ASSESSING THE GENETIC DIVERSITY BETWEEN AND WITHIN HOST POPULATIONS OF THE ROTAVIRUS A VP4 SPIKE PROTEIN

BY

LIZEL L. MENDOZA

A Thesis Submitted to the Graduate Faculty of

WAKE FOREST UNIVERSITY GRADUATE SCHOOL OF ARTS AND SCIENCES

in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

Biology

May 2023

Winston-Salem, North Carolina

Approved By:

James B. Pease, Ph.D., Advisor

David J. Anderson, Ph.D., Chair

Sarah McDonald Esstman, Ph.D.

ACKNOWLEDGEMENTS

Thank you to my graduate advisor, Dr. James Pease, and to my committee members: Dr. Sarah McDonald-Esstman and Dr. Dave Anderson for your guidance and support. I also want to thank my lab members David De La Cerda, Andria Harkey, and Ellen Weinheimer for welcoming me into the Wake Forest University Biology community, being there to answer all my questions, and letting me practice my talks. Lastly, I want to also thank the undergraduates who have helped organize and maintain RotaDB: Allie Lewis, Vincent Ling, Anesia Taylor, and Faye Fang for her research on VP4 and VP7.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS, EQUATIONS, AND TABLES

LIST OF ABBREVIATIONS

Ax	NSP1
BLAST	Basic Local Alignment Search Tool
C	cysteine
Ca	calcium
CD	conserved domain
Cx	VP2
D	aspartic acid
DLP	double layer particle
E	glutamic acid
Ex	NSP4
F	phenylalanine
FDA	The United States Food and Drug Administration
FLU	Influenza
G	glycine
GTR	general time reversal model
Gx	VP7
H	histidine

HBGAs	human blood group antigens
Hx	NSP5
I	isoleucine
Ix	VP6
K	lysine
L	leucine
M	methionine
MRCA	most recent common ancestor
Mx	VP3
N	asparagine
NCBI	The National Center for Biotechnology Information
nMDS	non-metric multidimensional scaling
NSP	non-structural protein
NT	nucleotide
Nx	NSP2
ORF	open reading frame
P	proline
PI-V	Genogroup I-V
Px/P[x]	VP4

Q	glutamine
QC	quartet concordance
QS	quartet sampling
R	arginine
RCWG	Rotavirus Classification Working Group
RNA	ribonucleic acid
RV	Rotavirus
Rx	VP1
S	serine
T*	the triangulation number
T	threonine
Tx	NSP3
USA	The United States of America
V	valine
VP	viral protein
W	tryptophan
WHO	World Health Organization
Y	tyrosine

ABSTRACT

Attachment and entry into the host cell are essential functions for viruses. The spike protein must have variation in its surface proteins to avoid detection by the host immune system. Rotavirus VP4 spike proteins are incredibly diverse, and several studies have attempted to classify VP4 genotypes based on various molecular features and host factors (Liu et al., 2012; Jiang et al., 2017; Sun et al., 2021), yet no one has compared host populations to genotype diversity. Furthermore, connecting VP4 genetic diversity based on host type to functional evolution is not commonly studied. My study aims to explore these knowledge gaps by conducting a thorough assessment of VP4 genetic variation using genotypes 1-57. With bioinformatic tools, I found that P11 may be more divergent than the avian clade. In addition, VP5* is a conserved subunit but may provide additional insight into cell entry and recognition. Reliance on prior knowledge of VP5* has exposed the limited understanding we possess of VP5* and its diversity. Poor surveillance in certain regions and limited wild host species sampling paints an incomplete picture of RV prevalence, host tropism, and host susceptibility. Studying the spike protein diversity will help enlighten what forces perpetuate RV evolution. Furthermore, VP4 can also serve as a model for other researchers studying viruses and provide clinicians with valuable information on vaccine research and development.

INTRODUCTION

Rotavirus (RV) is a double-stranded RNA virus in the Sedoreoviridae family. The pathogen was first discovered in mature epithelial cells by Ruth Bishop, Geoffrey Davidson, Ian Holmes, and Brian Ruck in 1973 at a children's hospital in Australia (Bishop, 2009; Crawford et al., 2017). RVs have been shown to infect a diverse set of mammal and bird species globally, including humans (Bishop, 2009; Crawford et al., 2017) and more than 600,000 people have died each year from RV infections (Bishop, 2009). Since the release of RV vaccines in 2006, the number of positive cases and deaths has dropped as much as 70% up until 2016 (Bishop, 2009; Crawford et al., 2017). However, outbreaks have still occurred in the post-vaccine era due to both lower vaccine efficacy in developing countries and the emergence of new strains (Donato and Bines, 2021). Studying what factors play a role in host susceptibility and RV diversity and constant surveillance are the first steps in controlling RV disease spread.

RV STRUCTURE

RVs have a triple-layered icosahedral capsid approximately 70 nm in diameter (Leung et al., 2005; Figure 1). RV outer and middle capsid layers have T*=13 icosahedral symmetry while the core has T*=1 symmetry, where the triangulation number (T*) represents the number of asymmetric capsid proteins used per face (Louten, 2016). RVs are also non-enveloped, lacking the lipid layer that protects the capsids in some other viruses.

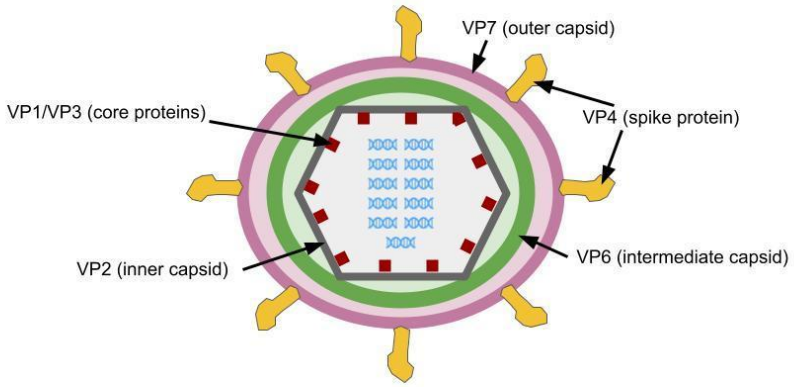


Figure 1: ***The RV Virion and Structural Proteins.*** *VP4 (gold) and VP7 (muted rose) proteins make the outer capsid and the intermediate and core capsids are formed using VP6 (green) and VP2 (gray) proteins, respectively. The dsRNA genome (blue) and VP1 and VP3 core proteins (red) are enclosed inside the core capsid.*

RV PROTEINS AND GENOME

The 12 proteins expressed from the RV genome are categorized as either structural viral proteins (VPs) or non-structural (NSPs). The VPs that comprise the virion are VP1, VP2, VP3, VP4, VP6, and VP7 (Figure 1). The non-structural proteins are NSP1, NSP2, NSP3, NSP4, NSP5, and NSP6. NSPs primarily facilitate RV genome replication (Hu et al., 2018), while VP4 and VP7 form the outer capsid, and VP6 make up the middle shell. VP1, VP2, and VP3 proteins make up the core capsid where the genome

resides (Estes and Cohen, 1989; Greenberg et al., 2009). The genome of an RV is comprised of 11 discrete genomic segments. Each segment contains one open reading frame (ORF) that encodes one protein, except for NSP5/6, which can translate two proteins from two overlapping frame-shifted ORFs (Estes and Cohen, 1989).

VP4, VP5*, AND VP8*

VP4 is the spike protein and is cleaved into two subunits (VP8* and VP5*). VP8* is considered the head of the protein and is formed by two twisted antiparallel β-sheets. VP5* makes up the body, stalk, and foot of the spike protein (Figure 2; Sun et al., 2021). The β-barrel structure makes up the body and stalk while the foot is formed by α helices (Settembre et al., 2011). Within VP5* lies the antigen domain called VP5CT ("CT" stands for chymotrypsin and trypsin) (Dormitzer et al., 2004) Together with VP8* and VP7, they contain most of the epitope sites for potential antibody neutralization (Dormitzer et al., 2004; Settembre et al., 2011).

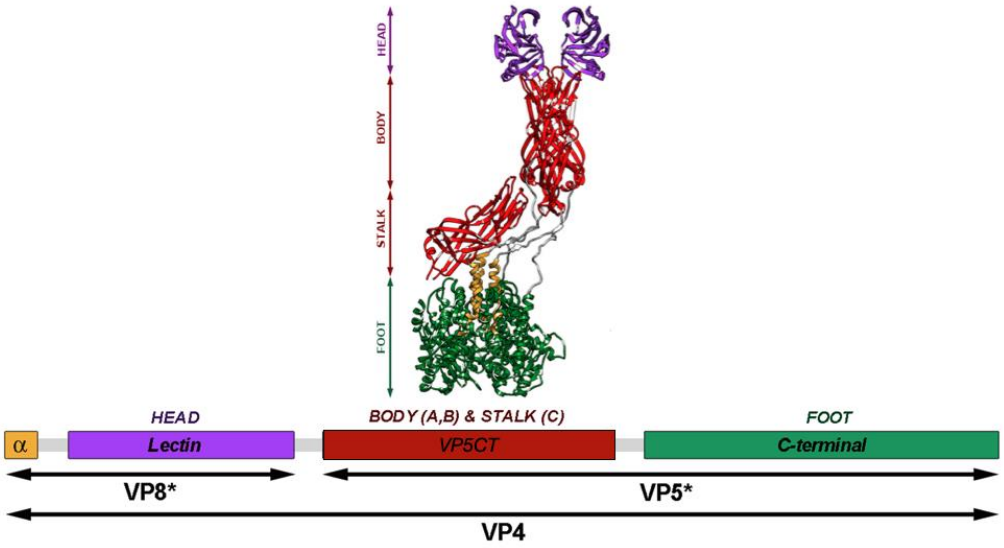


Figure 2: ***VP4 Spike Protein.*** *The head (purple) and n-terminal (gold) are found in VP8*. VP5* consists of the body and stalk (red) and foot (green). Together, VP8* and VP5* make up VP4. Created by Rodriguez et al. (2014). This figure is under the terms of the Creative Commons Attribution License and was adapted by Lizel Mendoza (2023).*

VP4 is also a protease-sensitive protein responsible for attachment and entry into the host cell. Before attachment, VP4 is cleaved by trypsin into VP5* and VP8*, priming the virus for infectivity (Figure 3; Amimo et al., 2021; Vetter et al., 2022). VP8* undergoes a conformational change to attach to the cellular receptors and engage with protein Debrin 1 and an RV restriction factor (Vetter et al., 2022). Once bound, VP5* and the glycoprotein VP7 interact with post receptors such as integrins. As the virus enters via receptor-mediated endocytosis using various endocytic pathways, it is contained inside an endosome (McDonald et al., 2009; Desselberger, 2014). VP4 undergoes a conformational change due to a lower pH and the loss of Ca2+ inside the endosomal

environment, helping the virion break out of the endosome (Suzuki, 2019). Another method RVs can enter is by membrane perforation (Suzuki, 2019). The process involves VP5* folding over as it penetrates the host cell and once inside, VP7 and VP8* dissociate (Jenni et al., 2022). Once the outer shell is shed, the virion becomes a double-layer particle (DLP) inside the cytosol (McDonald et al., 2009; Desselberger, 2014; Arias et al., 2015).

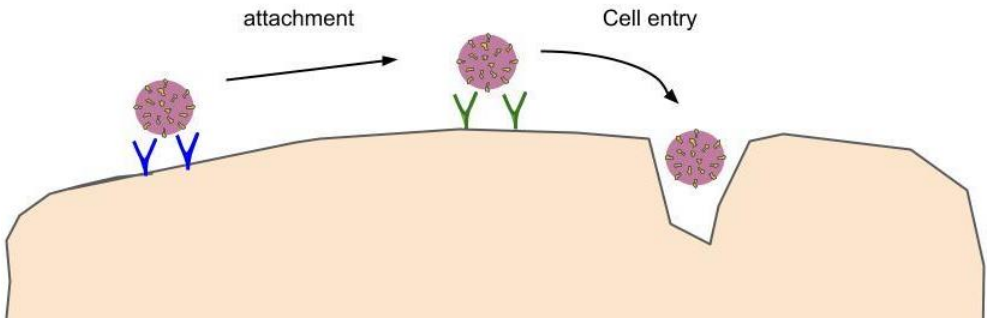


Figure 3: ***RV Attachment and Entry.*** *An RV virion (muted rose and yellow) interacts and attaches to host cellular attachment receptors (blue) with its spike proteins (yellow). Once bound, VP8* dissociates, and VP5* (yellow) and VP7 (muted rose) mediate entry into the host cell (peach) by binding with post-host attachment receptors (green).*

CLASSIFICATION SYSTEM

RV species are categorized alphabetically starting from A and are based on the antigenic diversity of VP6. There are 10 confirmed RV species and two potential subspecies (RVK and RVL; Zhao et al., 2022). RVA, RVB, and RVC have been found in

non-human vertebrates and humans and RVD, RVE, RVF, and RVG have been found strictly in non-human animals. RVA is the most prevalent human pathogen among all RV species and was given its own nomenclature based on a percent nucleotide similarity cutoff for each gene (Table 1; Matthijnssens et al., 2008; Matthijnssens et al., 2011).

Serotyping was used previously to classify RVAs but now has been replaced by genotyping. RVs are categorized by the surface properties of VP6, VP7, and VP4. Moderated by the Rotavirus Classification Working Group (RCWG), the system designates a specific genotype for each of the 11 segments and follows this format: Gx-P[x]-Ix-Rx-Cx-Mx-Ax-Nx-Tx-Ex-Hx (where "x" stands for a numbered genotype for each gene; Desselberger, 2014).While a large number of combinations of genotypes are possible, McDonald et al. (2009) indicated that actual RVA strains sampled are made up of relatively conserved genotype constellations, which they concluded are necessary to preserve protein functional interactions. Not including VP4 and VP7 genotypes, most human strains either possess a genome constellation of all "Genotype 1" genes (also called the "Wa-like" genogroup) or "Genotype 2" (also called the "DS-1-like" genogroup). McDonald et al. (2009) suggested that maintaining this set of genotypes may be necessary for maintaining high fitness in a human host (McDonald et al., 2009). Human strains belonging to the Wa-like genogroup are thought to originate from a common ancestor shared with porcine RVAs, while human DS-1 genogroup strains are derived from a common ancestor with bovine strains (Matthijnssens et al., 2008).

The genotypes of VP4 are labeled "Px," as in the P1, P2, P3, etc. genotypes of VP4. When a new strain of VP4 is sequenced, it is compared to a set of "type strains" that define each genotype. If the sequence is at least 80% similar by nucleotide sequence to at

least one existing type strain, then it is classified as a member of the most similar type strain's genotype. If the sequence is not 80% similar to any existing type, then it becomes a new type strain and receives the next available number. Currently, VP4 has 58 identified genotypes (ICTV, 2023). Each gene has its own cutoff similarity percentages, with VP4 and VP7 having the most permissive similarity cutoff (80%) as a reflection that they are the most variable proteins in sequence within RVA.

Gene	Segment	Genotype	Description	% cutoff
VP1	1	R	**R**NA-dependent RNA Polymerase	83%
VP2	2	C	**C**ore shell protein	84%
VP3	3	M	**M**ethyltransferase mRNA capping enzyme	81%
VP4	4	P	**P**rotease-sensitive cell attachment protein	~80%
VP5	*4*	-	*(fragment of VP4)*	
VP6	6	I	**I**ntermediate capsid shell, structural species-specific antigen	85%
VP7	9	G	**G**lycosylated surface neutralization antigen	~80%
VP8	*4*	-	*(fragment of VP4)*	
NSP1	5	A	5'RNA binding, interferon **A**ntagonist	79%
NSP2	8	N	**N**TPase involved in RNA packaging	85%
NSP3	7	T	**T**ranslation enhancer	85%
NSP4	10	E	**E**nterotoxin	85%
NSP5	11	H	p**H**osphoprotein, ssRNA and dsRNA binding modulator of NSP2	85%
NSP6	*11*	-	*Frame-offset protein from the NSP5 segment, ssRNA and dsRNA binding protein*	

Table 1: ***Information on RV Gene and Protein Function, Sequence Length, and Nucleotide Percent Similarity Cutoff.***

EPIDEMIOLOGY

RVA is a widespread pathogen causing acute gastroenteritis, particularly in children (Kuang et al., 2020; Donato and Bines, 2021). RVA infections can either be asymptomatic or symptomatic such as persistent fever, vomiting, and/or diarrhea. The most common transmission route is through fecal-to-oral contact, but close person-to-person contact and contaminated fomites are also possible transmission pathways (Crawford et al., 2017).

Vaccine development against the pathogen began in the 1980s. Still, it took almost 20 years for The United States Food and Drug Administration to approve the first live, oral rotavirus vaccine, RotaTeq® and then Rotarix® three years later (Greenberg and Estes, 2009). Many vaccines have since been launched overseas such as Rotavac® and Rotasiil® in India in 2014 and 2018, respectively. Rotarix® is monovalent meaning they were designed to elicit an immune response from one strain while RotaTeq® is pentavalent, which provides immunity from five strains (Glass et al., 2021). In 2003, close to 114 million RV cases were reported around the world in children under five years old, and currently there are more than 100 countries that have access to at least one RV vaccine (Glass et al., 2021).

Ten years after the first vaccine was licensed, rotavirus-related cases dropped by ~70% and deaths dropped by ~40% (Crawford et al., 2017). However, vaccine efficacy rates are disproportional across the world, with Europe, the USA, and Australia exhibiting the highest rates of effectiveness while developing countries in Africa and South America proved to have low to moderate protection. Some people hypothesize that innate host characteristics such as the differences in polysaccharides found on lipids and

proteins on the cell's surface (glycans) may be a factor in host susceptibility because successful attachment to host cell surface receptors mediate cell entry (Li et al., 2021; Donato and Bines, 2021).

RV GENETIC DIVERSITY

RV, like other segmented viruses, can increase its genetic diversity through various mechanisms. Viral RNA-dependent RNA polymerases are highly error-prone and lack proofreading capability, leading to high rates of mutation (Sanjuán and Domingo-Caley, 2021). With enough mutations, new lineages will emerge (Retel et al., 2019). Reassortment also plays a significant role in RV evolution. If two viruses infect the same cell and trade genomic segments, they can reassort and potentially produce a new genotype constellation (McDonald et al., 2016). Reassortment elevates genetic diversity even further beyond the accumulation of individual mutation. Furthermore, studies have demonstrated that different types of host cell receptors are recognized by different genotypes (Liu et al, 2012; Sun et al., 2018; Amino et al., 2021). Liu et al. (2012) also highlighted that genotypes P1–P35 can be categorized into five distinct clusters or "genogroups" based on the phylogenetic analysis conducted on VP8* (Figure 4). Later studies attempted to further classify these genogroups by host or glycan specificity but failed to show a direct link between either (Jiang et al., 2017; Xu et al., 2020; Sun et al., 2021).

