: 1) strains of individual viral species and 2) different viruses which infect the same host species. Thirdly the study predicted specific subunit motif-subunit motif interactions within the protomers of representative virus crystal structures. The study also included further prediction of the principle interacting residues and the corresponding types of interactions. The conservation of these residues across the strains of the respective viruses was calculated. The specific objectives included an *in silico* prediction of interacting residues within representative PDB files, the mapping of predicted residues to conserved motifs and the *in silico* analysis of interacting motifs within the subunit-subunit interface of representative crystal structures.

The phylogenies inferred with respect to the four structural proteins were found to have a substantial amount of congruency. As observed for each of the capsid proteins, the representative sequences of the cardioviruses, cosaviruses, aphthoviruses, erboviruses, teschoviruses and hunnivirus clustered as an out-group distinctively separate from a super-grouping comprised of sequences of the sapelovirus and bat and feline picornaviruses and the enteroviruses. Furthermore, a close relation between the cosavirus and cardiovirus proteins was observed across all capsid phylogenies. Similarly the aphthoviruses and teschoviruses and hunniviruses consistently clustered together while the clustering of the sapelovirus sequences with the unclassified bat, feline and pigeon picornaviruses was also observed across the respective subunit phylogenies. In regard to the VP1 and VP3 sequences, direct correlation was observed with respect to the common grouping of the LV and PeV sequences,

as well as the close relation between the AV, Salivirus and TV sequences. Viruses of the *Enterovirus* genus always formed a complex cluster, with sub-groupings representative of individual viral types. This group was consistently observed to have closest relation to the SV and bat, feline and pigeon picornaviruses. The topology inferred for all four of the capsid proteins was found to be directly dependent on viral type and genus and no significant host co-phylogeny observed. However the sub-clustering of virus subtypes according to host species was observed. Furthermore, no correlation was observed between viral pathogenicity and sub-clustering of the enterovirus VP1 sequences. Thus the study did not identify a correlation between sub-types and strains which are causative of the same diseases. The enteroviruses have the ability to infect multiple cell-types, as seen in PV with initial entry via the gastrointestinal or respiratory tract and subsequent infection cells of the central nervous system (Racaniello *et al*., 2007). Thus it is speculated that the variation in the symptoms induced by enterovirus infections may be dependent on the host's individual susceptibility to systemic infections.

Analysis of the external clustering within the main sub-groups revealed significant discrepancies in the monophyletic clustering across the four capsid proteins. Incongruences across the topologies with respect to closely related viral species, viral sub-types or viral strains were observed, with greatest disagreement pertaining to the enteroviruses. The study concluded that such incongruences may be indicative of inter-typic recombination of closely related picornavirus species. Thus the study supports findings by Heath *et al* (2006) that the viral capsid proteins may be functionally interchangeable across closely related FMDV species and further proposes that this might by archetypal of many picornaviruses. The monophyletic clustering within genera may also be supportive of horizontal gene transfer. Furthermore the topologies inferred in this study were not consistent with those of the replication proteins, as reported by Phelps *et al* (2013). Therefore the findings in this study support the notation that the capsid proteins are evolving independently from the replication proteins. It is suggested that since capsid proteins are more exposed to evolutionary pressures to evade host immune systems with simultaneous pressure to retain selectivity for the host cell receptor, they appear to have the ability to evolve without comprising host cell attachment. However the viral replication proteins may be more inclined to remain conserved, such that mutations do not comprise virus replication.

The phylogenetic reconstruction with respect to each of the different capsid proteins was found to be a complex procedure. Significantly low bootstrap values were observed, particularly in the VP1 and VP2 datasets. This iterates the limitations of phylogenetic analysis, with most current models representing general protein databases. As the picornavirus proteins are subject to the high mutation rate and genetic drift of RNA viruses, their evolutionary pattern may be unique and thus cannot be described according to current available models. Furthermore the continuous infection of new hosts, derivation of quasispecies and genetic recombination impose evolutionary pressures unique to these viruses. Thus the study may provide evidence that for accurate phylogenetic reconstruction of picornavirus proteins, the development of a highly specific evolutionary model is required as current models appear to be inapplicable to the evolution of these proteins.

The function of the structural proteins was further investigated through an exhaustive motif analysis which was performed individually for each structural protein dataset. Motif conservation was largely species specific, with no significant conservation observed across the structural proteins of viruses with the same host species. Significant conservation of motifs across sequences of individual subgroups of virus species rather than across the *Picornaviridae* family was particularly observed in the VP1 dataset. This study suggests that this is resultant of the increased selective pressure, imposed on the VP1 capsid protein, for host cell specificity. The analysis of the VP2, VP3 and VP4 datasets revealed a small proportion of motifs which were significantly more conserved across the *Picornaviridae* family. The study suggests that motifs which were observed to be uniquely conserved across individual virus species are more likely to be involved in host cell receptor binding or serve as virus-specific linear B-cell epitope regions, while motifs which were conserved across the strains of several viral species may be more likely to facilitate protein-protein interactions within the capsid subunit-subunit interface. However further analysis, incorporating the prediction and mapping of linear B-cell epitopes is required to support this suggestion.

Motif conservation correlated to the phylogenetic groupings with respect to all four protein datasets. Particular correlation was observed in motifs which were highly conserved across, and unique to, the *Enterovirus* genus as well as motifs which were conserved across the *Aphthovirus*, *Cardiovirus*, *Sapelovirus* and *Parechovirus* genera. The conservation of motifs across virus-subtypes, strains and closely related species may further support the theory of genetic recombination with functional regions of the capsid proteins possibly being interchangeable between virus sub-types. Specifically, it was suggested that the monophyletic

topology of the genera was indicative of genetic recombination and horizontal gene transfer, rather than vertical evolution from primitive virus strains (Heath *et al*., 2006).

As the study primarily aimed to identify interacting motifs within the protomers of picornavirus capsids, structural analysis was performed with respect to motifs which were significantly conserved across the viral family or genera. Moreover, structural analysis was performed with respect to representative capsid structures of the virus sub-groups which contained a significant number of sequences. Therefore the significant conservation of these motifs may serve as additional support of their functional importance. Additional analysis was performed with respect to the sub-groups: EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV sub-groups. Representative crystal structures of each sub-group were subjected to the PIC webserver for the prediction of interacting residues. Iterative Python scripted was incorporated to mapped predicted interacting residues to the corresponding residues of highly conserved motifs.

The interacting residues were mapped against a total of 41 conserved motifs. The study identified 33 motifs which may facilitate interactions between the subunits of capsid protomers. Highly conserved motifs which were not predicted to play a role in protomer assembly, may facilitate interactions between other capsid intermediates, thus may facilitate protomer-protomer or pentamer-pentamer interactions. Currently only the individual protomers, consisting of a single copy of each VP1, VP2, VP3 and VP4, have been crystallised. Thus further investigation into motifs involved in pentamer and procapsid assembly may be facilitated by complex homology modelling, incorporation with protein docking, of the 14S pentamer of different picornavirus capsids. Moreover, virus assembly is believed to be dependent on the interaction with host cellular proteins, particularly molecular chaperones (Geller *et al*., 2012). Therefore it is possible that these motifs are conserved for functional interaction with host cellular proteins. Similar this suggestion could be further investigated by means of homology modelling and protein docking as well as *in vitro* protein-protein interaction experiments.

The findings in this study suggest that picornavirus protomer formation is reliant on a network of multiple interactions between the capsid subunit proteins. Interacting motifs within these subunit proteins were identified to be highly conserved, particularly across representative the enteroviruses. Moreover the specific subunit motif-subunit motif interactions were also found to be conserved. Thus the study identified protagonist subunit-

subunit interactions which were predicted to play a critical role in the assembly of capsid protomers. In summary the study identified the principle interacting motifs as: VP1.1 across all seven representative virus species, VP1.2, VP1.4 and VP1.7 across all representative enteroviruses, VP1.9 in the representative FMDV virus,VP1.28 in the representative ThV virus, VP2.11 across all representative viruses, VP2.1 across the enteroviruses and ThV, VP2.6 in the representative FMDV virus, VP3.1-VP3.5 across the enteroviruses and ThV, VP3.8 in the representative FMDV, VP4.2 in the enterovirus representatives and VP4.4 in the FMDV and ThV representatives. Although comprehensive residue analysis was limited to six motifs, the study also identified principle partners of interacting residues which were 1) highly conserved across the *Picornaviridae* family, 2) highly conserved across the viruses of the *Enterovirus* genus or 3) highly conserved across sequences of individual virus species of EV-A, EV-B, EV-C, RV-A, RV-B, FMDV and ThV. The conservation of interacting motifs as well as the conservation of principle residues amongst closely related species may provide further support of inter-typic genetic recombination. Moreover the findings may suggest that the structural proteins are co-evolving to preserve the ability of capsid assembly though the network of interactions between these subunits. As previously stated this evolution appears to be independent from viral replication proteins.

This study has been based on *in silico* predictions. Thus the importance of the interacting motifs and residues identified in this study is subject to confirmation. Further *in silico* analysis could involve protein docking experiments, the prediction of positively selected residues and an analysis of co-evolving residues within the subunit proteins of picornavirus capsids. Furthermore *in vitro* protein-protein interactions, with the incorporation of site-directed mutagenesis, may serve to confirm principle interacting residues required for protomer assembly. As previously stated time constraints did not allow for a comprehensive analysis of all predicted interacting motifs and residues. However plots of residue specific interactions and residue conservation were generated for all highly conserved motifs (Appendices 6 and 7). The structural mapping of identified motifs was also limited by the availability of crystal structures. Thus further analysis could involve the mapping of generated homology models. Moreover, the study also identified motifs which were uniquely conserved across specific viral species. The function of such motifs could be further investigated through exhaustive structural analysis and examination of current motif databases. Heatmaps which identify motif conservation or absence in specific sequences of picornavirus species were also generated (Appendix 5). Thus future work involving a

comprehensive comparison of motif conservation across individual viral strains of the same species could further the understanding of the evolution of the capsid proteins amongst specific picornavirus species. This analysis may also serve to identify novel drug targets within individual virus species which could effectively treat specific picornavirus infections.