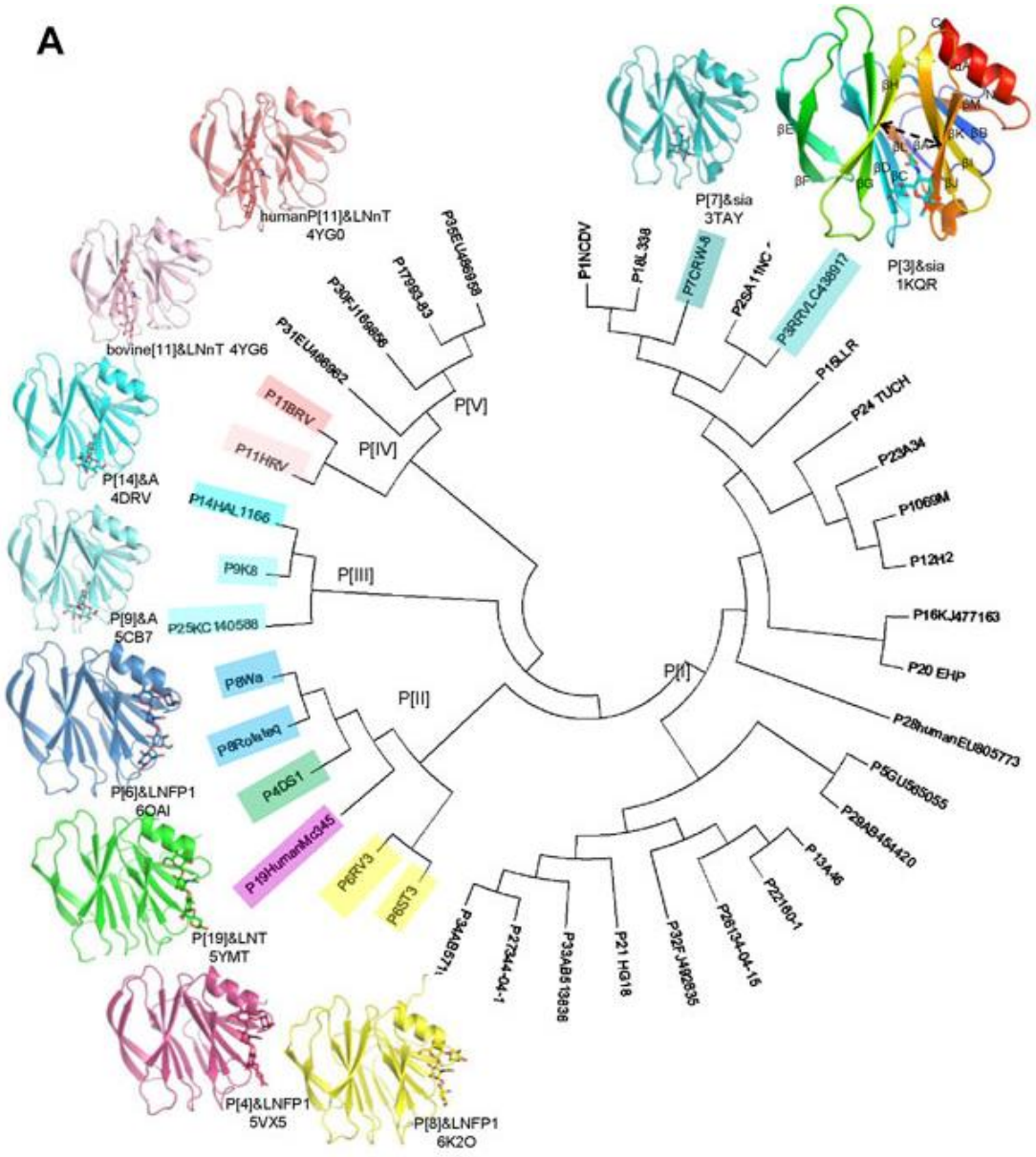


Figure 4: ***Ribbon Structures and Phylogenetic Analysis of VP8*'s Genotypes 1-35.*** *P1-35 are classified into five genogroups based on VP8* phylogenetic analysis using the neighbor-joining method. The ribbon structures in P[I] (teal), P[II] (cyan, green, magenta, and yellow), P[III] (turquoise), P[IV] (pink), and P[V] (salmon) are color-matched to the corresponding branch(es) on the phylogenetic tree. Created by Sun, Li,*

and Duan (2021). This figure is under the terms of the Creative Commons Attribution License (CC BY) and was adapted by Lizel Mendoza (2023).

MOTIVATION AND GOAL

RVA strains have been shown to be capable of infecting both mammals and birds, but previously were modeled as having a narrow host tropism. However, increasing evidence indicates that human and non-human strains share common origins. Close proximity and interactions between humans and animals increase the risk of coinfection and reassortment events between two strains of different origins, also raising the chance of zoonotic and reverse-zoonotic transmission (Malik et al., 2020). Coinfection rates due to reassortment events in developing countries are near 20% due to high levels of RV genetic variation in the population. Hybrids through reassortment are a cause of concern because some can mask the detection of distinct genotypes since VP4 epitopes can possess phenotypic differences or antigenic variation due to differences in their amino acid sequences (Patton 2012; Wiley et al., 2015). But due to the pressures of vaccines, new mutants that possess enough VP4 antigenic changes to evade the vaccine-induced host immune response may be positively selected for (George et al., 2018). In addition, greater advances in surveillance techniques have highlighted that emerging RV strains possess mutations that favor a broad host tropism. In turn, RV prevention becomes more challenging as this creates an

opportunity for strains to infect novel species, even more so, if species are asymptomatic (Malik et al., 2021).

Despite these endeavors, VP4 genotype tendencies have been linked to certain hosts. The most prevalent human strains are mainly found in P8, P4, or P6. P7- and P13-carrying strains predominantly infect pigs, while P5- and P11-carrying strains predominate in bovids (Papp et al., 2013). The genotypes that constitute Genogroup V [PV] (P17, P30, P31, and P35) consist of avian strains (Liu et al., 2012). It is imperative, then, to assess the genotype diversity in animal populations for the well-being of non-human animals and humans and improve my knowledge of RV interspecies transmission and reassortment events (Donato and Bines 2021). Evaluating host-genotype relationships may be a valuable proxy for understanding emerging strains and predicting their origins (Geletu et al., 2021).

Attachment and entry into the host cell are arguably the most essential evolutionary features for viruses because viruses require host cell machinery for genome replication and virion assembly (Amimo et al., 2021). This study aims to understand VP4 genotype diversity within and between host populations. I used a combination of bioinformatics and analysis of primary literature to analyze their population structure and identify molecular regions of functional significance. This is the first study to compare the genetic variation between different host types using P1-57. My results on genotype diversity shed light into RVA host tropism and VP4 sequence variation.

METHODS

DATA ACQUISITION AND CURATION

RV VP4 sequences were downloaded from NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) by identifying all sequences marked as RV and VP4. After exporting the Genbank records, the metadata underwent error-checking and correction, including going back to the primary literature to clear up uncertainties in sample source, location and date of collection, and host. The sequences were assigned a unique identifier using a standardized nomenclature: RVA/[HOST]-[SOURCE]/[COUNTRY]/[STRAIN]/[YEAR]. HOST comprises either five or seven letters to symbolize genus and species (Table 2). SOURCE is a two-letter identifier to describe if the strain is wild-type, tissue culture, an environmental sample, or of unknown origin. COUNTRY is the ISO 3-letter code for the country of sampling (XXX is added when the country of origin could not be resolved). STRAIN is the name given by the submitters. YEAR is the year when the strain was isolated (not sequenced or submitted (XXXX when YEAR is unknown).

Strains without a genotype or were suspected to be mislabeled with the wrong genotype were searched as queries in NCBI BLAST (https://blast.ncbi.nlm.nih.gov). For sequences with genotypes that did not meet the 80% nucleotide minimum cutoff, the genotype was changed from its submitted annotation on NCBI to the reference strain's genotype that has the highest query cover and percent identity match (usually both above 90%). I also used RotaC (now discontinued), an RVA whole genome classification tool to confirm the genotype (Maes et al., 2009).

The final sequence sets excluded all non-RVA species, *in vitro* cultured strains, vaccine-derived strains, and environmentally sampled strains (without a clear host). ow-quality sequences, defined as (1) containing at least five ambiguous characters, (2) being less than 25% of the complete VP4 coding sequence, (3) duplicate sequences, or (4) sequences that contained so many mutations relative to any other strain of RVA that homology was not able to be determined were also removed. Because of this criterion, sequences from genotypes P39, P44, P52, P53, P54, P55, and P58 were not included. The exceptions were the type strains used to define the VP4 P genotypes.

MULTIPLE SEQUENCE ALIGNMENT

Sequences were trimmed to the coding sequence (start to stop) and initially aligned as proteins using MUSCLE (Edgar, 2004) and then "untranslated" back to nucleotide sequences using a custom Python3 script. Manual editing of the alignment was necessary in many cases to repair obvious frameshifts due to missing bases. The final alignment was realigned with MAFFT v7.487 (Kazutaka and Standley, 2013) before being "untranslated" back to a nucleotide file using a custom Python3 script. The final sequence alignment has 4071 sequences and 2505 sites. The boundaries of the VP5* and VP8* regions followed Dormitzer et al. (2006), where VP8* is at the start of the VP4 (1–304 amino acid bases) while the remaining majority is occupied by VP5* (305–835 amino acid bases).

To minimize sampling bias, I also curated a reduced dataset where each genotype was allowed to have a maximum number of ten sequences. Since each genotype was allowed a maximum number of ten sequences and was chosen at random in my host

datasets using SeqKit v2.4.0 (Shen et al., 2016) and Google Sheets, each genotype and host population might not fully represent the full diversity since each site carries less information than the full dataset. Although still not ideal, this presented a more realistic comparison.

PHYLOGENETIC TREE RECONSTRUCTION

VP4 nucleotide- and amino acid-based phylogenetic trees were inferred using IQ-TREE v1.6.12 using the GTR (general time reversible) model for nucleotides and the "FLU" model for amino acids (Tavare, 1986; Dang et al., 2010). Trees were visualized with Figtree v.1.4.4 (Rambaut, 2007), with the avian clade used as the outgroup for rooted trees. Quartet Sampling (Pease, et al. 2018) was used to generate confidence scores using the complete VP4 nucleotide and proteins sequences. We tested 200 replicates using the default settings with the default ΔlnL cutoff of 2.0.

abbrev	Genus	Species	Common Name
AANACAS	*Anas*	*castanea*	chestnut teal
AANAPLA	*Anas*	*platyrhynchos*	mallard duck
AANASUP	*Anas*	*superciliosa*	Pacific black duck
AARRTAC	*Arremon*	*taciturnus*	pectoral sparrow
ABUTMER	*Buteogallus*	*meridionalis*	savanna hawk
ACOLLIV	*Columba*	*livia*	pigeon
ACOTCOT	*Coturnix*	*coturnix*	common quail
AGALGAL	*Gallus*	*gallus*	domesticated chicken
ALARCAN	*Larus*	*canus*	common gull
AMELFUS	*Melanitta*	*fusca*	velvet scoter duck
AMELGAL	*Meleagris*	*gallopavo*	turkey
ANUMMEL	*Numidia*	*meleagris*	domesticated guineafowl

APERPER	*Perdix*	*perdix*	grey partridge
APHACOL	*Phasianus*	*colchicus*	common pheasant
ASPICHI	*Spilopedia*	*chinensis*	spotted dove
ASPOANG	*Sporophila*	*angolensis*	chestnut-bellied seed finch
ATURRUF	*Turdus*	*rufiventris*	rufous-bellied thrush
AUNKNWN	*Unknown*	*unknown*	unknown avian
BCARPER	*Carollia*	*perspicillata*	Seba's short-tailed bat
BDESROT	*Desmodus*	*rotundus*	Common vampire bat
BEIDHEL	*Eidolon*	*helvum*	straw-coloured fruit bat
BGLOSOR	*Glossophaga*	*soricina*	Pallas's long-tongued bat
BHIPCAF	*Hipposideros*	*caffer*	Sundevall's roundleaf bat
BHIPGIG	*Hipposideros*	*gigas*	giant leaf-nosed bat
BHIPLAR	*Hipposideros*	*larvatus*	intermediate roundleaf bat
BHIPPOM	*Hipposideros*	*pomona*	Pomona roundleaf bat
BMINSCH	*Miniopterus*	*schreibersii*	common bent-wing bat
BMOLMOL	*Molossus*	*molossus*	velvety free-tailed bat
BMYODAS	*Myotis*	*dasycneme*	vesper bat
BMYODAU	*Myotis*	*daubentonii*	Daubenton's bat
BMYOMYS	*Myotis*	*mystacinus*	whiskered Myotis
BPIPKUH	*Pipistrellus*	*kuhlii*	Kuhl's pipistrelle
BPIPPIP	*Pipistrellus*	*pipistrellus*	common pipistrelle
BPTEGIG	*Pteropus*	*giganteus*	Indian flying fox
BRHIBLA	*Rhinolophus*	*blasii*	Blasius's horseshoe bat
BRHIEUR	*Rhinolophus*	*euryale*	Mediterranean horseshoe bat
BRHIFER	*Rhinolophus*	*ferrumequinum*	greater horseshoe bat
BRHIHAR	*Rhinopoma*	*hardwickii*	lesser mouse-tailed bat
BRHIHIP	*Rhinolophus*	*hipposideros*	lesser horseshoe bat
BRHISIM	*Rhinolophus*	*simulator*	Bushveld horseshoe bat
BROUAEG	*Rousettus*	*aegyptiacus*	Egyptian fruit bat
BROULES	*Rousettus*	*leschenaulti*	Leschenault's rousette
BSCOKUH	*Scotophilus*	*kuhlii*	lesser Asiatic yellow bat
BTAPMAU	*Taphozous*	*mauritianus*	Mauritian tomb bat
BUNKNWN			unknown bat
EAGFARM			agricultural farm
EFOODSM			food sample
EFRSHWT			freshwater

EGRNDWT			groundwater
EHOSPTL			hospital/clinic
EMARNWT			marine salwater
ERCYCWT			recycled water
ESEWAGE			sewage
EUNKNWN	*Unknown*	*unknown*	unknown environmental
HUMAN	*Homo*	*sapiens*	human
HUNKNWN	*Homo*	*sapiens*	presumably human, but unconfirmed
LLABSTR			lab strain
LVACCIN			vaccine-derived
MAILMEL	*Ailuropoda*	*melanoleuca*	giant panda
MBOSGRU	*Bos*	*grunniens*	domesticated yak
MBOSIND	*Bos*	*indicus*	zebu
MBOSTAU	*Bos*	*taurus*	domesticated cow
MBUBBUB	*Bubalus*	*bubalis*	water buffalo
MCALPEN	*Callithrix*	*penicillata*	black-tufted marmoset
MCAMDRO	*Camelus*	*dromedarius*	dromedary camel
MCANFAM	*Canis*	*familiaris*	domesticated dog
MCAPCAP	*Capreolus*	*capreolus*	roe deer
MCAPHIR	*Capra*	*hircus*	goat
MCAVPOR	*Cavia*	*porcellus*	domesticated guinea pig
MCHLPYG	*Chlorocebus*	*pygerythrus*	green vervet monkey
MEQUASI	*Equus*	*asinus*	domesticated donkey
MEQUCAB	*Equus*	*caballus*	domesticated horse
MFELCAT	*Felis*	*cattus*	domesticated cat
MGIRGIR	*Giraffa*	*giraffa*	southern giraffe
MHIPNIG	*Hippotragus*	*niger*	sable antelope
MLAMGLA	*Lama*	*glama*	llama
MLAMGUA	*Lama*	*guanicoe*	guanaco
MMACMUL	*Macaca*	*mulatta*	rhesus macacque
MMACNEM	*Macaca*	*nemestrina*	southern pig-tailed macaque
MMASNAT	*Mastomys*	*natalensis*	natal multimammate mouse
MMELURS	*Melursus*	*ursinus*	sloth bear
MMICAGR	*Microtus*	*agrestis*	field vole
MMUNREE	*Muntiacus*	*reevesi*	chinese muntjac

MMUSMUS	*Mus*	*musculus*	mouse
MNYCPRO	*Nyctereutes*	*procyonoides*	raccoon dog
MORYCUN	*Oryctolagus*	*cuniculus*	domesticated rabbit
MOVIARI	*Ovis*	*aries*	domesticated sheep
MPAGLAR	*Paguma*	*larvata*	masked palm civet
MPANUNC	*Panthera*	*uncia*	snow leopard
MPETBRE	*Petaurus*	*breviceps*	sugar glider
MPROLOT	*Procyon*	*lotor*	common raccoon
MRATNOR	*Rattus*	*norvegicus*	brown rat
MRATRAT	*Rattus*	*rattus*	black rat
MSCIVUL	*Sciurus*	*vulgaris*	red squirrel
MSORARA	*Sorex*	*araneus*	common shrew
MSUSSCR	*Sus*	*scrofa*	domesticated pig
MSYNCAF	*Syncerus*	*caffer*	African buffalo
MUNKNWN	*Unknown*	*unknown*	unknown mammalian
MVICPAC	*Vicugna*	*pacos*	alpaca
MVICVIC	*Vicugna*	*vicugna*	vicuna
MVULVUL	*Vulpes*	*Vulpes*	red fox
MZALCAL	*Zalophus*	*californianus*	California sea lion
UUNKNWN			unknown
ZMAGARI	*Magallana*	*ariakensis*	suminoe oyster
ZMAGBEL	*Magallana*	*belcheri*	oyster
ZMYTGAL	*Mytilus*	*galloprovincialis*	Mediterranean mussel
ZPERVIR	*Perna*	*viridis*	Asian green mussel
ZSACFOS	*Saccostrea*	*foskali*	rock oyster
ZTEGNOD	*Tegillarca*	*nodifera*	ark clam
ZUNKNWN	*Unknown*	*unknown*	unknown metazoa

Table 2: ***RVA Host Code.***

GENETIC DIVERSITY

To understand how populations are structured within VP4, I measured and compared the genetic distances between and within populations. A non-metric multidimensional scaling (nMDS) function was administered in Mothur v1.48.0 (https://mothur.org/) using a Phylip-formatted distance matrix.

In order to quantify the levels of genetic variation within each genotype, I calculated d_X (also known as π or Tajima's θ; Nei and Li 1979, Nei and Tajima 1981; Equation 1). This measure of genetic diversity is ordinarily calculated for populations of eukaryotic organisms but can be used generally to describe genetic diversity in any set of sequences. The value d_X is equal to the average pairwise sequence distance between all pairs of sequences within a population. In this case, I calculated d_X using a Python3 script for the set of sequences in each P-genotype of VP4.

Equation 1: $$d_X = \frac{\sum_{i<j} k_{ij}}{n(n-1)/2}$$

I also calculated d_{XY} for each pair of VP4 P-genotypes using a Python3 script (Nei and Li, 1979) . My d_{XY} Python3 script first computes the number of pairwise differences between one individual from population X and another individual from population Y (Figure 5). Once it has gone through each sequence comparison between the two populations, it sums the number of pairwise differences before averaging it (Equation 2; Nei and Li 1979).

Equation 2: $d_{XY} = \Sigma_{ij} x_i y_j k_{ij}$

The value of d_{XY} is equal to the average pairwise distance between all pairs of sequences where one member of each pair comes from two separate populations (X and Y). In this case, the populations were the sets of sequences in each P-genotype of VP4.