The findings in this study have served to broaden the understanding of the evolution of picornavirus structural proteins, with suggestions of independent evolution from the replication proteins through possible inter-typic recombination of functional protein regions. The study has also expanded the theories of the mechanisms responsible for virus assembly. The findings indicate that protomer assembly may be facilitated through a network of multiple subunit-subunit interactions, which have been predicted to be highly conserved across capsid proteins of closely related virus species. As capsid assembly is an integral process in the viral life-cycle, the principle interacting motifs may serve as novel drug targets for the antiviral treatment of picornavirus infections. Thus the findings in the study may be fundamental to the development of treatments which are more economically feasible or clinically effective than current vaccinations.

References

Acheson, N.H. (2007). Fundamentals of Molecular Journal of Virology, pp .169-180, John Wiley and Sons, United States.

Adachi, J. and M. Hasegawa. (1996). MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Computer Science Monographs of Institute of Statistical Mathematics* **28**,1-150.

Adachi, J., P. Waddell, W. Martin, and M. Hasegawa. (2000). Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *Journal of Molecular Evolution* **50**, 348-358.

Agol, V.I. (2001). Picornavirus genome: an overview. In: Semler, B., Wimmer, E. (Eds.), *Molecular Biology of Picornaviruses*, pp. 127–148, ASM Press, Washington, DC..

Amos, W., and J. Harwood. (1998). Factors affecting levels of genetic diversity in natural populations. *Philosophical Transactions of the Royal Society London* **353**,177-186.

Bachrach, H.L., Moore, D.M., McKercher, P.D. and Polatnick, J. (1975). Immune and antibody responses to an isolated capsid protein of foot-and-mouth disease virus. *Journal of Immunology* **115**, 1636-164l.

Bailey, T.L., Bodén, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching, *Nucleic Acids Research* **37**,W202-W208.

Bailey, T.L and Charles, E. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California.

Bailey, T.L and Gribskov, M. (1998). Methods and statistics for combining motif match scores. *Journal of Computational Biology* **5**, 211-221.

Bast, F. (2013). Sequence Similarity Search, Multiple Sequence Alignment, Model Selection, Distance Matrix and Phylogeny Reconstruction. *Nature Protocol Exchange*. doi:10.1038/protex.2013.065.

Bittle, J.L., Houghten, R.A., Alexander, H., Shinnink, J., Sutcliffe, G.J., Lerner, R., Rowlands, D. and Brown, F. (1982). Protection against foot and mouth disease by immunization with chemically synthesized peptides predicted from viral nucleotide sequence. *Nature* **298**, 30-33.

Bodewes, R., Ruiz-Gonzalez, A., Schapendonk, C.M., van den Brand, J.M., Osterhaus, A.D. and Smits, S.L. (2014). Viral metagenomic analysis of feces of wild small carnivores. *Journal of Virology* **11** (89).

Boros, A., Nemes, C., Pankovics, P., Kapusinszky, B., Delwart, E. and Reuter, G. (2013). Genetic characterization of a novel picornavirus in turkeys (Meleagris gallopavo) distinct from turkey galliviruses and megriviruses and distantly related to the members of the genus Avihepatovirus. *Journal of Virology*. **94**, 1496–1509.

Cameron, K., Zhang, X., Seal, B., Rodriguez, M., and Njenga, M.K. (2000). Antigens to viral capsid and non-capsid proteins are present in brain tissues and antibodies in sera of Theiler's virus-infected mice. *Journal of Virological Methods* **91**,11-19.

Cho, A. (2012). Constructing Phylogenetic Trees Using Maximum Likelihood. *Scripps Senior Theses*. Paper **46**.

Collen, T., Dimarchi, R. and Doel, T.R. (1991). A T cell epitope in VP1 of foot- and-mouth disease virus is immunodominant for vaccinated cattle. *Journal of Immunology* **146(2)**, 749-755.

Cooper, P.D., Agol, V.I., Bachrach, H.L., Brown, F., Ghendon, Y., Gibbs, A.J., Gillespie, J.H., Lonberg-Hnlm, K., Mandel, B., Melnick, J.L., Mohanty, S.B., Povey, R.C., Rueckert, R.R., Schafter, F.L. and Tyrrell, D.A.J. (1978) Picornaviridae: second report. *International Journal of Virology* **10**,165-180.

Couderc, T., Delpeyroux, J., Le Blay, H. and Blondel, B. (1996). Mouse adaptation determinants of poliovirus type 1 enhance viral uncoating. *Journal of Journal of Virology* **70**, 305-312.

Cunningham, C.W., Omland, K.E. and Oakley, T.H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution* **13**, 361-66.

Daleno, C., Piralla, A., Scala, A., Senatore,L., Principi, N. and Esposito, S. (2013). Phylogenetic analysis of human rhinovirus isolates collected from otherwise healthy children with community-acquired pneumonia during five successive years. *PLoS ONE* **8** p. e80614

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H. and Gibson, T. J. (2012). Attributes of short linear motifs. *Molecular BioSystems* **8**(1), 268–281.

Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt. (1978). A model of evolutionary change in proteins. In Atlas of protein sequence and structure. Vol. 5, Suppl. 3, pp. 345-352, National Biomedical Research Foundation, Washington, D.C.

Diella, F., Haslam, N. and Chica, C. (2008). Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Frontiers in Bioscience* **13**: 6580–603.

Dimmic, M.W., Rest, J.S., Mindell, D.P. and Goldstein, D. (2002). RArtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *Journal of Molecular Evolution* **55,** 65-73.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797.

Ehrenfeld, E., Domingo, E. and Roos, R.R. (2010). The Picornaviruses, pp 5-20, ASM Press, Washington, DC, USA.

Eun, H.M., Bae, Y.S., and Yoon, J.W. (1988). Amino acid differences in capsid protein, VP1, between diabetogenic and nondiabetogenic variants of encephalomyocarditis virus. *Journal of Virology* **163**, 369-373.

Farris, J.S. (1973). On the use of the parsimony criterion for inferring phylogenetic trees. *Systematic Zoology* **22**, 250–256.

Felsenstein, J. (1973). Maximum likelihood and minimumsteps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology* **22**, 240–249.

Gall, O.L., Christian, P., Fauquet, C.M., King, A.M.Q, Knowles, N.J., Nakashima, N., Stanway, G. and Gorbalenya, A.E. (2007). Picornavirales, a proposed order of positive-sense single-starned RNA vriuses with a pseudo-T =3 virion architecture, Spinger publishers, United Kingdom.

Geller, R., Taguwa, S. and Frydman, J. (2012). Broad action of Hsp90 as a host chaperone required for viral replication**.** *Biochimica et Biophysica Acta* **1823**, 698–706.

Grant, R.A., Filma, D.J., Fujinami, R.S., Icenogle, J.P. and Hogle, J.M. (1992). Three-dimensional structure of Theiler virus. *Proceedings of the National Academy of Science* **89**, 2061–65.

Hales, L.M., Knowles, N.J., Reddy, P.S., Xu, L., Hay, C. and Hallenbeck, P.L. (2008). Complete genome sequence analysis of Seneca Valley virus-001, a novel oncolytic picornavirus. *Journal of Virology* **89**, 1265–1275.

Harris, K.S., Xiang, W., Alexander, L., Lane, W.S., Paul, A.V. and Wimmer, E. (1994). Interaction of poliovirus polypeptide 3CDpro with the 5' and 3' termini of the poliovirus genome. Identification of viral and cellular cofactors needed for efficient binding. *Journal of Biological Chemistry* **269**, 27004-27014.

Hadfield, A.T., Lee, W.M., Zhao, R., Oliveira, M.A. and Minor, I. (1997). The refined structure of human rhinovirus 16 at 2.15Å resolution: implications for the viral life cycle. *Structure* **5**, 427–41.

Heath, L., E. van der Walt, A. Varsani, and D. P. Martin. (2006). Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *Journal of Virology* **80**, 11827–11832.

Hu, L., Zhang, Y., Hong, M., Zhu, S., Yan, D., Wang, D., Li, X., Zhu, Z., Tsewang, Z. and Xu, W. (2014). Phylogenetic evidence for multiple intertypic recombinations in enterovirus B81 strains isolated in Tibet, China. *Scientific Reports* **4**, 6035.

Hughes, A.L. (2004). Phylogeny of the Picornaviridae and differential evolutionary divergence of picornavirus proteins. *Infection, Genetics and Evolution* **4**, 143–152.

Johansson, S., Niklasson, B., Maizel, J., Gorbalenya, A.E. and Lindberg, A.M. (2002). Molecular analysis of three Ljungan virus isolates reveals a new close-to-root lineage of the Picornaviridae with a cluster of two unrelated 2A proteins. *Journal of Virology* **76**, 8920–8930.

Jones, D.T., W. R. Taylor, and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput.Appl*.Biosci**. 8,** 275-282.

Kapoor, A., Victoria, J., Simmonds, P., Wang, C., Shafer, R.W., Nims, R., Nielsen, O. and Delwart, E. (2008). A highly divergent picornavirus in a marine mammal. *Journal of Virology* **82**, 311–320.

Kaaden, O.R., Adam, K.H. and Stroi-Imaier, K. (1977). Induction of neutralizing antibodies and immunity in vaccinated guinea pigs by cyanogen-bromide-peptides of VP3 of foot-and-mouth disease virus. *Journal of Virology* **34**, 397-400.

King, A.M.Q., Brown, F., Christian, C., Hovi, T., Hyypia, T., Knowles, N.J., Lemon, S.M., Minor, P.D., Palmenberg, A.C., Skern, T. and Stanway, G. Picornaviridae. In: Van Regenmortel, M.H.V., Fauquet, C.M., Bishop, D.H.L., Calisher, C.H., Carsten, E.B., Estes, M.K., Lemon, S.M., Maniloff, J., Mayo, M.A.,McGeoch, D.J., Pringle, C.R., Wickner, R.B. (2000). Virus taxonomy. Seventh report of the International Committee for the Taxonomy of Viruses, pp. 657–678, Academic Press, New York.

Kirkegaard, K. (1990). Mutations in VP1 of poliovirus specifically affect both encapsidation and release of viral RNA. *Journal of Virology* **64**,195-206.

Knowles, N.J., Samuel, A.R. (2003). Molecular epidemiology of foot-and mouth disease virus. *Virus Reserach* **91**, 65-80.

Knox, C., Moffat, K., Ali, S., Ryan, M. and Wileman, T. (2005). Foot-and-mouth disease virus replication sites form next to the nucleus and close to the Golgi apparatus, but exclude marker proteins associated with host membrane compartments. *Journal of Virology* **86**, 687–696

Lau, S.K., Woo, P.C., Lai, K.K., Huang, Y. and Yip, C.C. (2011). Complete genome analysis of three novel picornaviruses from diverse bat species. *Journal of Virology* **85**, 8819–8828.

Le, S.Q. and Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular Biology and Evolution* **25**,1307-1320.

Lewis-Rogers, N., Crandall, K.A. (2009). Evolution of Picornaviridae: An examination of phylogenetic relationships and cophylogeny. *Molecular Phylogenetics and Evolution* **10**, 015.

Lim, E.S., Cao, S., Holtz, L.R., Antonio, M., Stine, O.C. and Wang, D. (2014). Discovery of rosavirus 2, a novel variant of a rodent-associated picornavirus, in children from The Gambia. *Journal of Virology* **454**, 25-33.

Lindberg, M. A., Andersson, P., Savolainen, C., Mulders, M. N. and Hovi, T. (2003). Evolution of the genome of Human enterovirus B: incongruence between phylogenies of the VP1 and 3CD regions indicates frequent recombination within the species. *Journal of Virology* **84**, 1223–1235.

Liu, X., Wang, Y., Zhang, Y., Fang, Y., Lu, J., Zhoe. P., Zhang, Z., Qi-wei, C., Wang, G., Wang, J., Lou. H., and Jiang, S. (2011). Cloning, codon optimization and homology modelling of structural protein VP1 from foot and mouth disease virus. *African Journal of microbiology research* **5**,486-495.

Lukashev AN., Shumilina EY., Belalov IS., Ivanova OE., Eremeeva TP., Reznik VI., Trotsenko OE., Drexler JF and Drosten, C. (2014). Recombination strategies and evolutionary dynamics of the Human enterovirus A global gene pool.

Miller, S.T., Hogle, J.M. and Filman, D.J. (2001). Ab initio phasing of high-symmetry macromolecular complexes: successful phasing of authentic poliovirus data to 3.0 A resolution. *Journal of Molecular Biology* **307**,499–512.

Murray, L., Luke, G., Ryan, M. D., Wileman, T. and Knox, C. (2009). Amino acid substitutions within the 2C coding sequence of Theiler's Murine Encephalomyelitis virus alter virus growth and affect protein distribution. *Virus Research* **144**, 74-82.

Neduva, V. and Russell, R.B. (2006). Peptides mediating interaction networks: new leads at last. *Current Opinion in Biotechnology* **17**(5), 465–71.

Nitayaphan, S., Toth, M. M. and Roos, R.P. (1985). Localization of a neutralization site of Theiler's murine encephalomyelitis viruses. *Journal of Virology* **56**, 887-895.

Oberste, M.S., Maher, K., Michele, S.M., Belliot, G., Uddin, M. and Pallansch, M.A. (2005). Enteroviruses 76, 89, 90, and 91 represent a novel group within the species Human enterovirus A. *Journal of Virology* **86**, 445–451.

Pei, J., Kim, B.-H. and Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research*, **36**(7), 2295–300.

Pellequer, J., Westhof, E., and Van Regenmortel, M. (1993). Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunology Letters* **36,** 83-99.

Phelps, N.B.D., Mor, S.K., Armien, A.G., Batts, W. and Goodwin, A.E. (2013) Isolation and Molecular Characterization of a Novel Picornavirus from Baitfish in the USA. *PLoS ONE* **9**(2), e87593.

Piralla, A., Baldanti, F., and Gerna, G. (2011)**.** Phylogenetic patterns of human respiratory picornavirus species, including the newly identified group C rhinoviruses, during 1-year surveillance of a hospitalized patient population in Italy. *Journal of Clinical Microbiology* **49**, 373–6.

Posada, D. and Crandall, K.A. (2001).Selecting the best-fit model of nucleotide substitution. *Systematic Biology* **50**,580-601.

Porta, C., Kotecha, A., Burman, A., Jackson, T., Ren, J., Loureiro, S.,Jones, I. M., Fry, E. E., Stuart, D. I. and Charleston, B. (2013). Rational engineering of recombinant picornavirus capsids to produce safe, protective vaccine antigen. *PLoS Pathogens* **9**, e1003255.

Racaniello, V.R. (2007). Picornaviridae: The viruses and their replication.In: Fields, B.N., Knippe, D.M., Chanock, R.M., Melnick, J.L., Rosizman, B. and Shope, R.E. (2007). *Fundamental Journal of Virology*, pp 357-390, New York Raven Press.

Rueckert, R. R. (2001). Picornaviridae: the viruses and their replication (Knipe, D. M. and Howley, P. M., 4th edn), pp. 685–715. Lippincott–Raven, New York.

Rodrigo, M.J and Dopazo, J. (1995). Evolutionary analysis of picornavirus family, *Journal of Molecular Evolution* **3**, 93662-3671.

Rossman, M. G., A. Arnold, J. W. Erickson, E. A. Frankenberger, J. P. Griffith, H.-J. Hecht, J. E. Johnson, G. Kamer, M. Luo, A. G. Mosser, R. R. Rueckert, B. Sherry, and G. Vriend. (1985). Structure of a human common cold virus and functional relationship to other picornaviruses. *Nature* **317,** 145–153.

Schrodinger LLC, 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

Simmonds, P. and Welch. J. (2006). Frequency and dynamics of recombination within different species of human enteroviruses. *Journal of Virology* **80,** 483–493.

Smura, T., Blomqvist, S., Vuorinen, T., Ivanova, O., Samoilovich, E., Al-Hello, H., Savolainen-Kopra, C., Hovi, T. and Roivainen, M. (2014). The Evolution of Vp1 Gene in Enterovirus C Species Sub-Group That Contains Types CVA-21, CVA-24, EV-C95, EV-C96 and EV-C99**.** *PLoS ONE* **9**, e94579.

Steel, M. and Penny, D. (200). Parsimony likelihood, and the role of models in molecular phylogenetics. *Molecular Biology and Evolution* **17**, 839–850.

Tina, K. G., Bhadra, R. and Srinivasan, N.. (2007). PIC: Protein Interactions Calculator, *Nucleic Acids Research* **35**, Web Server issue W473–W476.

Toyoda, H., Nicklin, M. J. H., Murray, M. G., Anderson, C. W., J.J Dunn, Studier, F. W. & Wimmer E. (1986). A second virus-encoded proteinase involved in proteolytic processing of poliovirus polyprotein. *Cell* **45**, 761-770.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0**.** *Molecular Biology and Evolution* **30,** 2725-2729.

Varrasso, A., Drummer, H. E., Huang, J. A., Stevenson, R. A., Ficorilli, N., Studdert, M. J. and Hartley, C. A. (2001). Sequence conservation and antigenic variation of the structural proteins of equine rhinitis A virus. *Journal of Virology* **75**, 10550–10556.

Van Phan, L., Tung, N., Kwang-Nyeong, L., Young-Joon, K., Hyang-Sim, L., Van Cam, N., Thuy Duong, M., Thi Hoa, D., Su-Mi, K., In-Soo, C. and Jong-Hyeon, P. (2010). Molecular characterization of serotype A foot-andmouth disease viruses circulating in Vietnam in 2009. *Veterinary Microbiology* **144**, 58-66.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution* **18**,691-699.

Wu, C.N., Lin, Y.C., Fann, C., Liao, N.S., Shih, S.R. and Ho, M.S. (2001). Protection against Enterovirus 71 infection in newborn mice by passive immunization with subunit VP1 vaccines and inactivated virus. *Vaccine* **20**, 895–904.

Yang, Z., R. Nielsen, and M. Hasegawa. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Molecular Biology and Evolution* **15**,1600-1611.

Yoder, J.D., Cifuente, J.O., Pan, J., Bergelson, J.M., and Hafenstein, S. (2012). The crystal structure of a coxsackievirus B3-RD variant and a refined 9-angstrom cryo-electron microscopy reconstruction of the virus complexed with decay-accelerating factor (DAF) provide a new footprint of DAF on the virus surface. *Journal of Virology* **86**,12571–12581.

Yausch, R.L., Kerekes, K., Saujani, K., and Kim, B.S. (1995). Indentification of major T-cell epitope within VP3 amino acid residues 24 to 37 of Theiler's virus in demyelination-susceptible SJL/J mice. *Journal of Virology*. **69**, 7315-7318.

Zhao, R., Pevear, D. C., Kremer, M. J., Giranda, V. L.,Kofron, J. A., Kuhn, R. J. and Rossman, M. G. (1996). Human rhinovirus 3 at 3.0 A resolution. Structure. *PubMed* **4**(10), 1205-20.

Zocher, G., Mistry, N., Frank, M., Hähnlein-Schick, I. and Ekström, J.I. (2014) A Sialic Acid Binding Site in a Human Picornavirus. *PubMed* **10**(10), e1004401.