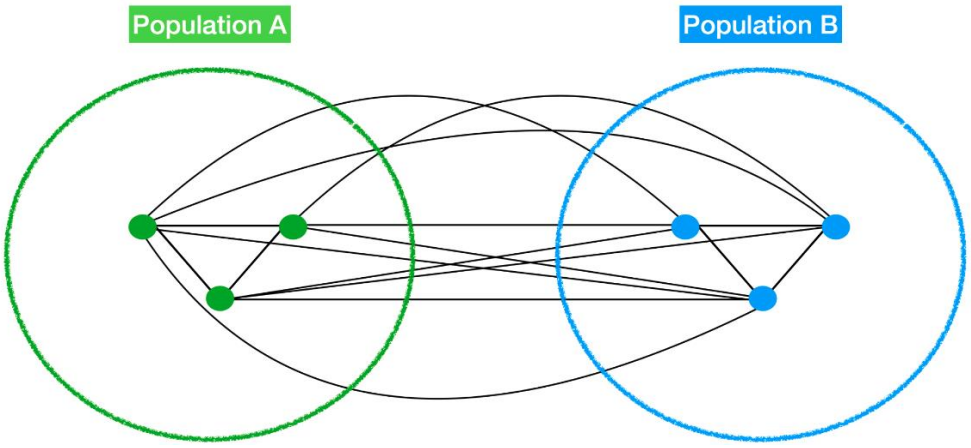


Figure 5: ***Examples of d_{XY} and d_X.*** *d_{XY} is the average pairwise sequence distances (black lines) between all pairs of individuals (circles) between two populations (green and blue rings). d_X is the average pairwise sequence distances (black lines) within a population (Hahn, 2019).*

IDENTIFYING CONSERVED DOMAINS

To attribute VP4's molecular variation to its phenotypic differences, consensus sequences were first created using SEAVIEW v5.0.4 (Gouy et al., 2021). The represented base must be found in at least 20% of the sequences at each site. Establishing this cutoff ensured no ambiguous code was included. A search for conserved domains using Motif (https://www.genome.jp/tools/motif/) was used to compare residue differences against hosts. To assess the amino acids at the conservative and variable regions, amino acid sequence logo plots were generated in WebLogo 3 (https://weblogo.threeplusone.com/create.cgi). Potential functional regions were found in the primary literature from Dormitzer et al., 2006.

PYTHON SCRIPTS

Python scripts used in this study were developed by the Pease lab and are open-source. Scripts can be found on my GitHub page (https://github.com/peaselab/rotaspike).

RESULTS

The goal of this study is to analyze VP4 genetic diversity between RVAs that infect various VP4 host populations and correlate their sequence homology to possible functions. Sequence information on the 57 VP4 genotypes was first collected before sequence alignment and analysis of 4071 RVA VP4 sequences from 51 out of the 57 genotypes (Figure 6). Two conserved domains and potential membrane interaction motifs were identified when comparing human, pig, cow, and avian consensus sequences. My work brings insight into the relationship between host and VP4 spike protein diversity and genotypes with small sample sizes emphasizing the importance of zoonotic strain surveillance.

The nucleotide- and amino acid-level phylogenetic trees were inferred from all 4701 sequences belonging to 51 genotypes (Figure 7 and Figure 8). The root generally separated the avian and mammalian clades with the exception of avian P37, which is more similar in sequence to mammalian genotypes than avian. A similar tree topology among the VP4 and VP8* nucleotide and amino acid trees was observed. Quartet Sampling (QS) scores reveal some discordance and only 21 nodes are 100% supported. For example, the P37 avian strains grouping with the mammalian sequences instead of the avian clade showed the highest concordance (1).

Several P4 strains share a more recent common ancestor with P8 than the P4 group in the nucleotide-based phylogeny. Furthermore, P4 and P8 form separate clades that then share an exclusive common ancestor in the amino acid-based phylogeny, but P8 is nested inside the P4 clade in the nucleotide-based phylogeny. Furthermore, there is a high amount of discordance among the P4 and P8 branches in the nucleotide tree but not

in the amino acid tree. RotaC confirmed that these genotypes are correctly labeled based on their relative genetic distances to the P4 and P8 type strains, despite the phylogenetic instability.

Based on the genogroups determined by Liu et al. (2012), P[II] through P[V] cluster together as expected apart from P[I] (Figure 4). For example, the consensus tree models P[V] (which are mostly avian strains) to be monophyletic, but a handful of P[I] genotypes (P5, P10, P12, P15, P20, P21, P23, P24, P27, P33, and P34) share a more recent common ancestor with genogroup P[II] (P4, P6, P8, and P19) than with other P[I] genotypes. Although P11 of P[IV] and P29 of P[I] belong to different genogroups, they form a clade and are most distantly related to the avian clade.

The consensus tree reconstructs most VP4 genotypes as monophyletic, apart from P4, P6, P13, P22, and P27. Strains from bat hosts are often divergent from other mammalian RVA strains (Bourdett-Stanziola, 2022). The nine VP4 genotypes that were sampled from bats (P2, P3, P10, 42, P43, P47, P48, P50, and P51) make up a paraphyletic group. Some genotypes are most genetically similar to human and non-human mammalian strains. Predominantly porcine strains P7 and P13 share an exclusive common ancestor, but predominantly bovine strains P5 and P11 do not.

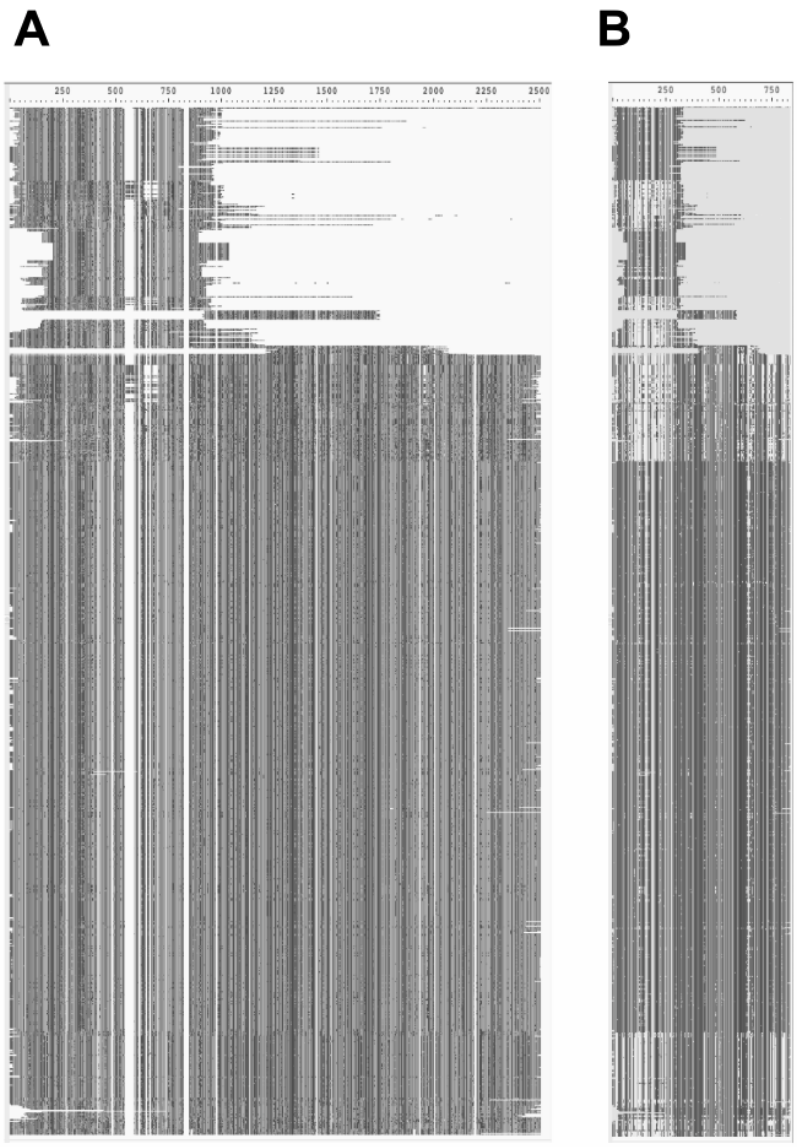


Figure 6: ***VP4 Alignments.*** *A. Nucleic acid alignment (left). B. Amino acid alignment (right) Both alignments comprise 4,701 sequences with 2505 and 835 sites, respectively. Viewed in Aliview.*

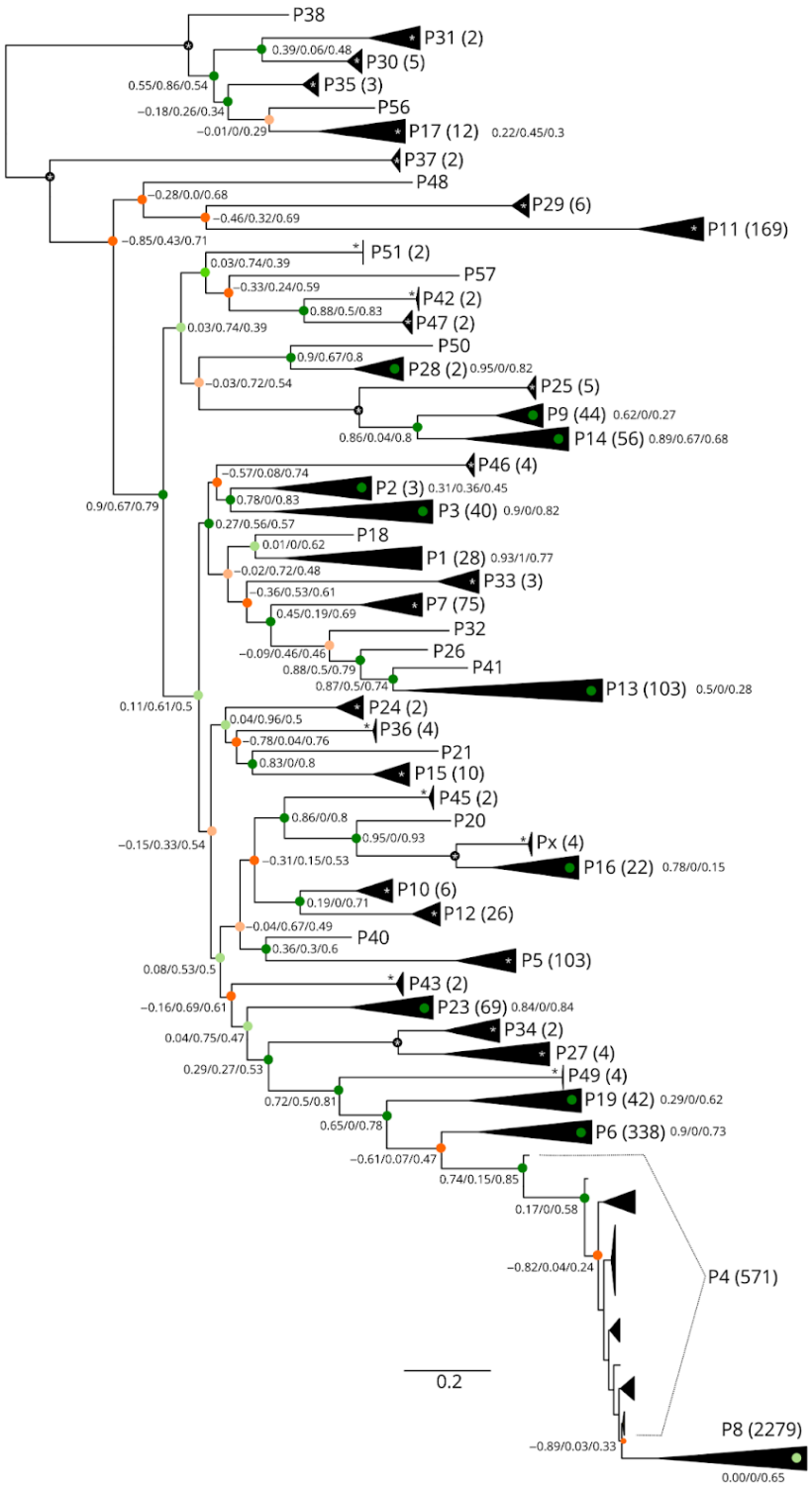


Figure 7: ***Nucleotide Phylogenetic Tree with Quartet Sampling (QS).*** *Nodes with either a negative or positive quartet concordance (QC) score are colored in orange or green,*

respectively. QC scores have a range of [-1,1], the closer the QC score is to zero the lighter the shade. Branches with 100% support are starred. Visualized in Figtree.

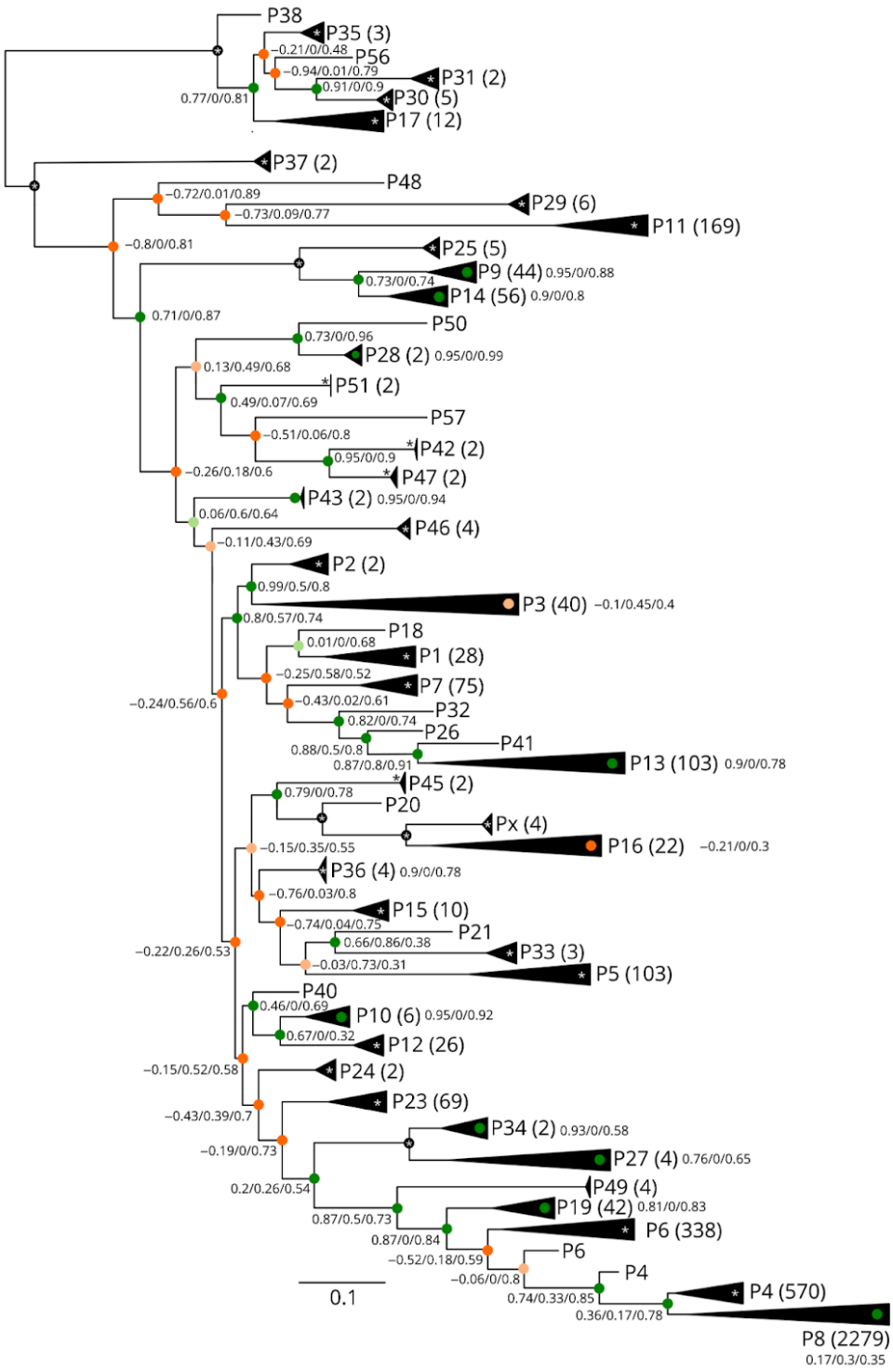


Figure 8: ***Amino Acid Phylogenetic Tree with Quartet Sampling (QS).*** *Nodes with either a negative or positive quartet concordance (QC) score are colored in orange or*

green, respectively. QC scores have a range of [-1,1], the closer the QC score is to zero the lighter the shade. Branches with 100% support are starred. Viewed in Figtree.

In order to further characterize the genotype relationships, an nMDS analysis using their pairwise distances were performed. There are some similarities shared between the protein and nucleotide nMDS (Figure 9 and Figure 10). Sequences generally cluster based on genotype, but only avian and horse strains exhibit clustering by host. The phylogenetic and nMDS models suggest crow P56 (RVA/ACORMAC-wt/JPN/JC-105/2019/VP4/P56/LC634425) and turkey P38 (RVA/AMELGAL-tc/IRL/Ty-1/1979/VP4/P38/LC088110) should be added to the avian genogroup P[V]. The nMDS further supports the close relationship between P4 and P8 sequences.

In the amino acid nMDS, P[II], P[IV], and P[V] are more distinct. P5 and genogroups II and IV are split into two clusters. Cow strains from P7 and P13 are adjacent to each other and the two avian P37 strains do not cluster with the other avian sequences.

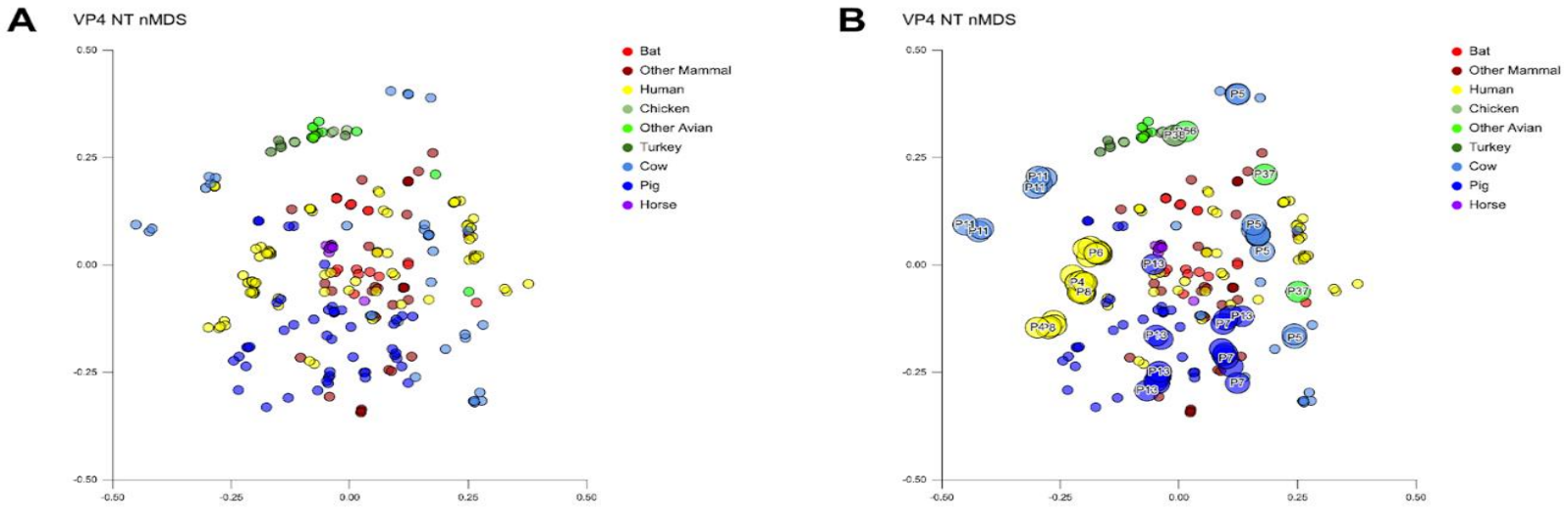


Figure 9: ***VP4 Nucleotide nMDS plots.*** *A. Points colored by host type. B. Points colored by host type and annotated with notable genotypes. nMDS models are made with a reduced dataset comprised of 254 sequences from Genotypes 1-57. Visualized in Google Sheets.*

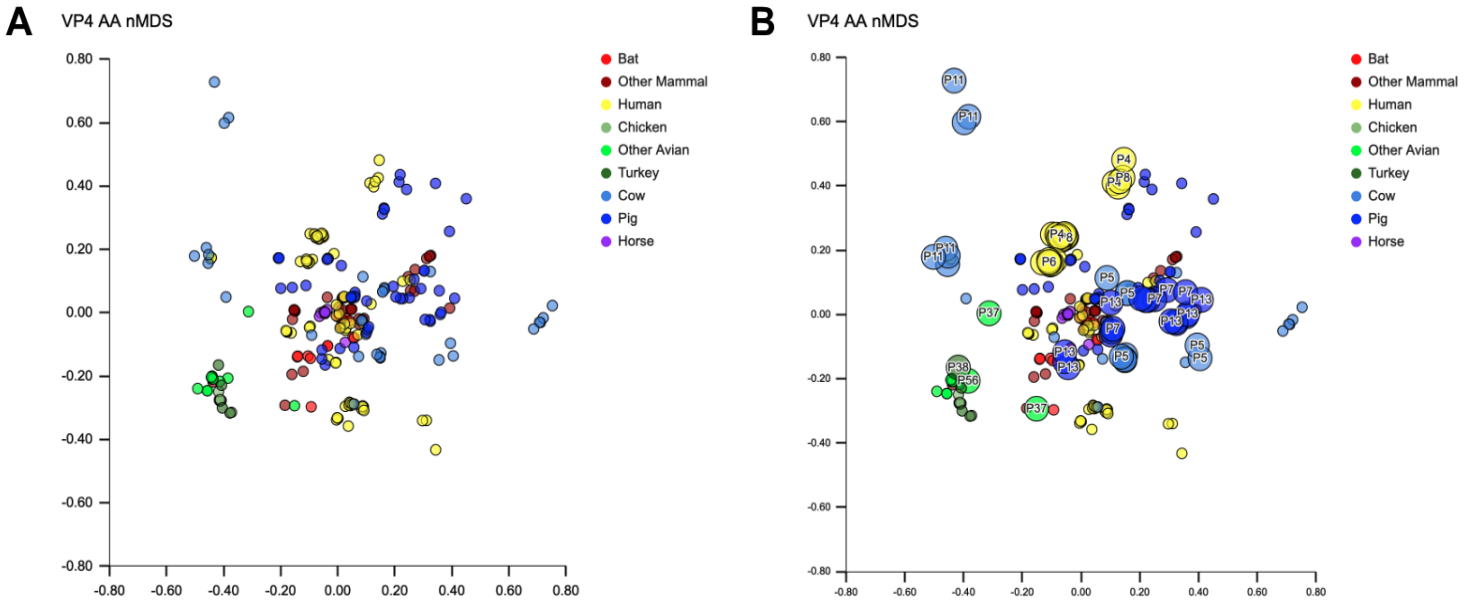


Figure 10: ***VP4 Protein nMDS plots.*** *A. Points colored by host type. B. Points colored by host type and annotated with notable genotypes. nMDS models are made with a reduced dataset comprised of 254 sequences from Genotypes 1-57. Visualized in Google Sheets.*

GENOTYPE HOST TENDENCIES

From the dataset, there are predominantly seven avian and 39 non-human mammalian genotypes (Table 3). Sequence abundance is highly varied among the genotypes. The most sampled genotypes are P8 (2279 sequences), P4 (571 sequences), P6 (339 sequences), P11 (169 sequences), P5 (103 sequences), P13 (102 sequences) and P7 (75 sequences). In contrast, 44 genotypes have fewer than 10 sequences. This highlights the unevenness in sampling and lack of broad sampling in wild species and strains of RVA. This varied sampling creates a technical challenge in studying and modeling virus diversity and evolution since the uneven sequence capture across time, host species, and global regions means I do not have a broadly sampled and representative set of sequences to understand the macroscopic picture.

Geno	HUMAN	MSUSSCR	MBOSTAU	MEQUCAB	Other Mammals	Bats	AGALGAL	AMELGAL	Other Avian	LLABSTR
1	4	3	9	0	0	0	0	0	0	0
2	2	0	0	0	0	1	0	0	0	0
3	11	3	0	1	0	11	0	0	0	0
4	568	1	2	0	0	0	0	0	0	0
5	0	0	103	0	0	0	0	0	0	0
6	285	53	0	0	0	0	0	0	0	1
7	0	68	6	0	0	0	0	0	0	0
8	2277	1	1	0	0	0	0	0	0	0
9	40	0	0	0	0	0	0	0	0	0
10	4	0	0	0	0	2	0	0	0	0
11	33	0	131	0	5	0	0	0	0	0
12	0	0	1	24	1	0	0	0	0	0
13	0	101	0	0	1	0	0	0	0	0
14	31	0	12	0	13	0	0	0	0	0
15	3	0	1	0	6	0	0	0	0	0
16	0	0	0	0	21	0	0	0	0	1
17	0	0	1	0	1	0	0	0	10	0
18	0	0	0	1	0	0	0	0	0	0
19	12	30	0	0	0	0	0	0	0	0
20	0	0	0	0	1	0	0	0	0	0
21	0	0	1	0	0	0	0	0	0	0
22	0	0	0	0	1	0	0	0	0	0
23	1	68	0	0	0	0	0	0	0	0
24	1	0	0	0	1	0	0	0	0	0
25	5	0	0	0	0	0	0	0	0	0
26	0	1	0	0	0	0	0	0	0	0
27	0	4	0	0	4	0	0	0	0	0
28	2	0	0	0	0	0	0	0	0	0
29	0	0	6	0	0	0	0	0	0	0
30	0	0	0	0	0	0	5	0	0	0

31	0	0	0	0	0	0	2	0	0	0
32	0	1	0	0	0	0	0	0	0	0
33	0	0	3	0	0	0	0	0	0	0
34	0	2	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	1	2	0	0
36	0	0	0	0	4	0	0	0	0	0
37	0	0	0	0	0	0	0	0	2	0
38	0	0	0	0	0	0	0	1	0	0
40	0	0	0	0	1	0	0	0	0	0
41	1	0	0	0	0	0	0	0	0	0
42	0	0	0	0	0	2	0	0	0	0
43	0	0	0	0	0	2	0	0	0	0
45	2	0	0	0	0	0	0	0	0	0
46	4	0	0	0	0	0	0	0	0	0
47	0	0	0	0	0	2	0	0	0	0
48	0	0	0	0	0	1	0	0	0	0
49	0	4	0	0	0	0	0	0	0	0
50	0	0	0	0	1	0	0	0	0	0
51	0	0	0	0	0	2	0	0	0	0
56	0	0	0	0	0	0	0	0	1	0
57	0	0	0	0	1	0	0	0	0	0

Table 3: ***VP4 Genotype Host Tendencies.*** *The number of sequences in each VP4 genotype (rows) and host category (columns) are shown for all 4071 sequences in my primary data set. Predominantly avian genotypes are marked with an asterisk. Lab Strains were only included if they were a genotype classification type strain. Visualized in Google Sheets.*

Viruses depend on their host to survive and replicate, so studying a virus's host tropism, or the ability to infect a host, can help us understand the mechanisms that determine viral virulence and the likelihood a virus will become pathogenic (Douam et al., 2015). In this study, the host tropisms of different RVA strains were analyzed to better understand the relationship between genotypes and host diversity. 21 out of the 51 genotypes sampled infect at least two species and genotypes P1, P3, P9, P11, P14, P15, and P17 have at least four hosts (Table 3). P17 is the only genotype with avian and mammalian strains and contains the largest number of known avian host species. The lone mammalian P17 comes from a fox sample that contains avian origins in Italy in 2011 (Busi et al., 2017). P3 has the largest host tropism and was sampled from eight out of the 11 bat species in my dataset. The second and third genotypes with the broadest host tropism are P14 and P1 with 10 and seven hosts, respectively. Out of the top seven most sampled strains, P4, P6, and P8 contain primarily human strains but also infect pigs and cows. In addition, there are 14 and 13 genotypes that have infected pigs and cows, respectively. P5, P13, and P7 carry the most pig strains while P11 and P5 have the most cow strains. The genotype with the most strains isolated from pigs is P13 followed by P7.

I was also interested to see if those closely related had a similar host range and vice versa. P9, the sister group to one of the genotypes with the broadest host tropism (P14), contrastingly infects mainly humans. Dissimilar host preferences are further witnessed among other sister taxa: P13/P22, P2/P3, and P11/P29.

POPULATION GENETICS

Quantifying levels of genetic diversity can help us understand relationship dynamics between and within populations. Furthermore, insight into genotypic molecular evolution can bring insight into their phenotypic diversity. The d_X and d_{XY} metric were used to describe the intra- and intergenotypic variation of the seven most sampled genotypes, respectively. Based on both populations' d_{XY}, the amount of genetic variation that can be ascribed to one population was also measured. On a 0–1 scale, it is expected for d_X values to fall under 0.20 and d_{XY} values to be above 0.20 for nucleotide sequence comparisons because of the 80% nucleotide identity genotype classification. VP4 genotypes that possess enough sequences and sites for genotype analysis are those that contain the most human (P8, P4, P6), pig (P13 and P7), and cow (P11 and P5) sequences. Most avian genotypes only contain one sample (Table 3).

The nucleotide d_X values for all seven genotypes are under 0.10 except for P13 (Figure 11). P4 has the lowest d_X values while P13 maintains the most amount of intragenotypic variation. P8 and P4 share the lowest d_{XY} values for each gene and sequence type followed by the remaining P4, P6, and P8 pairwise comparisons, and the pig-dominated genotypes, P7 and P13. As expected, VP8* is the less conserved gene compared to VP5*; however, most VP8* amino acid-based d_{XY} values are greater than its nucleotide-based d_{XY} values, but the opposite is seen for the VP8* d_X values where amino acid diversity exceeds nucleotide diversity. P11 appears to be the most divergent among the other six genotypes for each gene and sequence type. The largest d_{XY} values are between the genotypes with the most cow strains (P11 and P5) and P11 and P6, which also have cow strains.

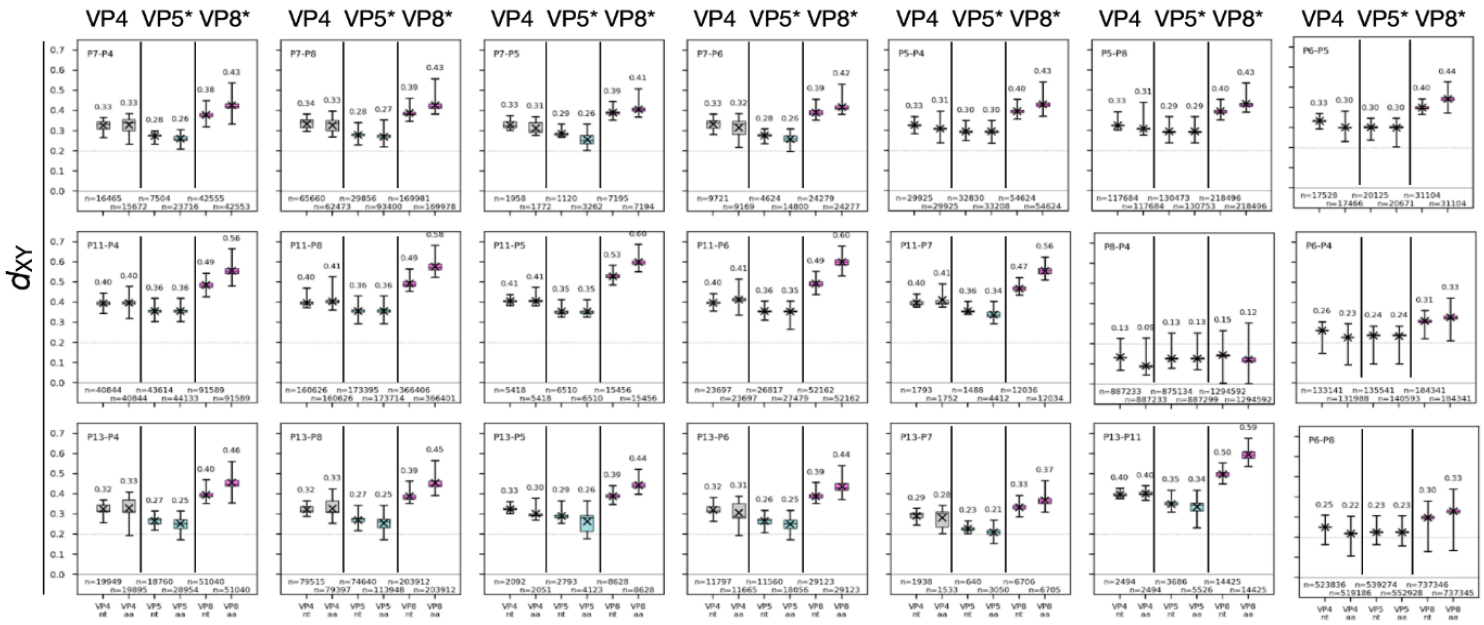


Figure 11: ***Nucleotide and Amino Acid Genotype Histogram Grid.*** *Percentage of Pairwise Comparisons on the vertical axis and dxy and dx values between intergenotypic and intragenotypic comparisons on the horizontal axis, respectively.*

Previous studies have demonstrated that VP8* is more genetically diverse than VP5* since the region consists of domains responsible for host glycan binding and recognition (Hu et al., 2012). The data in this study confirms this notion as there are fewer conserved regions in VP8* than its counterpart (Figure 12). From the amino acid dataset, most VP8* sites carry some variability such as in positions 80–130 and 223–237. VP5*, on the other hand, mostly contains pockets of non-conserved positions with only one region (positions 641-667) with high variability. A d_{XY} comparison between VP8* and VP5* further reinforces this idea since a larger VP8* d_{XY} value than VP4 and VP5* at the nucleotide and amino acid level has been consistently observed (Figure 11).

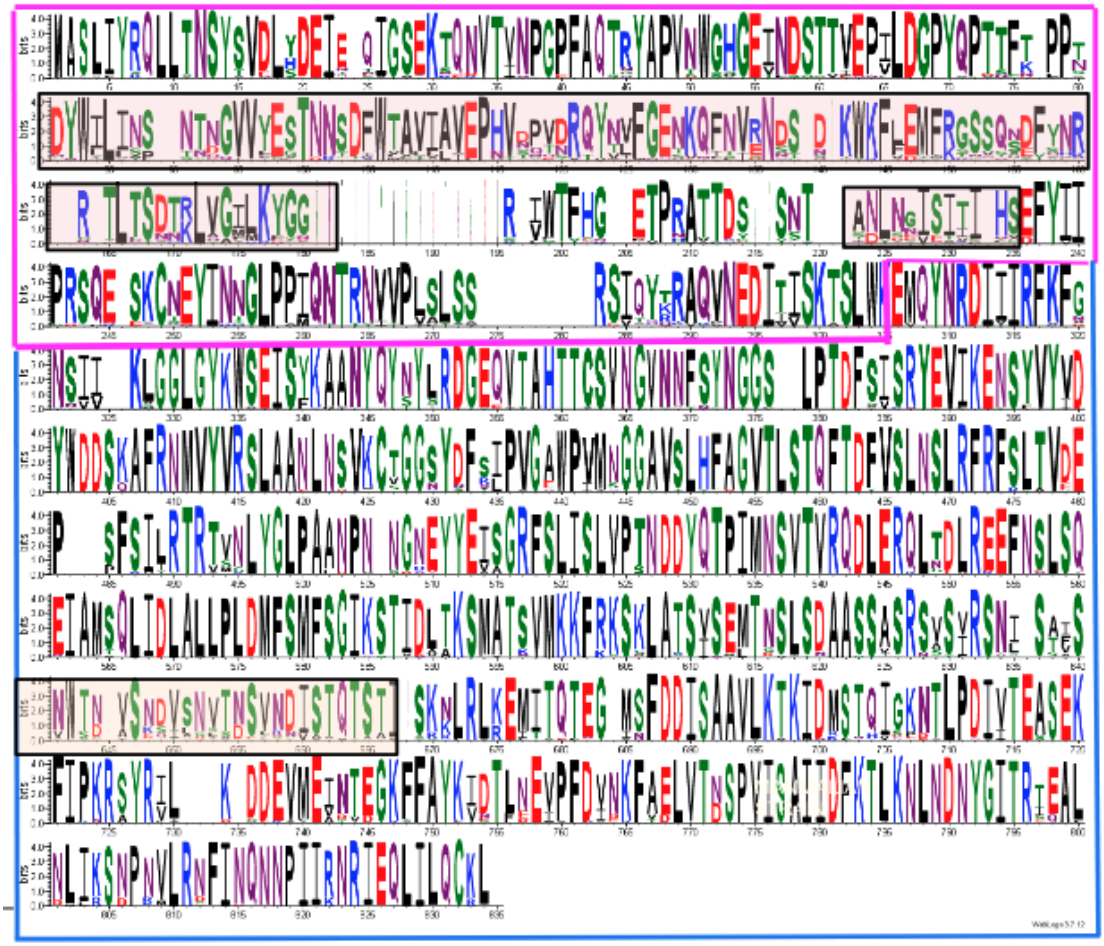


Figure 12: ***VP4 Amino Acid Sequence Logo Plot.*** *VP8* is boxed in magenta and VP5* is boxed in blue. Amino acid residues are colored according to biochemical traits: polar G, S, T, Y, and C (green), neutral Q and N (purple), basic K, R, and H (blue), acidic D and E (red), and hydrophobic A, V, L, I, P, W, F, and M (black). Highly variable regions are highlighted in orange. Created in WebLogo 3.*

The population structure within and between the four major host types (human, pig, cow, and avian) using a reduced dataset was also analyzed. The first three box plots compare two sequences at a time within each genotype. Once each intragenotypic pairwise comparison is made, the number of pairwise differences is averaged. For

example, in humans, pairwise sequence comparisons are first conducted within each genotype (i.e., P4, P8, etc.) and then the total number of pairwise differences between each comparison is averaged. Similarly, the latter three box plots analyze two sequences from two different genotype populations at a time and the total number of pairwise differences among all comparisons is then averaged at the end (Figure 13). The mean intragenotypic genetic diversity between- and within-host is under 0.2 for each gene and sequence type (Figure 14). Surprisingly, interhost comparisons with the avian population exhibit similar amounts of genetic differentiation when compared to mammalian host population comparisons. However, the bird and mammalian host populations are more genetically distinct at the protein level than at the nucleotide level, unlike the pairwise comparisons conducted between pigs, cows, and humans. Protein sequences vary more in the amount of genetic diversity in contrast to nucleotide sequences and the average VP5* intergenotypic diversity in PIG-PIG and AVIAN-AVIAN is near 0.2 at the amino acid level.