Appendix 1

1.1 Extract.py

```
import os

#Parse VirusHostList

hostfile = open("/home/caroline/PROJECT/Genomes/VirusHostList.txt",'r')
host_dic={}
for virus in hostfile:
	org = virus.split("|")[0]
	host = virus.split("|")[1].rstrip()
	if host=="Unknown":
		host = "Unknown Host"
	host_dic.update({org:host})
hostfile.close()
host_keys = host_dic.keys()




#folders = os.listdir("/home/caroline/PROJECT/Hosts")
#for folder in folders:
	#files = os.listdir("/home/caroline/PROJECT/Hosts/")
	#for fl in files:
f = open("/home/caroline/PROJECT/AllViruses/All/AllViruses.fasta",'r')
sequences =
{"VP0":[],"VP1":[],"VP2":[],"VP3":[],"VP4":[],"L":[],"2A":[],"2B":[],"2C":[],"3A":[],"3B":[
],"3C":[],"3D":[]}
keys = sequences.keys()
line = f.readline()

while line!='':
	if line[0]=='>':
		organism_start = line.find("Organism:")
		organism_end = line.find("|",organism_start)
		organism=line[organism_start+9:organism_end]
		pro_name_st = (line.find("Gene Symbol:"))
		if pro_name_st ==-1:
			pro_name_st = line.find("Protein Name:")+13
		else:
			pro_name_st+=12

		protein_name = line[pro_name_st:].rstrip()
		if protein_name =="3B(VPg)":
			protein_name="3B"
		if protein_name=="Lab":
			protein_name="L"
```

```
            if protein_name=="1D" or protein_name=="1D(VP1)":
                protein_name = "VP1"
            if protein_name=="1B" or protein_name=="1B(VP2)":
                protein_name = "VP2"
            if protein_name=="1A" or protein_name=="1A(VP4)":
                protein_name = "VP4"
            if protein_name=="1C" or protein_name=="1C(VP3)":
                protein_name = "VP3"

            if protein_name in keys:
                for key in host_keys:
                    if organism in key:
                        fasta = ('>'+organism+'|'+host_dic[key]+"|"+protein_name).replace("
","")+'\n'
                #fasta = ('>'+organism+'|'+protein_name).replace(" ","")+'\n'

                line = f.readline()
                while not line.isspace():
                    fasta=fasta+line
                    line=f.readline()
                sequences[protein_name].append(fasta)


        line=f.readline()
f.close()


for k in keys:
        #fname=f1.split('.')[0]
        fname="AllViruses_"+k+".fasta"
        #w=open("/home/caroline/PROJECT/Viruses/"+folder+"/"+fname,'w')
        w=open("/home/caroline/PROJECT/AllViruses/Proteins/"+fname,'w')
        data=sequences[k]
        w.writelines(data)
        w.close()
```

1.2 Filter.pyimport os

```
import copy
from Bio import pairwise2
from Bio.SubsMat import MatrixInfo as matlist

matrix = matlist.blosum62
gap_open = -3
gap_extend = -0.5

ifile = open(#read in fasta file,'r')

fname = #replace with file name
lines = ifile.readlines()
ifile.close()

seqs_key = ""
sequence = ""
seqs_dict = {}
copy_dict = {}

total_seqs = 0

for line in lines:                                    #makes a dictionary containing each
sequence in a multiple sequence alignment,
                                                      # indexed by the sequence header
(assumes fasta format).
   if line.startswith(">"):
         total_seqs += 1
         if seqs_key != "" and sequence != "":
            seqs_dict[seqs_key] = sequence
            sequence = ""   #originally not indented
         seqs_key = line.rstrip()


   else:
         sequence += line.strip()

if seqs_key != "" and sequence != "":
   seqs_dict[seqs_key] = sequence
   sequence = ""   #originally not indented



copy_dict =seqs_dict
final_dict = copy.deepcopy(seqs_dict)
removed=[]
a = ""
b = ""
c = 0
id_count = 0.0
```

```
id_dict={}

seqs_id = 0.0
id_string = ""


all_total = 0
for i in seqs_dict:
   if i in removed:

         continue
   for k in copy_dict:

         rev_seq = "%s,%s" % (k, i)
         if rev_seq in id_dict or k == i or k in removed:
            continue

         all_total += 1

         seqI = seqs_dict[i]
         seqK = seqs_dict[k]

         seq_ex_gapsI=''
         for base in seqI:
            if base!='-':
                  seq_ex_gapsI=seq_ex_gapsI+base
         seq_ex_gapsK=''
         for base in seqK:
            if base!='-':
                  seq_ex_gapsK=seq_ex_gapsK+base

         alns = pairwise2.align.globalds(seq_ex_gapsI, seq_ex_gapsK, matrix, gap_open,
gap_extend)

         top_aln = alns[0]
         aln_seqI, aln_seqK, score, begin, end = top_aln

         for j in range(len(aln_seqI)):
                                                #a and b represent residues the same position in
each of every combination of a pair
                                                # of aligned sequences (supposedly present in a
multiple sequence alignment).
            a = aln_seqI[j].lower()
            b = aln_seqK[j].lower()

            if a != "-" and b != "-":
                  if a == b:
                     id_count += 1.0

            if a != "-" or b != "-":    #keeps track of the length of the alignment, rather than the
```

```
longest sequence
                    c += 1


            seqs_id = (id_count/c)*100
            id_string = "%s,%s" % (i, k)
            #print c, seqs_id, id_string
            id_dict[id_string] = seqs_id

                                        #resets id and len counters
            id_count = 0.0
            c = 0

            print i,k
            if seqs_id>=100.00:
               if k in final_dict:
                    del final_dict[k]
                    removed.append(k)




w = open(#open new file,'w')
for key in final_dict:
   seq= final_dict[key]
   w.write(key+"\n")
   fasta=""
   for j in range(len(seq)):
            fasta+=seq[j]
            if (j+1)%60==0:
               fasta+='\n'
   w.write(fasta+"\n")
w.close()
```

1.3 SequenceHeader.py

```
import os

ifile = open(#read in fasta file,'r')
lines = ifile.readlines()
ifile.close()
data=[]

viruses =
{'teschovirus':'TeschV|','sapelovirus':'SV|','turdivirus':'TV|','ljunganvirus':'LV|','humanparecho
virus':'HPeV|','encephalomyocarditisvirus':'EMCV|','hepatitisavirus':'HAV|','foot-
mouthdiseasevirus':'FMDV|','foot-and-
mouthdiseasevirus':'FMDV|','theilovirus':'ThV|','enterovirus':'EV|','equinerhinitisavirus':'RAV|'
,
'equinerhinitisbvirus':'RBV|','rhinitisavirus':'RAV|','rhinitisbvirus':'RBV|','kobuvirus':'AV|','Aic
hivirus':'AV|','picornavirus':'PiV|','rhinovirus':'RV|','pasivirus':'PaV|','cosavirus':'CoSV|','rafivir
us':'RfV|',
'ovinehungarovirus':'','hunnivirus':'HuV|','encephalomyelitisvirus':'EMV|','porcineenterovirus':
'PEV|','aichivirus':'AV|','fatheadminnowpicornavirus':'FMiPV|','sicinivirus':'SiV'}

host_names =
{'Avian':'A|','Bat':'B|','Bovine':'Bo|','Human':'H|','Porcine':'P|','Pigeon':'Pi|','Equine':'E|','Feline':'
F|','Simian':'S|','Canine':'C|','Caprine':'Cp|','Tortoise':'Ts|','Seal':'Sl|'}

hosts =
{'UnknownHost':'U|','Swine':'P|','Cattle':'Bo|','Buffalo':'Bf|','Sheep':'O|','Bat':'B|','Human':'H|','P
ig':'P|','Pigeon':'Pi|','Mouse':'M|','Rat':'R|','Tick':'Tc|','Tiger':'Tg|','Turkey':'Tk|','Thrush':'Th|','Cat
':'F|','Dog':'C|','Goat':'Cp|','Alpaca':'Al|','Boar':'Br|','Cow':'Bo|','Bovine':'Bo|','Minnow':'Mi|','Chi
mpanzee':'Cz|','Monkey':'Mk|','Chicken':'Ck|','Horse':'E|','Avian':'A','Tortoise':'Ts|','Seal':'Sl|','Si
mian':'S'}


for line in lines:
    checked = False
    if line.startswith(">"):
            host = line.split('|')[1]
            if host =="UnknownHost":
                for k in host_names:
                        if k in line.split('|')[0]:
                            line = line.split('|')[0]
                            line = line.replace(k,host_names[k])
                            checked=True
            if not checked:
                host = hosts[host]
                line = line.split('|')[0].lstrip('>')
                line = '>'+host+line

            try:
                virus = line.split('|')[1]
```

```
except IndexError:
   print line
   print host
#Remove host names from virus name and abbreviates heading
for h in host_names:
   if h in virus:
         virus=virus.replace(h,'')

virus = virus.lower()
for v in viruses:
   if v in virus:
         virus = virus.replace(v,viruses[v])
         break
virus = virus.replace('ohuv','OHUV')
virus = virus.replace('iaioPiV','IaioPiV')
virus = virus.replace('bhcosv-b1','BHCoSV-B1')
virus = virus.replace('dhcosv-d1','DHCoSV-D1')
virus = virus.replace('ehcosv-e1','EHCoSV-E1')
virus = virus.replace('/homosapiens/','')
virus = virus.replace('fuyang.anhui.p.r.c/17.08','fu.an.08')
virus = virus.replace('/hokkaido.jpn/','/jpn/')
virus = virus.replace('queenmary/hongkong','hk')
virus = virus.replace('/shenzhen/08/china/hfmd/2008','')
virus = virus.replace('/shenzhen/08/china/hfmdfatal/2008','')
virus = virus.replace('/jingdezhen/china/hfmd_severe/2011','')
virus = virus.replace('/shenzhen/08/china/hfmd/2008','')
virus = virus.replace('//shenzhen/08/china/hfmdsevere/2008','')
virus = virus.replace('/gx/chn/2001','')
virus = virus.replace('/ningbo.chn/065/2010','chn/065/2010')
virus = virus.replace('bht-lykh202f/xj/chn/2011','bht-lykh202f')
virus = virus.replace('bhtps-mjh21f/xj/chn/2011','bhtps-mjh21f')
virus = virus.replace('bhtps-mklh04f/xj/chn/2011','bhtps-mklh04f')
virus = virus.replace('bhtyt-arl-afp02f/xj/chn/2011','bhtyt-arl-afp02f')
virus = virus.replace('bhtyt-arlh403f/xj/chn/2011','bhtyt-arlh403f')
virus = virus.replace('becho30/zhejiang/17/03/csf','becho30/zhejiang')
virus = virus.replace('cht-xebgh09f/xj/chn/2011','cht-xebgh09f')
virus = virus.replace('ckssc-alxhh01f/xj/chn/2011','ckssc-alxhh01f')
virus = virus.replace('sichuan/chn/2011','')
virus = virus.replace('/sd/chn/sewage','')
virus = virus.replace('fujian/93-8;chn-','')
virus = virus.replace('guangdong/92-2;chn-','')
virus = virus.replace('hainan/93-2;chn-','')
virus = virus.replace('hebei/91-2;chn-','')
virus = virus.replace('henan/91-3;chn-','')
virus = virus.replace('jiangxi/89-1;chn-','')
virus = virus.replace('yunnan/92;chn-','')
virus = virus.replace('/shenzhen/08/china/hfmdsevere/2008','')
virus = virus.replace('PEV|bovine/tb4-oev/2009/hun','PEV|tb-4')
virus = virus.replace('PEV|bswine/k23/2008/hun','PEV|b-k23')
virus = virus.replace('/pocheon/001/kor/2010','/pocheon/1')
```

```
        virus = virus.replace('lindholm1.3,pak3/2006','lindholm1.3')
        virus = virus.replace('asia1/jiangsu/china/2005','jiangsu')
        virus = virus.replace('asia1/bam/afg/l-590/2009','/bam/afg/l-590')
        virus = virus.replace('o1/bfs1860/uk/67(iah1)','o1//uk/67(iah1)')
        virus = virus.replace('o1/bfs1860/uk/67(iah2)','o1/uk/67(iah2)')
        virus = virus.replace('o1/bfs1860/uk/67(mah)','o1/uk/67(mah)')
        virus = virus.replace('persistent','')
        virus = virus.replace('murchisonfallsnationalpark','UK')
        virus = virus.replace('/2009/hunOHUV1/2009/hun','')
        virus = virus.replace('wildboar/wb2c-tv/2011/hun','wb2c-tv')
        virus = virus.replace('miniopterusschreibersiiPiV|1unknown-
jq814851','MiniPiV|jq814851')
        virus = virus.replace('/sichuan/chn/2012','')
        virus = virus.replace('bgal-7/2010/hungary','bgal-7')
        virus = virus.replace('unknown-','')
        virus = virus.replace('dog/an211d/usa/2009an211d','an211d')
        virus = virus.replace('swine/s-1-hun/2007/hungaryswine/s-1-hun/2007/hungary','s-1-
hun')


        line = line.split('|')[0]+"|"+virus+'\n'
        if len(line)>30:
           print line + '|'+str(len(line))

   data.append(line)

w = open(#open new file, w")
w.writelines(data)
w.close()
```

1.4 MotifConservation.py

```
import matplotlib

import matplotlib.pyplot as plt
import numpy as np
from matplotlib import cm as CM

mast = #read in MAST file
meme = #read in MEME file
out_path = #define path to save output to


def mast_list(path):
    fmast=open(path,'r')
    lines = fmast.readlines()
    fmast.close()
    for l in lines:
            if "Removing motifs" in l:
                bad_motifs = l[17:-34]
                bad_motifs = bad_motifs.replace(',','').replace('and ','').split()
            if "No overly similar pairs" in l:
                bad_motifs = []

    return bad_motifs

def parse_meme(meme):
    imeme = open(meme,'r')
    lines = imeme.readlines()
    t_set = lines[:lines.index("COMMAND LINE SUMMARY\n")]
    m_set = lines[lines.index("COMMAND LINE SUMMARY\n"):]
    imeme.close()
    return(t_set,m_set)


def data_groups(training_set):
    host_groups = {}
    virus_groups= {}

    L = iter(training_set)
    line = L.next() #goes to first line of file

    #Get all sequence names
    while not "Sequence name            Weight" in line:
            line = L.next()
    line = L.next()
    line = L.next()

    while
```

```
line !="**********************************************************************
***********\n":
        seq1 = line[:25].strip()
        seq2 = line[40:65].strip()
        host1=seq1.split('|')[0]
        virus1=seq1.split('|')[1]
        if virus1=='EV' or virus1=='RV':
           vtype = ((seq1.split('|')[2])[0]).upper()
           if vtype=='A' or vtype=='B' or vtype=='C':
                virus1=virus1+vtype
           else:
                if virus1=="EV":
                   virus1="Other EV"
                else:
                   virus1="Other RV"
        if virus1 in virus_groups:
           virus_groups[virus1]+=[seq1]
        else:
           virus_groups[virus1]=[seq1]
        if host1 in host_groups:
           host_groups[host1]+=[seq1]
        else:
           host_groups[host1]=[seq1]

        if seq2:
           host2=seq2.split('|')[0]
           virus2=seq2.split('|')[1]
           if host2 in host_groups:
                host_groups[host2]+=[seq2]
           else:
                host_groups[host2]=[seq2]
           if virus2=='EV' or virus2=="RV":
                vtype = ((seq2.split('|')[2])[0]).upper()
                if vtype=='A' or vtype=='B' or vtype=='C':
                   virus2=virus2+vtype
                else:
                   if virus2=="EV":
                        virus2="Other EV"
                   else:
                        virus2="Other RV"
           if virus2 in virus_groups:
                virus_groups[virus2]+=[seq2]
           else:
                virus_groups[virus2]=[seq2]

        line = L.next()
   return(host_groups, virus_groups)


def motif_info(motif_set):
```

```
    #Get motif info
    motifs = []
    motif_names=[]
    motif_indexes = [i for i, line in enumerate(motif_set) if "MOTIF" in line and "E-value" in
line]
    start_of_sequences = [i+4 for i, line in enumerate(motif_set) if "sites sorted by position p-
value" in line]
    for k,i in enumerate(motif_indexes,1):
            mots = []
            line = motif_set[i]
            e_value = line[line.index("E-value = ")+10:].strip().split('e')
            e = float(e_value[0])*10**float((e_value[1]))
        if e>0.05 or str(k) in bad_motifs:
                continue
            m_name=str(k)
            motif_names.append(m_name)
            start = start_of_sequences[k-1]
            end = motif_set.index("--------------------------------------------------------------------------------
-----\n",start)
            sequence_block = motif_set[start:end]
            for seq in sequence_block:
                p_value = seq[32:42].strip().split('e')
                p = float(p_value[0])*10**float((p_value[1]))
                if p>0.05:
                        continue
                mots+=[seq[:25].strip()]
        motifs+= [mots]
    return(motifs,motif_names) #returns a list of a list and a single list of names

def edit_host_groups(hosts):
    #Group hosts into larger sub-groups

    primates = ['H','S','Mk','Cz']
    ungulates = ['Bo','P','O','Cp','Bf','Al','Br','E']
    felines = ['F','Ti']
    avians = ['A','Pi','Tk','Ck']
    murines = ['R','M']
    #uncommon = ['Sl','Ti','Th','Mi']

    main_hosts=
{"Primates":[],"Ungulates":[],"Feline":[],"Bat":[],"Avian":[],"Murine":[],"Canine":[],"Uncom
mon Hosts":[]}

    for h in hosts:
            if h=="U":
                continue
            if h =="B":
                main_hosts["Bat"]+=hosts[h]
            else:
                if h=='C':
```

```
                main_hosts["Canine"]+=hosts[h]
            else:
                if h in primates:
                    main_hosts["Primates"]+=hosts[h]
                else:
                    if h in ungulates:
                        main_hosts["Ungulates"]+=hosts[h]
                    else:
                        if h in felines:
                            main_hosts["Feline"]+=hosts[h]
                        else:
                            if h in avians:
                                main_hosts["Avian"]+=hosts[h]
                            else:
                                if h in murines:
                                    main_hosts["Murine"]+=hosts[h]
                                else:
                                    main_hosts["Uncommon Hosts"]+=hosts[h]


#Remove empty host lists
    hosts={}
    for h in main_hosts:
        if main_hosts[h]:
            hosts.update({h:main_hosts[h]})
    return hosts

#def edit_viruses_groups(viruses):



def map_data(data, motifs):
    map_data={}
    for group in data:
        sub_groups = {}
        sequences = data[group]

        #makes sub-plot for larger groups
        num_sbplts = len(sequences)/50
        if len(sequences)%50 != 0:
            num_sbplts+=1
        for i in range(num_sbplts):
            sg_data=[]
            if i != num_sbplts-1:
                sub_sequences = sequences[i*50:(i+1)*50]
            else:
                sub_sequences = sequences[i*50:]

            for m in motifs: #m is a list of sequences with that motif
                m_data=[]
```

```
                count = 0
                for seq in m: #checks if seq is in the group of data
                    if seq in sequences:
                        count+=1.0
                conservation = count/len(sequences)

                for s in sub_sequences:
                    if s in m:
                        m_data+=[conservation]
                    else:
                        m_data+=[0]

                sg_data+=[m_data]
            sub_groups[i+1]=sg_data

        map_data[group]=sub_groups
    return map_data

def overall_conservation(data,motifs,name):
    overall_map = {name:[]}
    labels={name:data.keys()}
    for m in motifs:
        m_conserved = []
        for group in labels[name]:
            sequences = data[group]
            count=0.0
            for seq in sequences:
                if seq in m:
                    count+=1.0
            conservation = count/len(sequences)
            m_conserved.append(conservation)
        overall_map[name].append(m_conserved)
    return(labels,overall_map)


def plot_maps(g_map,m_labels,seqs,sizex,sizey,out_path):
    for g in g_map:
        sub_plts = g_map[g]
        num_sbplts = len(sub_plts)
        for i,sg in enumerate(sub_plts): #plot for each sub group
            sub_map = sub_plts[sg]
            plt.close("all")
            column_labels = m_labels
            if i!=num_sbplts-1:
                row_labels = seqs[g][i*50:(i+1)*50]
            else:
                row_labels = seqs[g][i*50:]
            data = np.array(sub_plts[sg])

            fig, ax = plt.subplots()
```

```
        colors = [('white')] + [(CM.jet(i)) for i in xrange(40,250)]
        new_map = matplotlib.colors.LinearSegmentedColormap.from_list('new_map',
colors, N=300)
        heatmap = ax.pcolor(data, cmap=new_map)
        fig = plt.gcf()
        fig.set_size_inches(12,22)
        fig.subplots_adjust(bottom=0.00,top=0.85,left=0.04,right=1)



        #turn off the frame
        ax.set_frame_on(False)


        # put the major ticks at the middle of each cell
        ax.set_yticks(np.arange(data.shape[0])+0.5, minor=False)
        ax.set_xticks(np.arange(data.shape[1])+0.5, minor=False)

        # want a more natural, table-like display
        ax.invert_yaxis()
        ax.xaxis.tick_top()

        ax.set_xticklabels(row_labels,fontsize=sizex, minor=False)
        ax.set_yticklabels(column_labels,fontsize=sizey, minor=False)

        # rotate the
        plt.xticks(rotation=90)
        ax.grid(False)

        # Turn off all the ticks
        ax = plt.gca()

        for t in ax.xaxis.get_major_ticks():
            t.tick1On = False
            t.tick2On = False
        for t in ax.yaxis.get_major_ticks():
            t.tick1On = False
            t.tick2On = False

        cbar = plt.colorbar(heatmap,orientation="horizontal")
        cbar.ax.set_xlabel('# of motif sites/total # of sequences',fontsize=20)
        plt.savefig(out_path+g+str(sg)+".png",dpi=300)
        plt.show()


def plot_overall(g_map,m_labels,seqs,sizex,sizey,out_path):
    for g in g_map:
        column_labels=m_labels
        row_labels = seqs[g]
        data = np.array(g_map[g])
```

```
        fig, ax = plt.subplots()
        colors = [('white')] + [(CM.jet(i)) for i in xrange(40,250)]
        new_map = matplotlib.colors.LinearSegmentedColormap.from_list('new_map',
colors, N=300)
        heatmap = ax.pcolor(data, cmap=new_map)
        fig = plt.gcf()
        fig.set_size_inches(10,15)
        fig.subplots_adjust(bottom=0.00,top=0.90,left=0.04,right=1)

        #turn off the frame
        ax.set_frame_on(False)


        # put the major ticks at the middle of each cell
        ax.set_yticks(np.arange(data.shape[0])+0.5, minor=False)
        ax.set_xticks(np.arange(data.shape[1])+0.5, minor=False)

        # want a more natural, table-like display
        ax.invert_yaxis()
        ax.xaxis.tick_top()

        ax.set_xticklabels(row_labels,fontsize=sizex, minor=False)
        ax.set_yticklabels(column_labels,fontsize=sizey, minor=False)

        # rotate the
        plt.xticks(rotation=90)
        ax.grid(False)

        # Turn off all the ticks
        ax = plt.gca()

        for t in ax.xaxis.get_major_ticks():
            t.tick1On = False
            t.tick2On = False
        for t in ax.yaxis.get_major_ticks():
            t.tick1On = False
            t.tick2On = False

        cbar = plt.colorbar(heatmap,orientation="horizontal")
        cbar.ax.set_xlabel('# of motif sites/total # of sequences',fontsize=20)
        plt.savefig(out_path+g+".png",dpi=300)
        plt.show()


data = parse_meme(meme)
bad_motifs = mast_list(mast)
print bad_motifs


training_set = data[0]
```

```
motif_set=data[1]

groups=data_groups(training_set)

hosts = edit_host_groups(groups[0])
viruses = groups[1]

motif_data = motif_info(motif_set)
motifs = motif_data[0]
m_labels = motif_data[1]



'''v_labels = All_V[0]
all_v_map = All_V[1]
plot_overall(all_v_map,m_labels,v_labels,14,12,out_path+"/Viruses/")

All_H = overall_conservation(hosts,motifs,"All_Hosts")
h_labels = All_H[0]
all_h_map = All_H[1]
plot_overall(all_h_map,m_labels,h_labels,14,12,out_path+"/Hosts/")


virus_map = map_data(viruses,motifs)
host_map = map_data(hosts,motifs)

plot_maps(host_map,m_labels,hosts,12,9,out_path+"/Hosts/")
plot_maps(virus_map,m_labels,viruses,12,9,out_path+"/Viruses/")'''


#12 and 9 font size for vp1,2,3
#12 and 12 for vp4
```