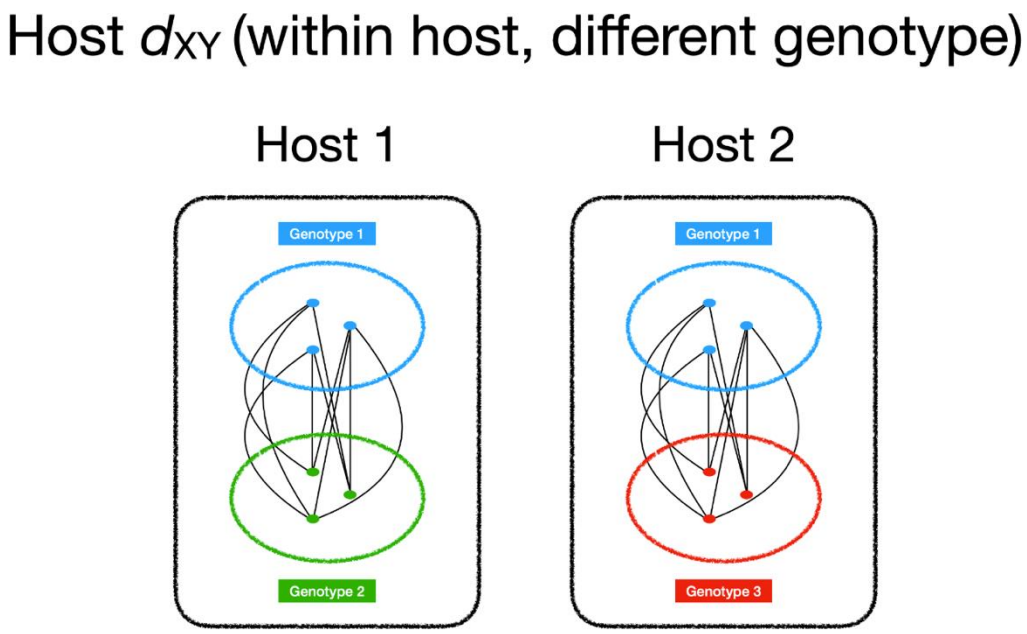


Figure 13: ***Within Host, Different Genotype d_{XY} Demonstration.*** *Blue, green, and red rings represent the genotype populations within each host population (black rectangles). The pairwise distances between two individuals (circles) are illustrated as black lines.*

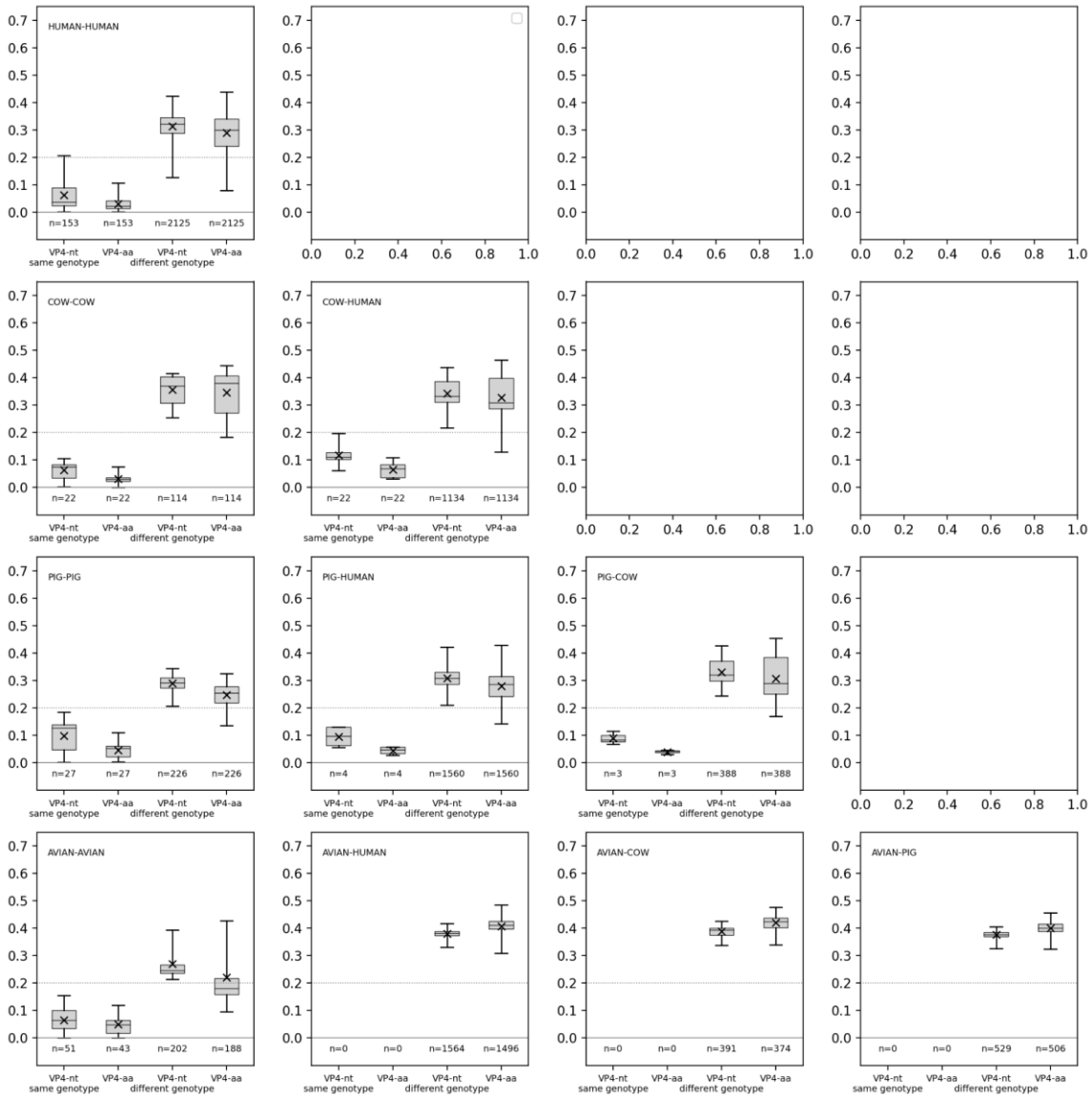


Figure 14: ***Host-Host Nucleotide and Amino Acid Box Plot Grid.*** *Percentage of Pairwise Comparisons on the vertical axis and dxy and dx values between intergenotypic and intragenotypic comparisons on the horizontal axis, respectively.*

The mean PIG-HUMAN intergenotypic variation is lower than the diversity shared between P11 and the most prevalent human genotypes (P8, P4, and P6). Intergenotypic variation within the cow and human populations are higher than the

comparisons between P7 and P13 and P4, P6, and P8, respectively. On the other hand, the genetic diversity between P11 and P5 is greater than the intergenotypic diversity within the pig population. The P11 genogroup may be more genetically diverse than the avian genogroup since P11 contains the largest d_{XY} values.

GENETIC DIVERSITY

Quantifying d_X for both sequence data types within each host type population and for all sequences, both domains follow similar trends. The more conserved regions are within 800–1700 bp (or 267–567 bp at the amino acid level), while spikes in d_X occur around 400 bp and 1800 bp in the nucleotide sequences and 150 bp and 640 bp in the amino acid dataset (Figure 15 and Figure 16). Human and cow sequences have the least and most amount of genetic diversity, respectively.

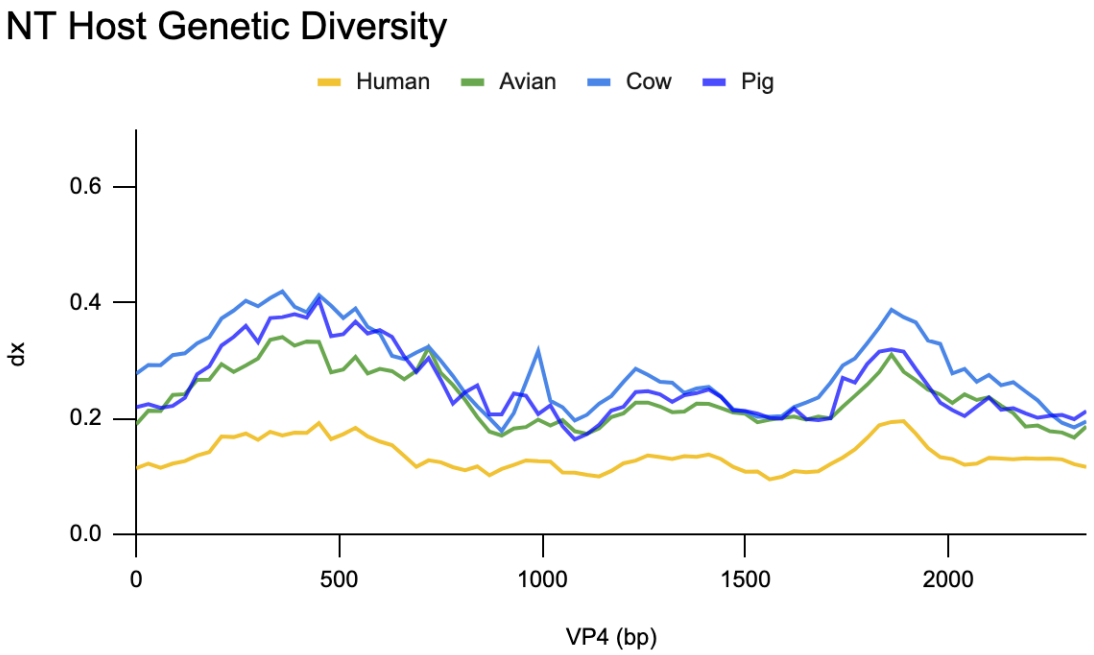


Figure 15: ***VP4 Nucleotide Genetic Diversity Between Hosts.*** d_X *is on the y-axis and the length of VP4 is on the x-axis. 50 bp sliding windows for every 30 bases were used to generate the plot. Visualized in Google Sheets.*

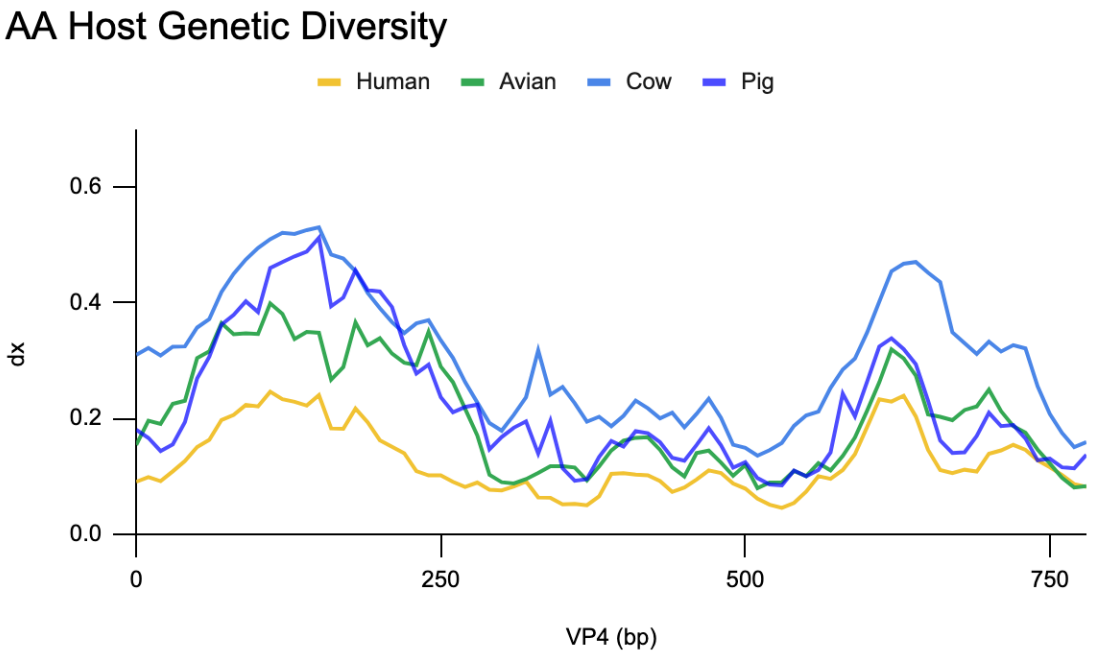


Figure 16: ***VP4 Protein Genetic Diversity Between Hosts.***

$\boldsymbol{d_X}$ is on the y-axis and the length of VP4 is on the x-axis. 50 bp sliding windows for every 30 bases were used to generate the plot. Visualized in Google Sheets.

CONSERVED DOMAIN ANALYSIS

To assess and predict the functional similarities and differences among the four host populations, I searched for conserved domains that exist among all four host types. The two conserved domains found belong to VP5*: 'VP4 helical' and 'Rota VP4 MID' using Motif (https://www.genome.jp/tools/motif/).

As expected, the avian consensus sequence is the most genetically divergent across VP5*. Sites that are 100% variable are at positions 350, 425, 430, 434, 444, 452,

and 490 as well as 549, 638, 648, 653, 670, 675, 706, 740, and 812 in VP5CT and the foot, respectively. Dormitzer et al. (2006) found two motifs in VP5CT of a purified recombinant VP4 protein: the integrin binding region and the membrane interaction loop (Figure 17).

Figure 17: **Host Type Consensus Sequence VP5CT Alignment.** The integrin-binding motif and membrane interaction loop are shaded in orange. Made in Chimera.

The integrin-binding motif is conserved among each host type, and aspartate 100% maintained for all consensus sequences (Figure 18). In contrast, the membrane interaction loop is less conserved and contains five sites that are 100% variable among the four host populations. Interestingly, the pig population has the most sites with the most polymorphisms. Both glycine pairs and proline residues are 100% conserved, but a valine-to-glutamine substitution is found at position 442 in the avian dataset. Even though the domain is mostly conserved, there are 16 positions that carry host-specific amino acid residues.

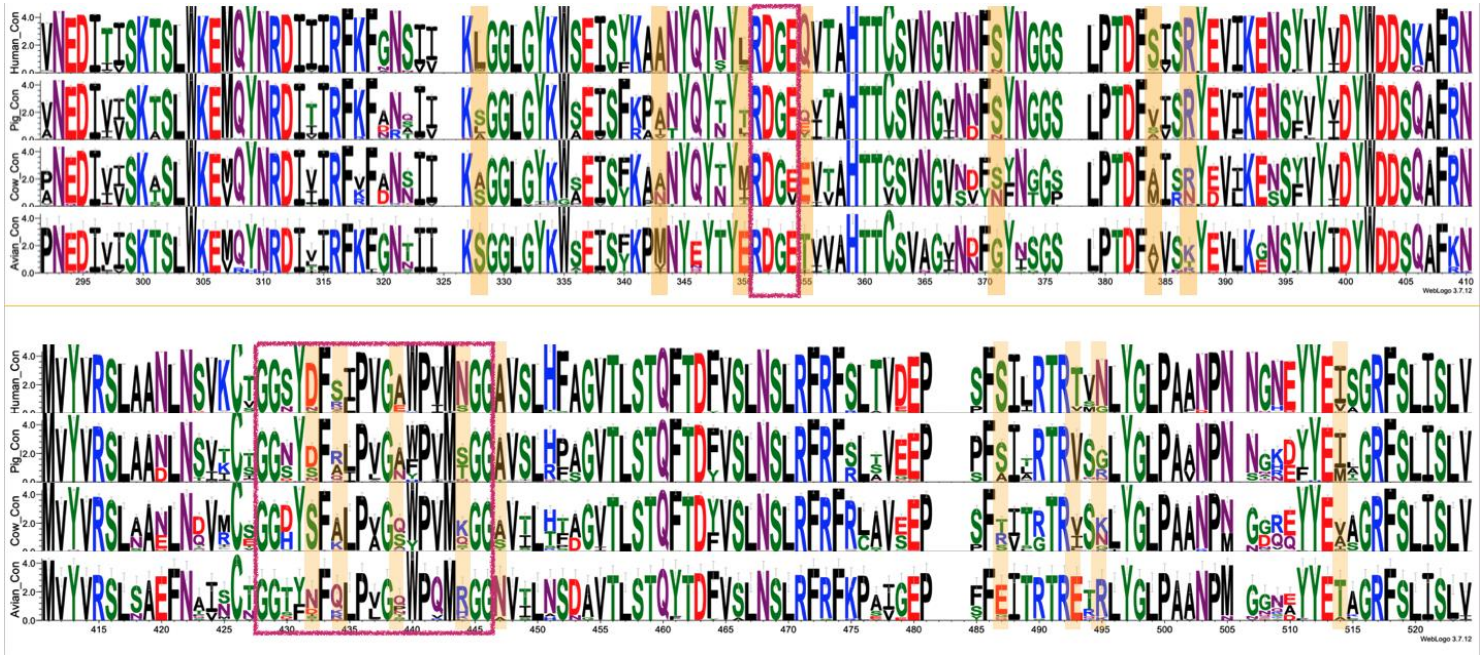


Figure 18: **Host Type Consensus VP5CT Sequence Logo Plots**. Amino acid residues are colored according to biochemical traits: polar G, S, T, Y, and C (green), neutral Q and N (purple), basic K, R, and H (blue), acidic D and E (red), and hydrophobic A, V, L, I, P, W, F, and M (black). The integrin-binding motif and membrane interaction loop are boxed in orange. Made in WebLogo3.

DISCUSSION

RVAs remain the leading cause of acute gastroenteritis in children globally, despite the introduction of vaccines (World Health Organization [WHO], 2023). Genetic diversity plays an important role in their prevalence and a major driver of RV genetic diversity is host tropism and interspecies transmission. Vaccines can place selective pressures on RVs to acquire mutations that help them evade the vaccine-induced host immune response. The avian and mammalian strains exhibit strong genetic distinctiveness from one another, however, the clear genetic separation between P11 and other genotypes suggests that P11 may be more divergent than the avian clade. Although VP8* is studied more than its counterpart, VP5*, VP5* possesses a range of diverse regions that further encourages the idea that more VP5* functional and structural research is needed. Within VP5CT, there is consistency in the location of variable and conserved sites. Sites of high variability and change may be associated with membrane interaction. My project seeks to evaluate intra- and interpopulation differentiation on multiple levels by comparing hosts, genotypes, VP4 subunits, and sequence types to address the relationship between host tropism and genotype diversity.

PHYLOGENETIC AND nMDS ANALYSES

Nucleotide sequences are generally more informative than amino acid sequences because many third-position codon nucleotide changes are silent and do not change the amino acid. As the number of nucleotide substitutions increases over time, the phylogenetic signal decreases since certain substitutions, especially those that happen at

the same site, cannot be tracked (Townsend et al., 2008). In addition, highly conserved taxa would be hard to discern at the protein level and partial sequences may not provide enough informative sites, further dampening the phylogenetic signal (Pease et al., 2018). For these reasons, inferring both a nucleotide and amino-acid phylogenetic tree and measuring the amount of concordance of each branch through QS provides a more complete picture of the evolutionary history.