1.5 ProtomerInterface.py

```
import pylab as pl

import numpy as np
Virus = ‘FMDV’ #[Change as required]
sequence_id = 'U|FMDV|a22iraq' #[Change as required]
name = "FMDV A22-Iraq " #[Change as required]

# conserved motifs identified by heat maps [Change as required]
vp1 = ['1','9','10']
vp2 = ['1','5','6','11','21','22']
vp3 = ['1','2','3','9']
vp4 = ['4','5']

# protein interactions file:
p = open('/home/caroline/PROJECT/StructuralMapping/'+Virus+'/PIC','r')
pic = p.readlines()
p.close()

#parsing PIC file into different types of interactions
start = pic.index('Hydrophobic Interactions within 5 Angstroms\n')+2
end = pic.index('\n',start+1)
hydrophobic = pic[start:end]

start = pic.index('Protein-Protein Main Chain-Main Chain Hydrogen Bonds\n')+3
end = pic.index('\n',start+1)
MCMC = pic[start:end]

start = pic.index('Protein-Protein Side Chain-Side Chain Hydrogen Bonds\n')+3
end = pic.index('\n',start+1)
SCSC = pic[start:end]

start = pic.index('Protein-Protein Main Chain-Side Chain Hydrogen Bonds\n')+3
end = pic.index('\n',start+1)
MCSC = pic[start:end]

start = pic.index('Ionic Interactions within 6 Angstroms\n')+2
end = pic.index('\n',start+1)
ionic = pic[start:end]


#motif files for each protein
m = open('/home/caroline/PROJECT/AllViruses/VP1/MEME/meme.txt','r')
meme1 = m.readlines()
m.close()
m = open('/home/caroline/PROJECT/AllViruses/VP2/MEME/meme.txt','r')
meme2 = m.readlines()
m.close()
```

```
m = open('/home/caroline/PROJECT/AllViruses/VP3/MEME/meme.txt','r')
meme3 = m.readlines()
m.close()
m = open('/home/caroline/PROJECT/AllViruses/VP4/MEME/meme.txt','r')
meme4 = m.readlines()
m.close()


#gets the exact residues of the motif for the specific structural sequence
def get_block(meme,mot,seq):
    start = meme.index("          Motif "+mot+" in BLOCKS format\n")+3
    end = meme.index("//\n",start+1)
    blocks = meme[start:end]
    block = ""
    for b in blocks:
        if b.split(' ')[0] == seq:
            block=b[33:].split(" ")[0]
    return block

#gets the exact postions of the residues of the motif in a specifc sequence
def get_res_pos(meme,mot,seq):
    start = meme.index("          Motif "+mot+" in BLOCKS format\n")+3
    info = meme[start-1]
    width =int((info[info.index("width=")+6:info.index("width=")+9].strip().split(' '))[0])
    end = meme.index("//\n",start+1)
    blocks = meme[start:end]
    residues = []
    for b in blocks:
        if b.split(' ')[0] == seq:
            res1 = int(b[26:31].strip())
            for i in range(width):
                res = res1+i
                residues.append(res)
    return residues


#Instantiates Dictionaries of the residues of each motif in each chain
Res_VP1={}
Res_VP2={}
Res_VP3={}
Res_VP4={}

for m in vp1:
    residues = get_res_pos(meme1,m,sequence_id)
    Res_VP1[m]=residues
for m in vp2:
    residues = get_res_pos(meme2,m,sequence_id)
    Res_VP2[m]=residues
for m in vp3:
    residues = get_res_pos(meme3,m,sequence_id)
```

```
    Res_VP3[m]=residues
for m in vp4:
    residues = get_res_pos(meme4,m,sequence_id)
    Res_VP4[m]=residues

All_res = {'A':Res_VP1,'B':Res_VP2,'C':Res_VP3,'D':Res_VP4}

# Methods to determine which motifs from which proteins were predicted to interact by PIC
def interactions(interaction,t,mot_residues,m1,c1,c2):
    motifs = [] #returns motifs which are interacting from chain1 and chain2
    positions = [] #returns positions of the interacting residues
    residues = [] #returns residues of the interacting residues
    if t == 'h' or t== 'i': #Position tables in PIC file vary by type of interaction
        pos1 = int(interaction[0].strip())
        res1 = (interaction[1].strip())
        chain1 = interaction[2].strip()
        pos2 = int(interaction[3].strip())
        res2 = (interaction[4].strip())
        chain2 = interaction[5].strip()
    else:
        pos1 = int(interaction[0].strip())
        chain1 = interaction[1].strip()
        res1 = (interaction[2].strip())
        pos2 = int(interaction[4].strip())
        chain2 = interaction[5].strip()
        res2 = (interaction[6].strip())

    if pos1 in mot_residues and chain1==c1 and chain2==c2:
        chain2_motifs = All_res[chain2] #gets a dictionary of motifs corresponding to the chain
with which the residue is interacting

        #calculate position of residue in motif:
        mot_pos1 = (pos1-int(mot_residues[0]))+1
        res1=res1+str(mot_pos1)
        for m2 in chain2_motifs:
            mot2_residues = chain2_motifs[m2]
            if pos2 in mot2_residues:
                #calculate position of residue in motif:
                mot_pos2 = (pos2-int(mot2_residues[0]))+1
                res2=res2+str(mot_pos2)
                motifs.append(m1)
                motifs.append(m2)
                positions.append(str(pos1))
                positions.append(str(pos2))
                residues.append(res1)
                residues.append(res2)

    else:
        if pos2 in mot_residues and chain2==c1 and chain1==c2:
            chain1_motifs = All_res[chain1] #gets a dictionary of motifs corresponding to the
```

```
chain with which the residue is interacting
            #calculate position of residue in motif:
            mot_pos2 = (pos2-int(mot_residues[0]))+1
            res2=res2+str(mot_pos2)
            for m2 in chain1_motifs:
               mot2_residues = chain1_motifs[m2]
               if pos1 in mot2_residues:
                  #calculate position of residue in motif:
                  mot_pos1 = (pos1-int(mot2_residues[0]))+1
                  res1=res1+str(mot_pos1)
                  motifs.append(m1)
                  motifs.append(m2)
                  positions.append(str(pos2))
                  positions.append(str(pos1))
                  residues.append(res2)
                  residues.append(res1)

   return (motifs,residues,positions)


def
map_interface(mot_residues,hydrophobic,ionic,MCMC,SCSC,MCSC,interacting_table,intera
cting_motifs,c1,c2,detailed_residue_interactions):
   for line in hydrophobic:
      interaction = line.split('\t')
      i_info = interactions(interaction,'h',mot_residues,m1,c1,c2)
      i_motifs = i_info[0]
      i_res = i_info[1]
      i_pos = i_info[2]
      if i_motifs:
         if not i_motifs[0] in interacting_motifs[c1]:
            interacting_motifs[c1].append(i_motifs[0])
         if not i_motifs[1] in interacting_motifs[c2]:
            interacting_motifs[c2].append(i_motifs[1])
         if not 'VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n' in interacting_table:
            interacting_table.append('VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n')

detailed_residue_interactions.append('VP'+c1+'\t'+i_motifs[0]+'\t'+i_res[0]+'\t'+i_pos[0]+'\tV
P'+c2+'\t'+i_motifs[1]+'\t'+i_res[1]+'\t'+i_pos[1]+'\tHydrophobic\n')

   for line in ionic:
      interaction = line.split('\t')
      i_info = interactions(interaction,'i',mot_residues,m1,c1,c2)
      i_motifs = i_info[0]
      i_res = i_info[1]
      i_pos = i_info[2]
      if i_motifs:
         if not i_motifs[0] in interacting_motifs[c1]:
            interacting_motifs[c1].append(i_motifs[0])
         if not i_motifs[1] in interacting_motifs[c2]:
```

```
            interacting_motifs[c2].append(i_motifs[1])
        if not 'VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n' in interacting_table:
            interacting_table.append('VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n')

detailed_residue_interactions.append('VP'+c1+'\t'+i_motifs[0]+'\t'+i_res[0]+'\t'+i_pos[0]+'\tV
P'+c2+'\t'+i_motifs[1]+'\t'+i_res[1]+'\t'+i_pos[1]+'\tIonic\n')

    for line in MCMC:
        interaction = line.split('\t')
        i_info = interactions(interaction,'p',mot_residues,m1,c1,c2)
        i_motifs = i_info[0]
        i_res = i_info[1]
        i_pos = i_info[2]
        if i_motifs:
            if not i_motifs[0] in interacting_motifs[c1]:
                interacting_motifs[c1].append(i_motifs[0])
            if not i_motifs[1] in interacting_motifs[c2]:
                interacting_motifs[c2].append(i_motifs[1])
            if not 'VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n' in interacting_table:
                interacting_table.append('VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n')

detailed_residue_interactions.append('VP'+c1+'\t'+i_motifs[0]+'\t'+i_res[0]+'\t'+i_pos[0]+'\tV
P'+c2+'\t'+i_motifs[1]+'\t'+i_res[1]+'\t'+i_pos[1]+'\tMCMC H-Bond\n')

    for line in SCSC:
        interaction = line.split('\t')
        i_info = interactions(interaction,'p',mot_residues,m1,c1,c2)
        i_motifs = i_info[0]
        i_res = i_info[1]
        i_pos = i_info[2]
        if i_motifs:
            if not i_motifs[0] in interacting_motifs[c1]:
                interacting_motifs[c1].append(i_motifs[0])
            if not i_motifs[1] in interacting_motifs[c2]:
                interacting_motifs[c2].append(i_motifs[1])
            if not 'VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n' in interacting_table:
                interacting_table.append('VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n')

detailed_residue_interactions.append('VP'+c1+'\t'+i_motifs[0]+'\t'+i_res[0]+'\t'+i_pos[0]+'\tV
P'+c2+'\t'+i_motifs[1]+'\t'+i_res[1]+'\t'+i_pos[1]+'\tSCSC H-Bond\n')

    for line in MCSC:
        interaction = line.split('\t')
        i_info = interactions(interaction,'p',mot_residues,m1,c1,c2)
        i_motifs = i_info[0]
        i_res = i_info[1]
        i_pos = i_info[2]
        if i_motifs:
            if not i_motifs[0] in interacting_motifs[c1]:
                interacting_motifs[c1].append(i_motifs[0])
```

```
        if not i_motifs[1] in interacting_motifs[c2]:
            interacting_motifs[c2].append(i_motifs[1])
        if not 'VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n' in interacting_table:
            interacting_table.append('VP'+c1+'\t'+i_motifs[0]+'\tVP'+c2+'\t'+i_motifs[1]+'\n')

detailed_residue_interactions.append('VP'+c1+'\t'+i_motifs[0]+'\t'+i_res[0]+'\t'+i_pos[0]+'\tV
P'+c2+'\t'+i_motifs[1]+'\t'+i_res[1]+'\t'+i_pos[1]+'\tMCSC H-Bond\n')
    return(interacting_table,interacting_motifs,detailed_residue_interactions)



#determine list of interacting motifs
interacting_motifs = {'A':[],'B':[],'C':[],'D':[]}
interacting_table=['Protein\tMotif\tProtein\tMotif\n']
detailed_residue_interactions=['Protein\tMotif\tResidue\tPostion\tProtein\tMotif\tResidue\tPo
stion\tInteraction\n']
for m1 in vp1:
    mot_residues = Res_VP1[m1]
    c1 ='A'
    c2 ='B'
    interface_interactions =
map_interface(mot_residues,hydrophobic,ionic,MCMC,SCSC,MCSC,interacting_table,intera
cting_motifs,c1,c2,detailed_residue_interactions)
    interacting_table = interface_interactions[0]
    interacting_motifs = interface_interactions[1]
    detailed_residue_interactions = interface_interactions[2]
for m1 in vp1:
    mot_residues = Res_VP1[m1]
    c1 ='A'
    c2 ='C'
    interface_interactions =
map_interface(mot_residues,hydrophobic,ionic,MCMC,SCSC,MCSC,interacting_table,intera
cting_motifs,c1,c2,detailed_residue_interactions)
    interacting_table = interface_interactions[0]
    interacting_motifs = interface_interactions[1]
    detailed_residue_interactions = interface_interactions[2]
for m1 in vp1:
    mot_residues = Res_VP1[m1]
    c1 ='A'
    c2 ='D'
    interface_interactions =
map_interface(mot_residues,hydrophobic,ionic,MCMC,SCSC,MCSC,interacting_table,intera
cting_motifs,c1,c2,detailed_residue_interactions)
    interacting_table = interface_interactions[0]
    interacting_motifs = interface_interactions[1]
    detailed_residue_interactions = interface_interactions[2]
for m1 in vp2:
    mot_residues = Res_VP2[m1]
    c1 ='B'
    c2 ='C'
```

```
    interface_interactions =
map_interface(mot_residues,hydrophobic,ionic,MCMC,SCSC,MCSC,interacting_table,intera
cting_motifs,c1,c2,detailed_residue_interactions)
    interacting_table = interface_interactions[0]
    interacting_motifs = interface_interactions[1]
    detailed_residue_interactions = interface_interactions[2]
for m1 in vp2:
    mot_residues = Res_VP2[m1]
    c1 ='B'
    c2 ='D'
    interface_interactions =
map_interface(mot_residues,hydrophobic,ionic,MCMC,SCSC,MCSC,interacting_table,intera
cting_motifs,c1,c2,detailed_residue_interactions)
    interacting_table = interface_interactions[0]
    interacting_motifs = interface_interactions[1]
    detailed_residue_interactions = interface_interactions[2]
for m1 in vp3:
    mot_residues = Res_VP3[m1]
    c1 ='C'
    c2 ='D'
    interface_interactions =
map_interface(mot_residues,hydrophobic,ionic,MCMC,SCSC,MCSC,interacting_table,intera
cting_motifs,c1,c2,detailed_residue_interactions)
    interacting_table = interface_interactions[0]
    interacting_motifs = interface_interactions[1]
    detailed_residue_interactions = interface_interactions[2]


#Makes tables in text file
w = open('/home/caroline/PROJECT/StructuralMapping/'+Virus+'/mapping.txt','w')
w.writelines(interacting_table)
w.writelines(detailed_residue_interactions)
w.write('Protein\tInteracting Motifs\n')

chains = ["A","B","C","D"]
for chain in chains:
    motifs = interacting_motifs[chain]
    motifs_ints=[]
    for m in motifs:
        motifs_ints.append(int(m))
    motifs_ints.sort()
    w.write(chain)
    for m in motifs_ints:
        w.write('\t'+str(m))
    w.write('\n')
w.close()

#The next section of the script analyses the motifs which have been identified as interacting
in more detail.
```

```
def count_res_interactions(protein,pos,hydrophobic,ionic,MCMC,SCSC,MCSC):
    count=0
    for line in hydrophobic:
        interaction = line.split('\t')
        pos1 = int(interaction[0].strip())
        chain1 = interaction[2].strip()
        pos2 = int(interaction[3].strip())
        chain2 = interaction[5].strip()
        if (pos==pos1 and protein ==chain1) or (pos==pos2 and protein ==chain2):
            count+=1
    for line in ionic:
        interaction = line.split('\t')
        pos1 = int(interaction[0].strip())
        chain1 = interaction[2].strip()
        pos2 = int(interaction[3].strip())
        chain2 = interaction[5].strip()
        if (pos==pos1 and protein ==chain1) or (pos==pos2 and protein ==chain2):
            count+=1
    for line in MCMC:
        interaction = line.split('\t')
        pos1 = int(interaction[0].strip())
        chain1 = interaction[1].strip()
        pos2 = int(interaction[4].strip())
        chain2 = interaction[5].strip()
        if (pos==pos1 and protein ==chain1) or (pos==pos2 and protein ==chain2):
                count+=1
    for line in SCSC:
        interaction = line.split('\t')
        pos1 = int(interaction[0].strip())
        chain1 = interaction[1].strip()
        pos2 = int(interaction[4].strip())
        chain2 = interaction[5].strip()
        if (pos==pos1 and protein ==chain1) or (pos==pos2 and protein ==chain2):
                count+=1
    for line in MCSC:
        interaction = line.split('\t')
        pos1 = int(interaction[0].strip())
        chain1 = interaction[1].strip()
        pos2 = int(interaction[4].strip())
        chain2 = interaction[5].strip()
        if (pos==pos1 and protein ==chain1) or (pos==pos2 and protein ==chain2):
                count+=1
    return count

for protein in interacting_motifs:
    meme=''
    if protein == 'A':
        meme = meme1
    else:
        if protein == 'B':
```

```
            meme = meme2
        else:
            if protein == 'C':
                meme=meme3
            else:
                if protein=='D':
                    meme=meme4
    motifs = interacting_motifs[protein]
    residue_dic = All_res[protein]
    for m in motifs:
        block = get_block(meme,m,sequence_id)
        positions = residue_dic[m]
        residue_count = []
        for pos in positions:
            count = count_res_interactions(protein,pos,hydrophobic,ionic,MCMC,SCSC,MCSC)
            residue_count.append(count)

        #plotting graph for motif showing number of interactions per residue
        fig = pl.figure()
        fig.set_size_inches(12,3)
        fig.subplots_adjust(bottom=0.3,top=0.95,left=0.06,right=1)
        ax = pl.subplot(111)
        width=0.8
        ax.bar(range(len(block)), residue_count, width=width)
        ax.set_xticks(np.arange(len(block)) + width/2)
        ax.set_xticklabels(block,fontsize=12.5)
        # Turn off all the ticks
        ax = pl.gca()
        for t in ax.xaxis.get_major_ticks():
            t.tick1On = False
            t.tick2On = False
        for t in ax.yaxis.get_major_ticks():
            t.tick1On = False
            t.tick2On = False

        pl.xlabel(name+'- '+"VP"+protein+": Motif "+m, fontsize=14)
        pl.ylabel('Number of interactions', fontsize=12)

pl.savefig('/home/caroline/PROJECT/StructuralMapping/'+Virus+'/'+protein+'_'+m+'.png',dpi
=1200)
```