Rooting a tree defines the evolutionary history of the taxa and the location of the root signals where the most recent common ancestor (MRCA) is for all descendants (Kinene et al., 2016). While the model allows for various root placements, choosing the outgroup rooting method seemed the most fitting for this study since studies have shown that mammalian RVA sequences are distantly related to avian RVA strains (Johne et al., 2016). Phylogenetic analysis revealed an early split between the avian and mammalian RVA strains that explains the low genetic similarity between the two classes (Ito et al., 2001) and the distinct separation from other genogroups (Figure 7 and Figure 8). The tight clustering among RVAs from different bird species suggests that interspecies transmission of the virus among other birds is common. Much of the RV avian strains sampled were found in farms or live animal markets (Pauly et al., 2017). Avian RVA VP4 genotypes P56 and P38 are grouped in the avian clade, suggesting that the two genotypes belong to avian genogroup V (Figure 4). Interestingly, there is some separation between the two avian P37 strains and the avian clade on the trees and in the aa nMDS analysis, signaling that the genotype may contain evidence for an interspecies transmission event between a mammal and bird. Though rare, it is not impossible since a

P17 avian strain was sampled from a red fox (*Vulpes vulpes*) in Italy in 2011 (Pauly et al., 2017).

Since the genogroups were established using one strain from each VP8* genotype, it is not surprising that the intergenotypic relationships are not conserved between Liu et al.'s (2012) and this study since this project uses both subunits to infer the phylogeny of VP4. Although VP8* is more genetically diverse than VP5*, the nonconserved tree topology across studies suggests that VP5* also plays a major role in RVA evolution. Furthermore, high levels of discordance and low support for some branching patterns may be evidence of convergent evolution. Reassortment and interspecies transmission events can lead to the emergence of highly virulent novel strains that independently evolve to accrue similar mutations that are positively selected for (Nunes et al., 2022).

P4 AND P8 GENOTYPES

P8 is the most prevalent RVA VP4 genotype globally, followed by P4 (Zeller et al., 2012; Heylen et al., 2016; Hu et al., 2018). P4 and P8 sequences are rarely more than 80% different from each other and are the two most genetically similar of the seven most common genotypes (Figure 11). The overlap exhibited in the nMDS plots suggests that both genotypes are almost identical to one another, apart from a few divergent strains (Figure 9 and Figure 10). It is common to find both P8 and P4 circulating in the same cities and targeting similar demographic populations (Liu et al., 2012). And these shared preferences in population demographics, host tropism, and host cell receptors may explain why it may be difficult to discern P8 from P4. The nesting of P8 inside P4 in the

amino acid phylogeny suggests that P8 may have diverged from P4. Although P4 has the least amount of intragenotypic diversity, geographic isolation and other factors may have resulted in the divergence of several P4 strains from the founding population. P8 and P4 may be incredibly genetically similar at the nucleotide and amino-acid levels; however, RV serotyping has confirmed the presence of several diagnostic antigens between the two most prevalent VP4 genotypes (Matthijnssens et al., 2008).

Genome-wide assessments are also widely used to distinguish P8 from P4. In Brazil, a six-year (2010–2016) surveillance documented that genotype combinations with P8 were mainly part of the Wa-like constellation, and combinations with P4 were strictly DS-1-like (Silva-Sales et al., 2020). An assessment of the genome is an accurate method of surveillance as certain genotype combinations are found more often than others (Matthijnssens et al. 2008; Silva-Sales et al., 2020). However, pairing genomic analysis with serotyping for antigenic differences provides the most comprehensive analysis for studying RV evolution.

THE FOUR HOST POPULATIONS

The top seven most sampled genotypes exhibit a narrow host tropism (Table 3) but they are capable of infecting other species, which hints at possible host jumping and/or reassortment. Higher intergenotypic genetic variation within a host versus the overall genetic differentiation between two genotypes is expected; however, that is not the case between the two most sampled predominantly cow genotypes (P11 and P5) and the intergenotypic diversity within the pig sequences. P11 is the most divergent genotype in this study and shares one of the largest d_{XY} values with P5. P11's low genetic

similarity to other genogroups suggests that its host tendencies to cattle and neonates may be the reason since it poses unique host binding specificities. Crystal structure analysis reveals P11 exhibiting a noticeably different VP8* conformation than the other genotypes in Sun et al.'s (2021) study. Moreover, human P11 sequences possess a unique binding site where it specifically binds to poly-LacNAc, a type I precursor of human blood group antigens or HBGAs only found in newly born human infants (Liu et al., 2013; Sun et al., 2021). Previous studies have also retrieved P11 cattle-neonate RV reassortants, an unusual host jump mostly sampled in India (Ramani et al., 2013) and in the red fox, which is known to carry other P genotypes: P3, P23, P13, P21, and P14. The genetic similarity between red fox P11 sequences and P11 cattle sequences hints at interspecies transmission events between the two hosts have occurred (Čolić et al., 2021). The distinctive zoonotic patterns of P11 emphasize the concern for the increased risk of zoonotic transmission and reassortment events since high genetic variation can lead to host tropism expansion and improved virulence.

The use of d_{XY} helped explain some observations made from the VP4 phylogeny and the nMDS analysis. The d_{XY} values of 11 out of the 21 genotype pair comparisons are in the low 0.30s. Close interactions leading to interspecies transmission between the three species since the transition from hunting to farming may have played a role in their genetic similarity (Papp et al., 2013). Genomic analysis reveals RVA pig and cow strains share a common ancestor between the dominant Wa-like and DS-1-like human genogroups, respectively (Matthijnssens et al., 2008). Similar levels of genetic differentiation seen between all four host populations suggest that host is not a major driver of VP4 evolution. This does not mean that host does not play a role in the overall

evolution of rotaviruses, however, for VP4, studying the underlying mechanism that drive VP8* diversity and VP4 interactions with other proteins such as VP7 may prove to be useful in understanding VP4 evolution.

The most predominant genotype is P7 in pigs and P5 in cows. However, this project did not account for annual geographical fluctuations. Although they are the most sampled, these genotypes were found less frequently in some regions and in some years (Papp et al., 2013). Porcine is the second most common host in my dataset yet P7 and P13 share one of the lowest d_{XY} values. Pigs are known to be reservoir hosts and their dense populations lend for easy RV interspecies transmission (Papp et al., 2013).

Another point of interest is the low intragenotypic variation exhibited within and between populations. RNA viruses exhibit large amounts of genetic variation due to their high mutation rates, short generation times, and large population sizes (Rubio et al., 2013). However, low intragenotypic variation is not uncommon and may be due to extrinsic factors such as natural selection or rapid transmission cycles leading to genetic bottlenecks (Li and Roossinck, 2004).

DOMAIN ANALYSES

Due to the number of partial sequences, any domain hits not found in all four host types were excluded to reduce the chance of inaccurate functional site calling. Fragmented sequences cannot provide enough information to report accurate search returns and have a higher chance of matching to unrelated domains by chance. In addition, domains are annotated at the domain level, meaning that Pfam predictions do not account for domain shuffling, which can result in incorrect descriptions (Friedberg

2006). For these reasons any matches within VP8* were ignored. A small region of VP8* was only returned in one consensus sequence (pig). Due to the low sequence similarity among the consensus sequences (Figure 19) and that most of the genetic diversity lies in VP8* (Figure 12), a lack of hits within VP8* was expected.

The virion's reliance on the initial interactions between its host and VP8* to attach to the cell suggests that VP8* is the determinant of host tropism (Stencel-Baerenwald et al., 2014; Hu et al., 2018). Sialic acid, mucin core, HBGAs, and heat shock proteins are just some known classes of host receptors and the polymorphisms within each group have made it impossible to exclusively characterize VP8* genotypes by host or binding preferences (Liu et al., 2012; Hu et al., 2018). Although many papers focus on the functional significance of VP8* over VP5*, I was unable to compare the various motifs found within VP8* due to this reason. Furthermore, these functional regions are based on a multitude of factors and not just genetic similarity or host tropism alone (Liu et al., 2012; Hu et al., 2018; Sun et al., 2021). Also, assessing VP5* and its conserved domains (CD) lends a fair analysis between populations with different evolutionary histories as they can be used to predict function across different hosts.

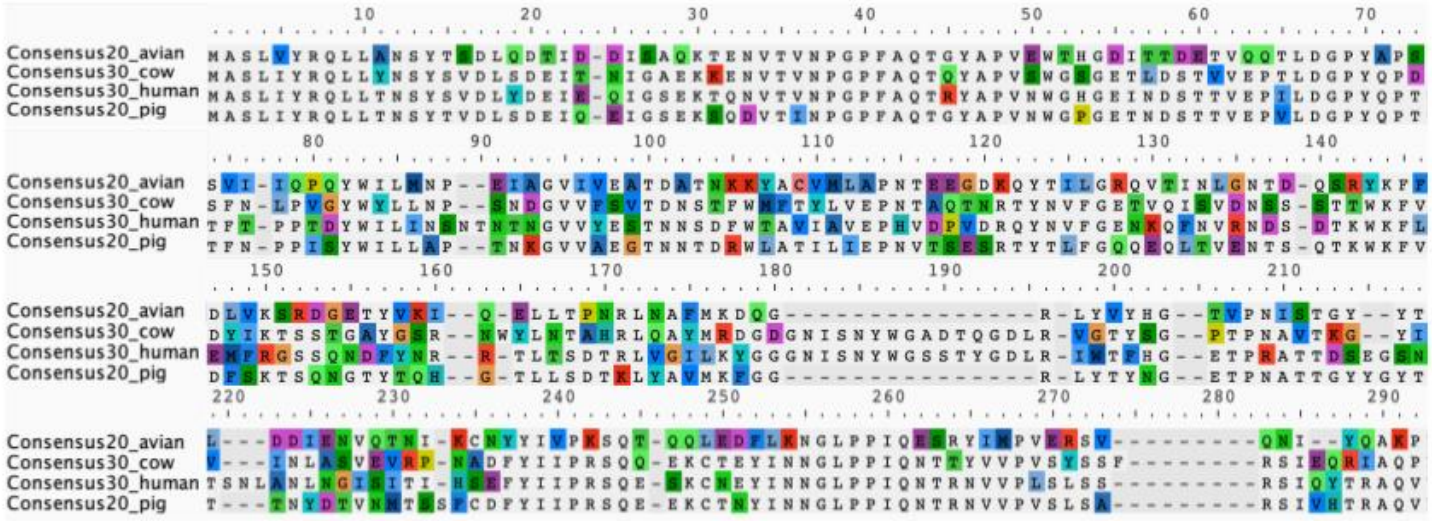


Figure 19: ***Host Type Consensus Sequence VP8* Alignment****. Non-conserved positions are colored. Visualized in Aliview.*

The conservation of VP5* suggests that it is under strong selection to preserve essential functions such as virion stability and intergenotypic interactions (Liu et al., 2008). The conserved glycine pairs (428-429 and 445-446) may be responsible for membrane adhesion before penetration (Dormizter et al., 2004). VP5CT contains an α2β1 integrin site and is conserved in all four host types. Integrins consist of an α and β subunit and belong to a family of glycoproteins used for cell attachment and recognition. Previous studies have shown this site to be conserved in human and non-human mammals but is speculated to be also maintained in avian species (Coulson et al., 1997; Seo et al., 2008). The multitude of variable sites in VP5* further implies that there are more binding sites than we know. Other integrins such as α4β1, αxβ2, and αvβ3 and Hsc70, a member of the heat shock protein family are theorized to serve as post-receptors (Zárate et al., 2003; Amimo et al., 2021).

On the other hand, the membrane interaction loop also possesses sites with high variability suggesting that different species may have different interaction mechanisms within VP5*. The CD loop may be involved with integrin binding and motifs that may interact with the membrane include the BC, ED, and FG loops (See Figure 2b in Dormizter et al., 2004). Amino acid substitutions are thought to be uncommon but if they do happen, they are more likely to happen with residues with similar biochemical characteristics since a change in amino acid can significantly affect protein function (Zhang, 2000). There are several instances of host specific amino acid residues in VP5CT such as the polar and uncharged Q replaced by a nonpolar V among the avian sequences. A polar to non-polar amino acid substitution and the reverse does not always result in

noticeable effects on function (Clarke et al., 1990) but the motifs denoted by Dormitzer et al. (2004) have not been tested in this study. Avian strains possess different antigenic properties than mammalian strains (Dhama et al., 2015), therefore, a deeper assessment of this mutation and its effects on membrane interaction is required for further interpretation.

GENBANK SEQUENCES

This study analyzed 4071 VP4 sequences, however, many available sequences are not full-length coding sequences. Therefore, the inclusion of partials may have hampered the phylogenetic signal. Through filtering and metadata correction, several mislabeled strains were uncovered either with the wrong genotype or species of RV. After undergoing an extensive filtering process, I presume any potential reassortants were excluded since there are no lone genotype sequences suspiciously branched to a clade or genotype away from its own population.

SAMPLING BIAS

I do my best to decrease sampling bias, but I am aware that some sampling bias is inevitable. My goal is to retain as many genotypes as possible, but that comes with the tradeoff of having under- and over-represented populations. There are also limitations to using consensus sequences. Consensus sequences lose some of their heterogeneity to offset the possibility of an ambiguity code so areas with the least conservation may be underrepresented. Heavy reliance on the accuracy of GenBank entries to represent the current global geospatial distribution of genotypes will also present some bias. Most strains were from positive human cases; therefore, the overall animal genotype diversity

and its mechanisms for interspecies transmission is still largely unknown. Moreover, parts of the world have done little to no surveillance. Having a better sense of the living and shared conditions among humans and animals in the developing world and documenting annual RV prevalence will provide needed sentiments regarding host susceptibility and host tendencies. I hypothesize that population size and the number of comparisons are the reasons for this contradiction.

CONCLUSION

Expanding my understanding of intra- and intergenotypic diversity with massive gene alignment analyses poses several implications. Visualizing genotypic sequence diversity grows my knowledge of how strain abundance and level of divergence might correspond to host shifts. If we compare sequence diversity between populations with shared attributes such as host tropism, we can determine and better predict which least conserved sites may lead to lineage splits. Similarly, mapping regions prone to change in key domains of VP4 will bring insight into the changes associated with their complex structure, function, and interactions (Bordin et al., 2021). Ultimately, connecting genetic diversity to functional diversity will deepen my capacity to understand the relationship between nucleotide and protein diversity and how their evolving rates overall shape RV evolution.

Studying VP4-VP7 interactions will allow us to obtain a clearer portrayal of the mechanisms that drive genotype and overall RV diversification. Likewise, studying the two proteins in tandem may shed light on means to improve vaccine development and can enhance my predictive capabilities in strain prevalence. Also, measuring their genetic diversity before and after the distribution of vaccines may help us rationalize if vaccines drive diversity or not.

My project relied on the knowledge and work of other scientists, which limits the depth of my analysis and discussion. With little research on VP5* compared to VP8*, this study emphasizes the growing need for more analysis to be done in this region. The more conserved domain may prove to be a potential alternative for classifying RV genotypes. In addition, determining VP5*'s crystal structure and conducting binding

assays to locate functional sites across different hosts is another avenue I would like to pursue in the future. Surveillance in underreported countries and more animal RV testing across a wider array of species is a critical step toward understanding the full extent of RV prevalence and host susceptibility worldwide.

REFERENCES

Arias, Carlos F., et al. "Rotavirus Entry: A Deep Journey into the Cell with Several Exits." *Journal of Virology*, vol. 89, no. 2, Jan. 2015, pp. 890–93, https://doi.org/10.1128/jvi.01787-14.

Abdelhakim, Aliaa H., et al. "Structural Correlates of Rotavirus Cell Entry." *PLoS Pathogens*, vol. 10, no. 9, Sept. 2014, https://doi.org/10.1371/journal.ppat.1004355.

Amimo, Joshua Oluoch, et al. "Rotavirus Interactions With Host Intestinal Epithelial Cells." *Frontiers in Immunology*, vol. 12, no. 793841, Dec. 2021, https://doi.org/10.3389/fimmu.2021.793841.

Bishop, Ruth. "Discovery of Rotavirus: Implications for Child Health - PubMed." *Journal of Gastroenterology and Hepatology*, vol. 24 Suppl 3, no. S81-5, Oct. 2009, https://doi.org/10.1111/j.1440-1746.2009.06076.x.

Bordin, Nicola, et al. "Tracing Evolution Through Protein Structures: Nature Captured in a Few Thousand Folds." *Frontiers in Molecular Biosciences*, vol. 8, May 2021, https://doi.org/10.3389/fmolb.2021.668184.

Bourdett-Stanziola, Lurys. "Genetic Diversity of Bat Rotaviruses." *Microbiology Research Journal International*, vol. 32, no. 1, May 2022, pp. 17–31, https://doi.org/10.9734/mrji/2022/v32i130365.

Busi, Chiara, et al. "Group A Rotavirus Associated with Encephalitis in Red Fox." *Emerging Infectious Diseases*, vol. 23, no. 9, Sept. 2017, https://doi.org/10.3201/eid2309.170158.

Clarke, D. M., et al. “Functional Consequences of Alterations to Polar Amino Acids Located in the Transmembrane Domain of the Ca2(+)-ATPase of Sarcoplasmic Reticulum - PubMed.” *The Journal of Biological Chemistry*, vol. 265, no. 11, Apr. 1990.

Čolić, Daniel, et al. “A Remarkable Genetic Diversity of Rotavirus A Circulating in Red Fox Population in Croatia.” *Pathogens (Basel, Switzerland)*, vol. 10, no. 4, Apr. 2021, p. 485, https://doi.org/10.3390/pathogens10040485.

Coulson, B. S., et al. “Rotavirus Contains Integrin Ligand Sequences and a Disintegrin-like Domain That Are Implicated in Virus Entry into Cells.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 10, May 1997, pp. 5389–94, https://doi.org/10.1073/pnas.94.10.5389.

Crawford, Sue E., et al. “Rotavirus Infection.” *Nature Reviews. Disease Primers*, vol. 3, no. 17083, Nov. 2017, https://doi.org/10.1038/nrdp.2017.83.

Crooks, Gavin E., et al. “WebLogo: A Sequence Logo Generator.” *Genome Research*, vol. 14, no. 6, June 2004, pp. 1188–90, https://doi.org/10.1101/gr.849004.

de Barros, Bruno de Cássio Veloso, et al. “Rotavirus A in Wild and Domestic Animals from Areas with Environmental Degradation in the Brazilian Amazon.” *PLOS ONE*, vol. 13, no. 12, Dec. 2018, https://doi.org/10.1371/journal.pone.0209005.

Dhama, Kuldeep, et al. “Avian Rotavirus Enteritis – an Updated Review.” *Veterinary Quarterly*, vol. 35, no. 3, May 2015, pp. 142–58, https://doi.org/10.1080/01652176.2015.1046014.

Donato, Celeste M., and Julie E. Bines. “Rotaviruses and Rotavirus Vaccines.” *Pathogens*, vol. 10, no. 8, Aug. 2021, https://doi.org/10.3390/pathogens10080959.