1.6 PymolMapping.py

```
from pymol import cmd


def map_cbv_conserved(chain,residues,colors):
    for i, res in enumerate(residues):
        if res!="*":
            cmd.select('m', '(chain %s) and (resi %s)' %((chain),(res)))
            col = colors[i]
            cmd.color(col,'m')


def get_resi(motifs,chain):
    residues=[]
    f = open('/home/caroline/PROJECT/AllViruses/'+chain+'/MEME/meme.txt','r')
    data = f.readlines()
    f.close()
    for mot in motifs:
        if mot != '*':
            m_id="Motif "+mot
            for i,line in enumerate(data):
                if m_id+" in BLOCKS format" in line:
                    info = data[i+2]
                    width =int((info[info.index("width=")+6:info.index("width=")+9].strip().split('
'))[0])
                    seqs = int(info[info.index("seqs=")+5:info.index("seqs=")+9].strip())
                    mset = data[i+3:i+3+seqs]
                    for m in mset:
                        if 'M|ThV|da' in m:
                            start = int(m[26:31].strip())
                            #print start
                            end = start+width-1
                            resi = str(start)+'-'+str(end)
                            residues.append(resi)
        else:
            residues.append("*")
    return residues

colors1 = ['red','raspberry','deepsalmon','chocolate','firebrick','brown','wampink']
colors2 =
['yellow','green','limegreen','forest','palegreen','smudge','limon','splitpea','paleyellow','lime']
colors3=['blue','lightblue','purpleblue','marine','deeppurple','deepteal','cyan','purple','density','d
eepteal']
colors4 =['orange','tv_orange']

structure = '1TME' #[Change as required]
cmd.load('/home/caroline/PROJECT/StructuralMapping/Structures/'+structure+'.pdb')
cmd.show_as('cartoon','all')
```

```
cmd.select('c','chain 1')
cmd.color('grey70','c')
cmd.select('c','chain 2')
cmd.color('grey70','c')
cmd.select('c','chain 3')
cmd.color('grey70','c')
cmd.select('c','chain 4')
cmd.color('grey70','c')


#[Change as required]
conserved_motifs_vp1 = ['1','*','*','*','28']
residues = get_resi(conserved_motifs_vp1,"VP1")
map_cbv_conserved('1',residues,colors1)
conserved_motifs_vp2 = ['1','2','*','*','5','10','11','*','21','22']
residues = get_resi(conserved_motifs_vp2,"VP2")
map_cbv_conserved('2',residues,colors2)
conserved_motifs_vp3 = ['1','2','3','4','5','*','*','8','9']
residues = get_resi(conserved_motifs_vp3,"VP3")
map_cbv_conserved('3',residues,colors3)
conserved_motifs_vp4 = ['4']
residues = get_resi(conserved_motifs_vp4,"VP4")
map_cbv_conserved('4',residues,colors4)

#printing key of figure
f = open('/home/caroline/PROJECT/StructuralMapping/'+structure+'_key.txt','w')
lines = "VP1 Motifs:\n"
for i,m in enumerate(conserved_motifs_vp1):
   lines+= m+": "+colors1[i]+'\n'
lines +="VP2 Motifs:\n"
for i,m in enumerate(conserved_motifs_vp2):
   lines+=m+": "+colors2[i]+'\n'
lines+= "VP3 Motifs:"+'\n'
for i,m in enumerate(conserved_motifs_vp3):
   lines+= m+": "+colors3[i]+'\n'
lines+= "VP4 Motifs:\n"
for i,m in enumerate(conserved_motifs_vp4):
   lines+= m+": "+colors4[i]+'\n'
f.write(lines)
f.close()
```

1.7 ResidueConservation.py

```
import pylab as pl

import numpy as np
f = open('/home/caroline/PROJECT/AllViruses/VP3/MEME/meme.txt','r') #[Change as
required]
data = f.readlines()
f.close()
motifs = ['9']

virus_ids=['EV-A','EV-B','EV-C','RV-A','RV-B','RV-C'] #[Change as required]

for motif in motifs:
    for virus_id in virus_ids:
        #get regular expression
        for k, line in enumerate(data):
            if "Motif "+motif+" regular expression" in line:
                reg = data[k+2].strip()
                break



        # parse regular expression
        lens = len(reg)
        pos_list=[]
        i=0
        while i<lens:
            res=reg[i]
            if res != "[" and res !="]":
                pos_list.append(res)
                i=i+1
            else:
                if res=='[':
                    i=i+1
                    pos_res=[]
                    while reg[i] != "]":
                        pos_res.append(reg[i])
                        i=i+1
                    pos_list.append(pos_res)
                else:
                    i=i+1


        #get block digrams
        start = data.index("         Motif "+motif+" in BLOCKS format\n")+3
        print start
        end = data.index("//\n",start+1)
        print end
```

```
blocks = data[start:end]

#get virus specific blocks
virus_blocks=[]
for seq in blocks:
    virus = seq[:25].split("|")[1].upper()
    if virus == "EV" or virus =="RV":
        v_type = (seq[:25].split("|")[2])[0].upper()
        virus = virus+"-"+v_type
    if virus==virus_id:
        v_block=seq[33:].split(" ")[0]
        virus_blocks.append(v_block)

#Calculate conservation of each residue in each position according to virus
lines=[]
total = len(virus_blocks)
residues=[]
conservation_list=[]
for i,pos in enumerate(pos_list):
    exceptions=[]
    for res in pos: #Counts conservation of regular expressions
        count = 0.0
        for b in virus_blocks:
            if b[i]==res:
                count=count+1
        conservation = count/total
        residues.append(res)
        conservation_list.append(conservation)
        #lines.append(res+'\t'+str(conservation)+'\n')

    for bl in virus_blocks:
        if bl[i] in pos or bl[i] in exceptions:
            continue
        else:
            e_count = 0.0 #count exceptions to regular expression
            exc = bl[i]
            exceptions.append(exc)
            for m in virus_blocks:
                if m[i]==exc:
                    e_count=e_count+1
            e_con = e_count/total
            residues.append(exc.lower())
            conservation_list.append(e_con)
            #lines.append(exc+'*\t'+str(e_con)+'\n')
    if i<len(pos_list)-1:
        residues.append(" ")
        residues.append(" ")
        conservation_list.append(0)
        conservation_list.append(0)
    #lines.append('\n')
```

```
    #plotting graph
    fig = pl.figure()
    fig.set_size_inches(12,3)
    fig.subplots_adjust(bottom=0.3,top=0.95,left=0.06,right=1)
    ax = pl.subplot(111)
    width=0.8
    ax.bar(range(len(residues)), conservation_list, width=width)
    ax.set_xticks(np.arange(len(residues)) + width/2)
    ax.set_xticklabels(residues,fontsize=12.5)
    # Turn off all the ticks
    ax = pl.gca()
    for t in ax.xaxis.get_major_ticks():
       t.tick1On = False
       t.tick2On = False
    for t in ax.yaxis.get_major_ticks():
       t.tick1On = False
       t.tick2On = False

    pl.xlabel(virus_id, fontsize=14)
    pl.ylabel('Conservation', fontsize=12)

pl.savefig("/home/caroline/PROJECT/AllViruses/VP3/MEME/Viruses/Residues/Conservatio
n/"+virus_id+"_"+motif+".png",dpi=1200)
"""w =
open('/home/caroline/PROJECT/AllViruses/VP1/MEME/Viruses/Residues/'+virus_id+'_'+mo
tif+'.txt','w')
w.writelines(lines)
w.close()"""
```

Appendices 2-7

Appendices 2-7 have been included as digital appendices on the accompanying CD.

Appendix 2: FASTA Protein Files

2.1 All VP4 sequences

2.2 All VP2 sequences

2.3 All VP3 sequences

2.4 All VP1 sequences

2.5 VP4 sequences filtered at 100% for motif analysis

2.6 VP2 sequences filtered at 100% for motif analysis

2.7 VP3 sequences filtered at 100% for motif analysis

Appendix 6: Motif specific interacting residue plots in representative viruses

6.1 EV-A plots
6.2 EV-B plots
6.3 EV-C plots
6.4 RV-A plots
6.5 RV-B plots
6.6 FMDV plots
6.7 ThV Plots

Appendix 7: Conservation plots of motif specific residues across respective virus sub-groups

7.1 Residue conservation of VP4 motifs

7.2 Residue conservation of VP3 motifs

7.3 Residue conservation of VP2 motifs

7.4 Residue conservation of VP1 motifs