Dormitzer, Philip R., et al. “Structural Rearrangements in the Membrane Penetration Protein of a Non-Enveloped Virus.” *Nature*, vol. 430, no. 7003, Aug. 2004, https://doi.org/10.1038/nature02836.

Douam, Florian, et al. “Genetic Dissection of the Host Tropism of Human-Tropic Pathogens.” *Annual Review of Genetics*, vol. 49, no. 21–45, Jan. 2015, https://doi.org/10.1146/annurev-genet-112414-054823.

Edgar, Robert C. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research*, vol. 32, no. 5, Jan. 2004, https://doi.org/10.1093/nar/gkh340.

Estes, M. K., and J. Cohen. “Rotavirus Gene Structure and Function.” *Microbiological Reviews*, vol. 53, no. 4, Dec. 1989, pp. 410–49, https://doi.org/10.1128/mr.53.4.410-449.1989.

Friedberg. “Automated Protein Function Prediction—the Genomic Challenge.” *Briefings in Bioinformatics*, vol. 7, no. 3, Sept. 2006, pp. 225–42, https://doi.org/10.1093/bib/bbl004.

Geletu, Umer Seid, et al. “Rotavirus in Calves and Its Zoonotic Importance.” *Veterinary Medicine International*, vol. 2021, no. 6639701, Apr. 2021, https://doi.org/https://doi.org/10.1155/2021/6639701.

George, Sincy, et al. “Genetic Diversity of Rotavirus Strains in the Era of Vaccination: A Pilot Study from Central Kerala, India.” *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH*, 2018, https://doi.org/10.7860/jcdr/2018/35403.11964.

Glass, et al. “The Rotavirus Vaccine Story: From Discovery to the Eventual Control of Rotavirus Disease.” *The Journal of Infectious Diseases*, vol. 224, no. Supplement_4, Sept. 2021, pp. S331–42, https://doi.org/10.1093/infdis/jiaa598.

Gorziglia, M., et al. “Conservation of Amino Acid Sequence of VP8 and Cleavage Region of 84-kDa Outer Capsid Protein among Rotaviruses Recovered from Asymptomatic Neonatal Infection.” *Proceedings of the National Academy of Sciences*, vol. 83, no. 18, Sept. 1986, pp. 7039–43, https://doi.org/10.1073/pnas.83.18.7039.

Gouy, Manolo, et al. “Seaview Version 5: A Multiplatform Software for Multiple Sequence Alignment, Molecular Phylogenetic Analyses, and Tree Reconciliation - PubMed.” *Methods in Molecular Biology (Clifton, N.J.)*, vol. 2231, Jan. 2021, pp. 241–60, https://doi.org/10.1007/978-1-0716-1036-7_15.

Greenberg, Harry B., and Mary K. Estes. “Rotaviruses: From Pathogenesis to Vaccination.” *Gastroenterology*, vol. 136, no. 6, May 2009, https://doi.org/10.1053/j.gastro.2009.02.076.

Hahn, Matthew William. *Molecular Population Genetics*. Sinauer Associates, Incorporated, 2018.

Han, Mira V., and Christian M. Zmasek. “phyloXML: XML for Evolutionary Biology and Comparative Genomics.” *BMC Bioinformatics*, vol. 10, no. 356, Oct. 2009, https://doi.org/10.1186/1471-2105-10-356.

Hewish, Marilyn J., et al. “Integrins A2β1 and A4β1 Can Mediate SA11 Rotavirus Attachment and Entry into Cells.” *Journal of Virology*, vol. 74, no. 1, Jan. 2000, https://doi.org/10.1128/jvi.74.1.228-236.2000.

Heylen, et al. “Human P[6] Rotaviruses From Sub-Saharan Africa and Southeast Asia Are Closely Related to Those of Human P[4] and P[8] Rotaviruses Circulating Worldwide.” *The Journal of Infectious Diseases*, vol. 214, no. 7, July 2016, pp. 1039–49, https://doi.org/10.1093/infdis/jiw247.

Hu, et al. “Glycan Recognition in Globally Dominant Human Rotaviruses.” *Nature Communications*, vol. 9, no. 1, July 2018, pp. 1–12, https://doi.org/10.1038/s41467-018-05098-4.

International Committee on Taxonomy of Viruses. “Genus: Rotavirus.” *ICTV*, 2023, https://ictv.global/report/chapter/sedoreoviridae/sedoreoviridae/rotavirus. Accessed 21 Mar. 2023.

Ito, H., et al. “Complete Nucleotide Sequence of a Group A Avian Rotavirus Genome and a Comparison with Its Counterparts of Mammalian Rotaviruses - PubMed.” *Virus Research*, vol. 75, no. 2, June 2001, https://doi.org/10.1016/s0168-1702(01)00234-9.

Jenni, Simon, et al. “Rotavirus VP4 Epitope of a Broadly Neutralizing Human Antibody Defined by Its Structure Bound with an Attenuated-Strain Virion.” *Journal of Virology*, vol. 96, no. 16, Aug. 2022, https://doi.org/10.1128/jvi.00627-22.

Jiang, Xi, et al. “Histo-Blood Group Antigens as Receptors for Rotavirus, New Understanding on Rotavirus Epidemiology and Vaccine Strategy.” *Emerging Microbes & Infections*, vol. 6, no. 4, Apr. 2017, https://doi.org/10.1038/emi.2017.30.

Johne, Reimar, et al. "Generation of an Avian-Mammalian Rotavirus Reassortant by Using a Helper Virus-Dependent Reverse Genetics System." *Journal of Virology*, vol. 90, no. 3, Feb. 2016, https://doi.org/10.1128/JVI.02730-15.

Katoh, Kazutaka, and Daron M. Standley. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution*, vol. 30, no. 4, Apr. 2013, pp. 772–80, https://doi.org/10.1093/molbev/mst010.

Kim, Irene S., et al. "Effect of Mutations in VP5* Hydrophobic Loops on Rotavirus Cell Entry." *Journal of Virology*, vol. 84, no. 12, June 2010, https://doi.org/10.1128/JVI.02461-09.

Kinene, T., et al. "Rooting Trees, Methods For." *Encyclopedia of Evolutionary Biology*, Elsevier, 2016, pp. 489–93, http://dx.doi.org/10.1016/b978-0-12-800049-6.00215-8. Accessed 21 Mar. 2023.

Kuang, Xiaozhou, et al. "Genetic Diversity of Group A Rotavirus in Acute Gastroenteritis Outpatients in Shanghai from 2017 to 2018." *BMC Infectious Diseases*, vol. 20, no. 1, Aug. 2020, https://doi.org/10.1186/s12879-020-05279-x.

Kyoto University Bioinformatics Center. "MOTIF: Searching Protein Sequence Motifs." *GenomeNet*, Sept. 1991, https://www.genome.jp/tools/motif/. Accessed 24 Mar. 2023.

Laboratory of Viral Metagenomics. "Rotavirus Classification Working Group: RCWG." *Laboratory of Viral Metagenomics*, Rega Instituut KU Leuven, 25 Mar. 2021, https://rega.kuleuven.be/cev/viralmetagenomics/virus-classification/rcwg. Accessed 26 Mar. 2023.

Larsson. "AliView: A Fast and Lightweight Alignment Viewer and Editor for Large Datasets." *Bioinformatics*, vol. 30, no. 22, Aug. 2014, pp. 3276–78, https://doi.org/10.1093/bioinformatics/btu531.

Leung, Alexander K. C., et al. "Rotavirus Gastroenteritis - PubMed." *Advances in Therapy*, vol. 22, no. 5, Oct. 2005, https://doi.org/10.1007/BF02849868.

Li, Yuqing, et al. "The Importance of Glycans of Viral and Host Proteins in Enveloped Virus Infection." *Frontiers in Immunology*, vol. 12, no. 638573, Apr. 2021, p. 638573, https://doi.org/10.3389/fimmu.2021.638573.

Liu, et al. "Natural Selection of Protein Structural and Functional Properties: A Single Nucleotide Polymorphism Perspective." *Genome Biology*, vol. 9, no. 4, Apr. 2008, pp. 1–17, https://doi.org/10.1186/gb-2008-9-4-r69.

Liu, Yang, Pengwei Huang, Baoming Jiang, et al. "Poly-LacNAc as an Age-Specific Ligand for Rotavirus P[11] in Neonates and Infants." *PLOS ONE*, vol. 8, no. 11, Nov. 2013, https://doi.org/10.1371/journal.pone.0078113.

Liu, Yang, Pengwei Huang, Ming Tan, et al. "Rotavirus VP8*: Phylogeny, Host Range, and Interaction with Histo-Blood Group Antigens." *Journal of Virology*, vol. 86, no. 18, Sept. 2012, https://doi.org/10.1128/JVI.00979-12.

Louten, Jennifer. "Virus Structure and Classification." *Essential Human Virology*, Elsevier, 2016, pp. 19–29, http://dx.doi.org/10.1016/b978-0-12-800947-5.00002-8. Accessed 21 Mar. 2023.

Maes, et al. “RotaC: A Web-Based Tool for the Complete Genome Classification of Group A Rotaviruses.” *BMC Microbiology*, vol. 9, no. 1, Nov. 2009, pp. 1–4, https://doi.org/10.1186/1471-2180-9-238.

Malik, Yashpal S., et al. “Evolving Rotaviruses, Interspecies Transmission and Zoonoses ~ Fulltext.” *The Open Virology Journal*, vol. 14, no. 1, Mar. 2020, https://doi.org/doi:10.2174/1874357902014010001.

Matthijnssens, et al. “Uniformity of Rotavirus Strain Nomenclature Proposed by the Rotavirus Classification Working Group (RCWG).” *Archives of Virology*, vol. 156, no. 8, May 2011, pp. 1397–413, https://doi.org/10.1007/s00705-011-1006-z.

Matthijnssens, Jelle, Max Ciarlet, Erica Heiman, et al. “Full Genome-Based Classification of Rotaviruses Reveals a Common Origin between Human Wa-Like and Porcine Rotavirus Strains and Human DS-1-Like and Bovine Rotavirus Strains.” *Journal of Virology*, vol. 82, no. 7, Apr. 2008, https://doi.org/10.1128/JVI.02257-07.

Matthijnssens, Jelle, Max Ciarlet, Mustafizur Rahman, et al. “Recommendations for the Classification of Group A Rotaviruses Using All 11 Genomic RNA Segments.” *Archives of Virology*, vol. 153, no. 8, Jan. 2008, https://doi.org/10.1007/s00705-008-0155-1.

McDonald, Sarah M., et al. “Evolutionary Dynamics of Human Rotaviruses: Balancing Reassortment with Preferred Genome Constellations.” *PLoS Pathogens*, vol. 5, no. 10, Oct. 2009, p. e1000634, https://doi.org/10.1371/journal.ppat.1000634.

Nei, M., and W. H. Li. “Mathematical Model for Studying Genetic Variation in Terms of Restriction Endonucleases.” *Proceedings of the National Academy of Sciences of*

the United States of America, vol. 76, no. 10, Oct. 1979, https://doi.org/10.1073/pnas.76.10.5269.

Nei, M., and F. Tajima. "DNA Polymorphism Detectable by Restriction Endonucleases." *Genetics*, vol. 97, no. 1, Jan. 1981, pp. 145–63, https://doi.org/10.1093/genetics/97.1.145.

Nguyen, Lam-Tung, et al. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution*, vol. 32, no. 1, Jan. 2015, https://doi.org/10.1093/molbev/msu300.

Nunes, Danilo Rosa, et al. "Deep Phylogenetic-Based Clustering Analysis Uncovers New and Shared Mutations in SARS-CoV-2 Variants as a Result of Directional and Convergent Evolution." *PLOS ONE*, vol. 17, no. 5, May 2022, https://doi.org/10.1371/journal.pone.0268389.

Papp, Hajnalka, et al. "Review of Group A Rotavirus Strains Reported in Swine and Cattle." *Veterinary Microbiology*, vol. 165, no. 3, Aug. 2013, https://doi.org/10.1016/j.vetmic.2013.03.020.

Pauly, et al. "Molecular Epidemiology of Avian Rotaviruses Group A and D Shed by Different Bird Species in Nigeria." *Virology Journal*, vol. 14, no. 1, June 2017, pp. 1–10, https://doi.org/10.1186/s12985-017-0778-5.

Patton, John T. "Rotavirus Diversity and Evolution in the Post-Vaccine World." *Discovery Medicine*, vol. 13, no. 68, Jan. 2012.

Pease, James, B. "RotaDB Guide v2.0.1." *Google Docs*, 19 Oct. 2021, https://docs.google.com/document/d/1bVRKpMRmeyq4shPmWZ-

Vplw41D32GH31UIu_STuvYG4/edit#heading=h.etqqhyzb4tj6. Accessed 21 Mar. 2023.

Pease, James B., et al. “Quartet Sampling Distinguishes Lack of Support from Conflicting Support in the Green Plant Tree of Life - PubMed.” *American Journal of Botany*, vol. 105, no. 3, Mar. 2018, https://doi.org/10.1002/ajb2.1016.

Python Software Foundation. “The Python Standard Library.” *Python Documentation*, 21 Mar. 2023, https://docs.python.org/3.11/library/index.html. Accessed 21 Mar. 2023.

Ramani, Sasirekha, Nicolas W. Cortes-Penfield, et al. “The VP8(Star) Domain of Neonatal Rotavirus Strain G10P[11] Binds to Type II Precursor Glycans.” *American Society for Microbiology*, 24 Apr. 2013, https://www.researchgate.net/publication/236329043_The_VP8star_Domain_of_Neonatal_Rotavirus_Strain_G10P11_Binds_to_Type_II_Precursor_Glycans. Accessed 21 Mar. 2023.

Rambaut, Andrew. “FigTree.” *Molecular Evolution, Phylogenetics and Epidemiology*, 2007, http://tree.bio.ed.ac.uk/software/figtree/. Accessed 21 Mar. 2023.

Retel, Cas, et al. “Ecological and Evolutionary Processes Shaping Viral Genetic Diversity.” *Viruses*, vol. 11, no. 3, Mar. 2019, p. 220, https://doi.org/10.3390/v11030220.

Rodríguez, Javier M., et al. “New Insights into Rotavirus Entry Machinery: Stabilization of Rotavirus Spike Conformation Is Independent of Trypsin Cleavage.” *PLOS Pathogens*, vol. 10, no. 5, May 2014, https://doi.org/10.1371/journal.ppat.1004157.

Rubio, Luis, et al. “Genetic Variability and Evolutionary Dynamics of Viruses of the Family Closteroviridae.” *Frontiers in Microbiology,* vol. 4, no. 151, Jan. 2013, https://doi.org/10.3389/fmicb.2013.00151.

Sanjuán, Rafael, and Pilar Domingo-Calap. “Genetic Diversity and Evolution of Viral Populations.” *Encyclopedia of Virology*, Elsevier, 2021, pp. 53–61, http://dx.doi.org/10.1016/b978-0-12-809633-8.20958-8. Accessed 24 Mar. 2023.

Schloss, Patrick D. “Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology*, vol. 75, no. 23, Dec. 2009.

Seo, Neung-n Seo, et al. “INAUGURAL ARTICLE by a Recently Elected Academy Member: Integrins A1β1 and A2β1 Are Receptors for the Rotavirus Enterotoxin.” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 26, July 2008, https://doi.org/10.1073/pnas.0803934105.

Settembre, Ethan C., et al. “Atomic Model of an Infectious Rotavirus Particle.” *The EMBO Journal*, vol. 30, no. 2, 2011, pp. 408–16, https://doi.org/10.1038/emboj.2010.322.

Shen, Wei, et al. “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.” *PLOS ONE*, vol. 11, no. 10, Oct. 2016, https://doi.org/10.1371/journal.pone.0163962.

Silva-Sales, Marcelle, et al. “Genomic Constellation of Human Rotavirus A Strains Identified in Northern Brazil: A 6-Year Follow-up (2010-2016).” *Revista Do*

Instituto de Medicina Tropical de São Paulo, vol. 62, Jan. 2020, https://doi.org/10.1590/S1678-9946202062098.

Stencel-Baerenwald, Jennifer E., et al. "The Sweet Spot: Defining Virus-Sialic Acid Interactions." *Nature Reviews. Microbiology*, vol. 12, no. 11, Nov. 2014, pp. 739–49, https://doi.org/10.1038/nrmicro3346.

Sun, Xiaoman. "Glycan Binding Specificity and Mechanism of Human and Porcine P[6]/P[19] Rotavirus VP8*s." *Journal of Virology*, vol. 92, no. 14, May 2018.

---. "Structural Basis of Glycan Recognition of Rotavirus." *Frontiers in Molecular Biosciences*, vol. 8, no. 658029, July 2021, https://doi.org/10.3389/fmolb.2021.658029.

Suzuki, Hiroshi. "Rotavirus Replication: Gaps of Knowledge on Virus Entry and Morphogenesis." *The Tohoku Journal of Experimental Medicine*, vol. 248, no. 4, 2019, pp. 285–96, https://doi.org/10.1620/tjem.248.285.

Townsend, et al. "The Phylogenetic Informativeness of Nucleotide and Amino Acid Sequences for Reconstructing the Vertebrate Tree." *Journal of Molecular Evolution*, vol. 67, no. 5, Aug. 2008, pp. 437–47, https://doi.org/10.1007/s00239-008-9142-0.

Vetter, Janine, et al. "Rotavirus Spike Protein VP4 Mediates Viroplasm Assembly by Association to Actin Filaments." *Journal of Virology*, vol. 96, no. 17, Sept. 2022, https://doi.org/10.1128/jvi.01074-22.

World Health Organization. "Rotavirus." *World Health Organization*, 2023, https://www.who.int/teams/health-product-policy-and-standards/standards-and-specifications/vaccines-quality/rotavirus. Accessed 21 Mar. 2023.

Wylie, Kristine M., et al. “Emergence of Rotavirus G12P[8] in St. Louis During the 2012–2013 Rotavirus Season.” *Journal of the Pediatric Infectious Diseases Society*, vol. 4, no. 4, Sept. 2014, pp. e84–89, https://doi.org/10.1093/jpids/piu090.

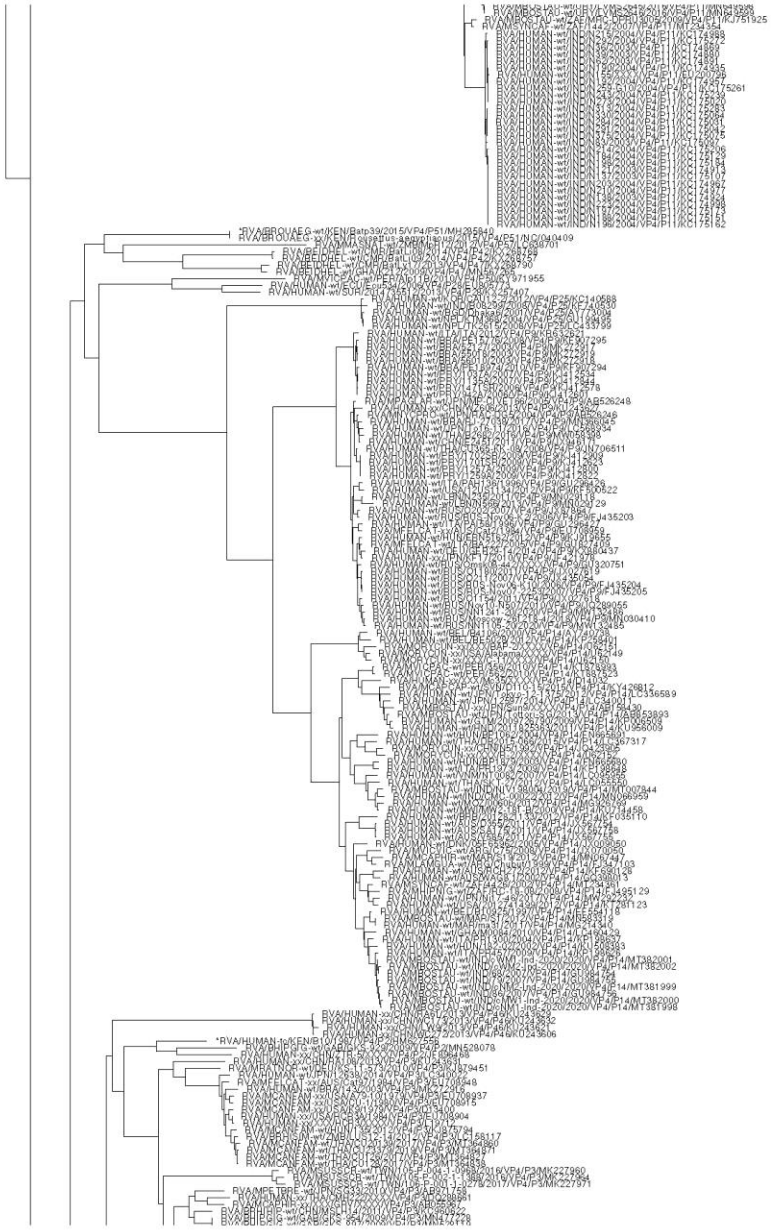
Zárate, Selene, et al. “Interaction of Rotaviruses with Hsc70 during Cell Entry Is Mediated by VP5.” *Journal of Virology*, vol. 77, no. 13, July 2003, https://doi.org/10.1128/JVI.77.13.7254-7260.2003.

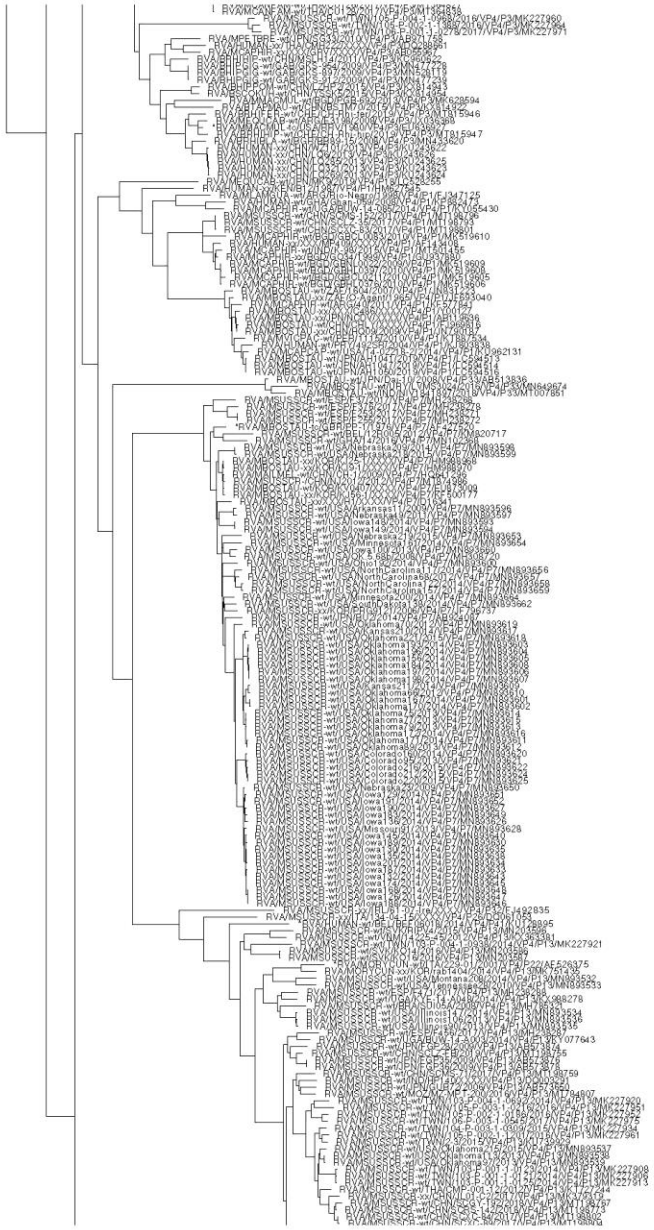
Zeller, Mark, et al. “Genetic Analyses Reveal Differences in the VP7 and VP4 Antigenic Epitopes between Human Rotaviruses Circulating in Belgium and Rotaviruses in Rotarix and RotaTeq.” *Journal of Clinical Microbiology*, vol. 50, no. 3, Mar. 2012, https://doi.org/10.1128/JCM.05590-11.

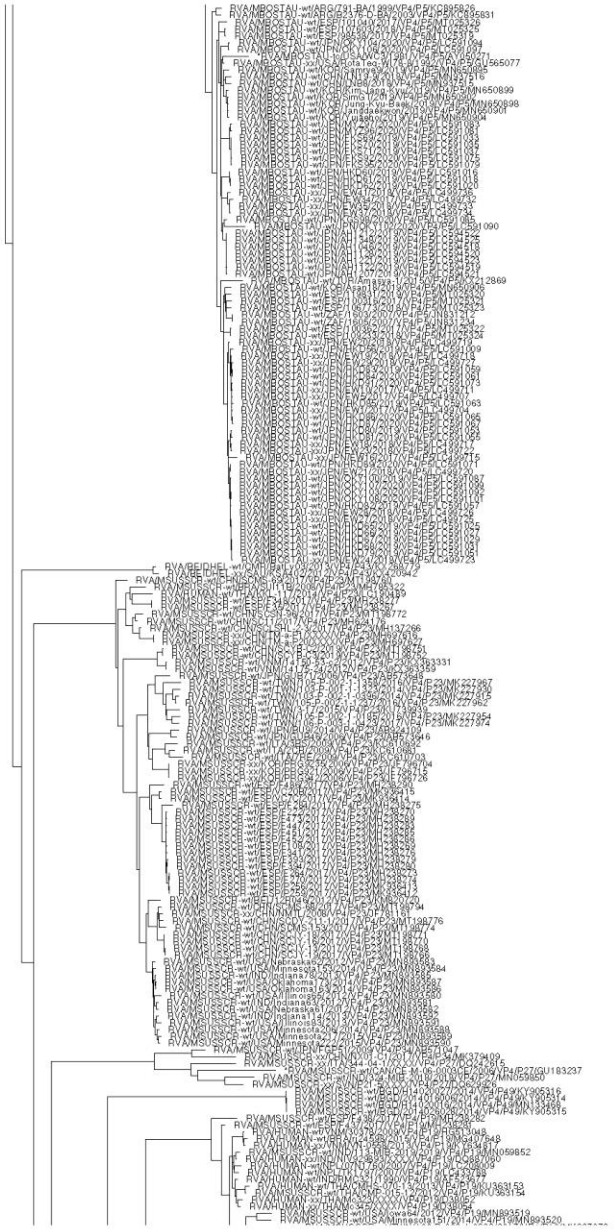
Zhang. “Rates of Conservative and Radical Nonsynonymous Nucleotide Substitutions in Mammalian Nuclear Genes.” *Journal of Molecular Evolution*, vol. 50, no. 1, Jan. 2000, pp. 56–68, https://doi.org/10.1007/s002399910007.

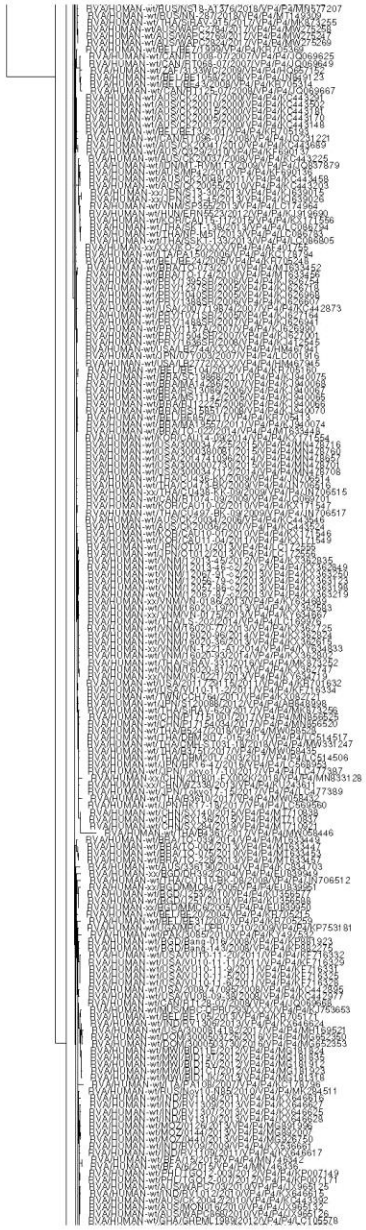
Zhao, Simiao, et al. “Global Infection Rate of Rotavirus C during 1980-2022 and Analysis of Critical Factors in the Host Range Restriction of Virus VP4.” *Viruses*, vol. 14, no. 12, Dec. 2022, p. 2826, https://doi.org/10.3390/v14122826.

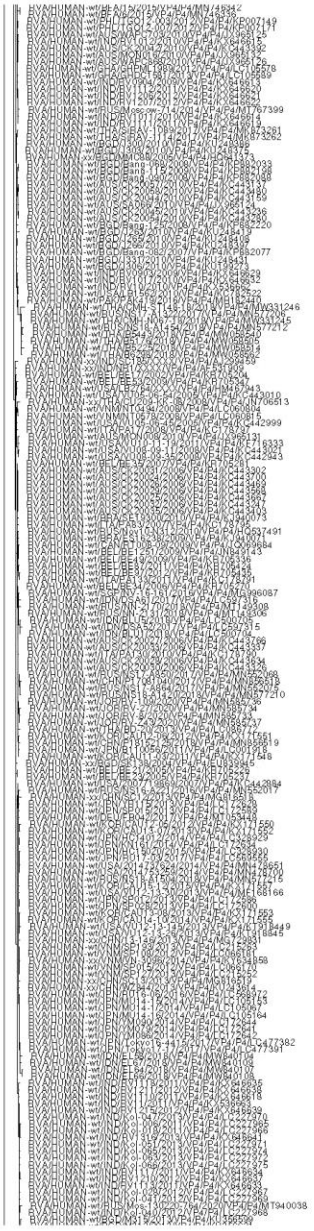
APPENDIX

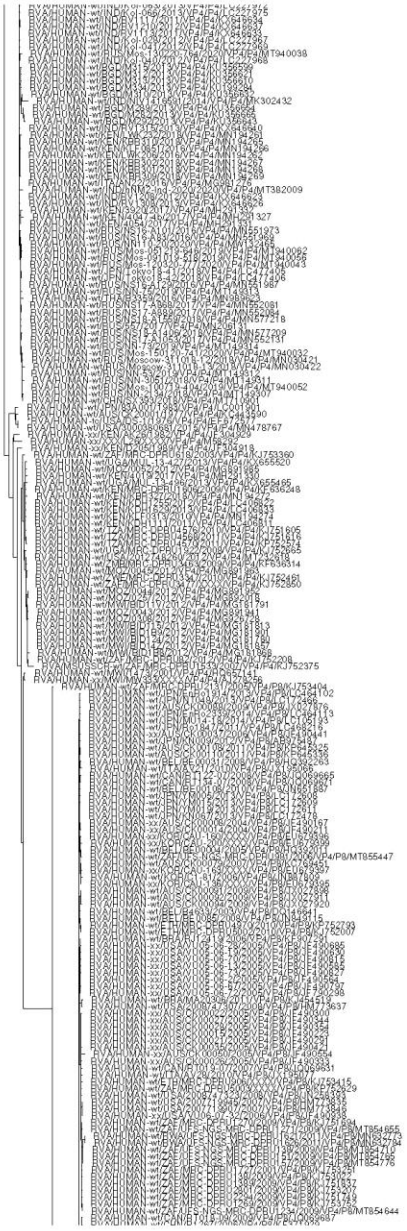


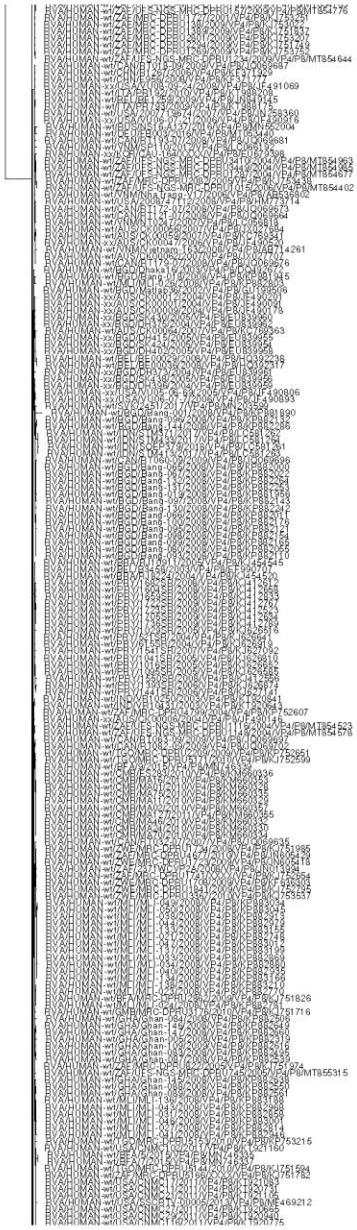


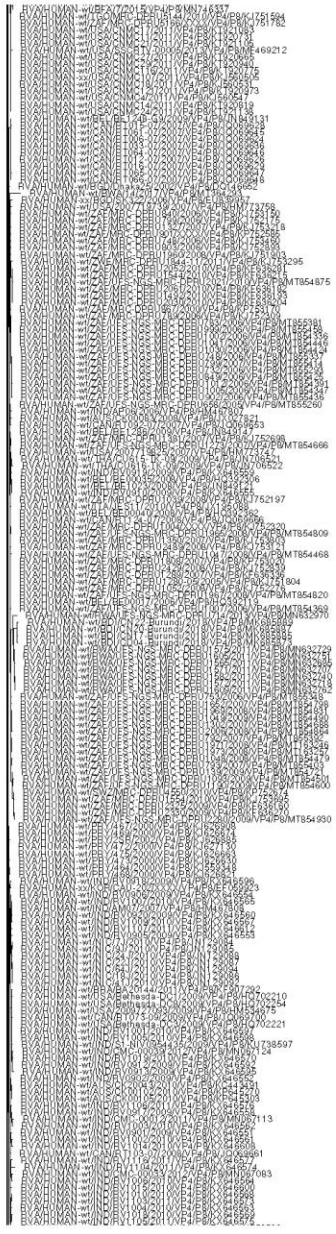


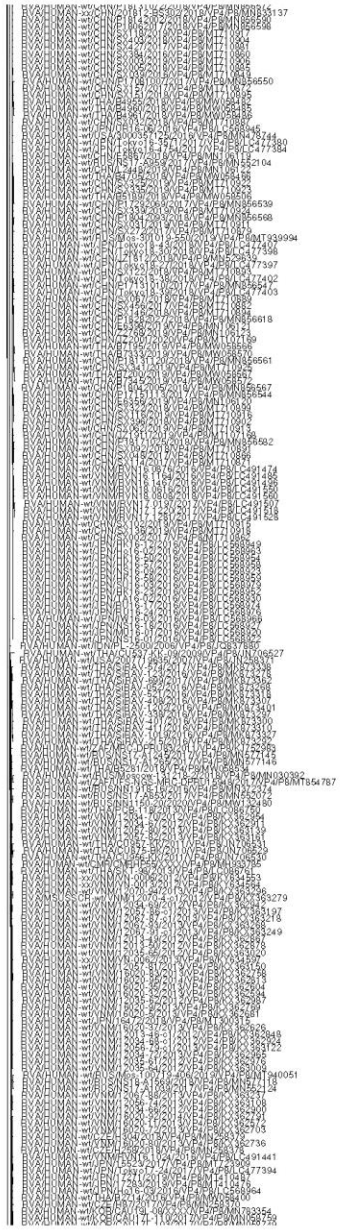


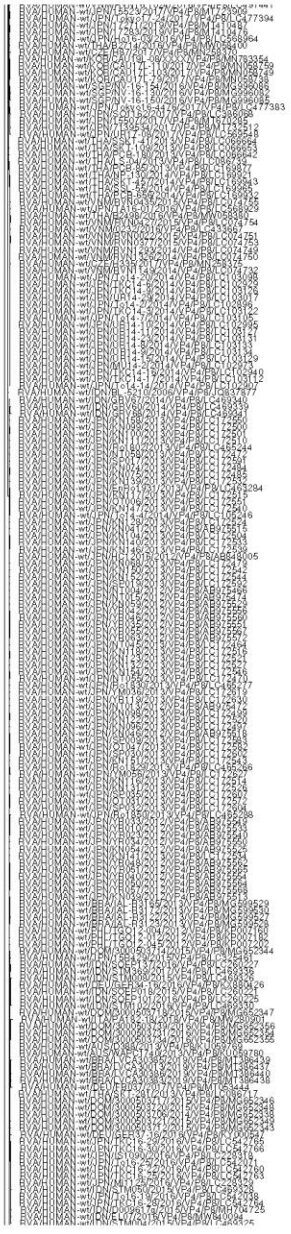


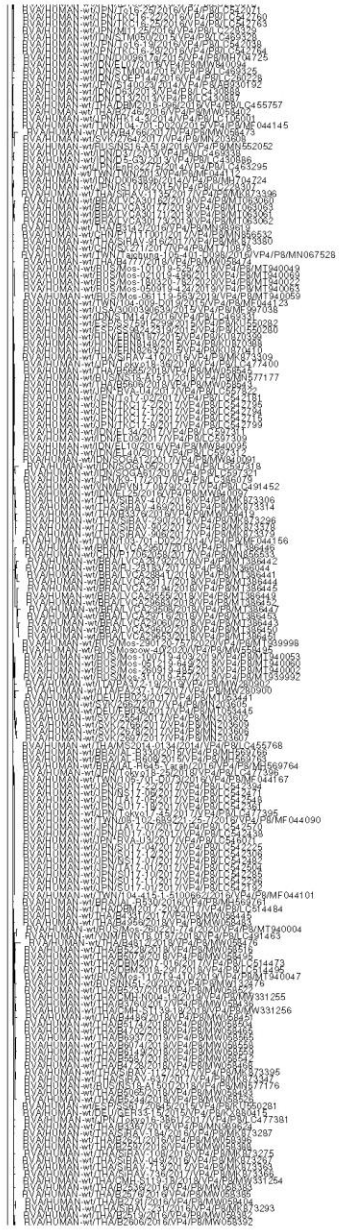


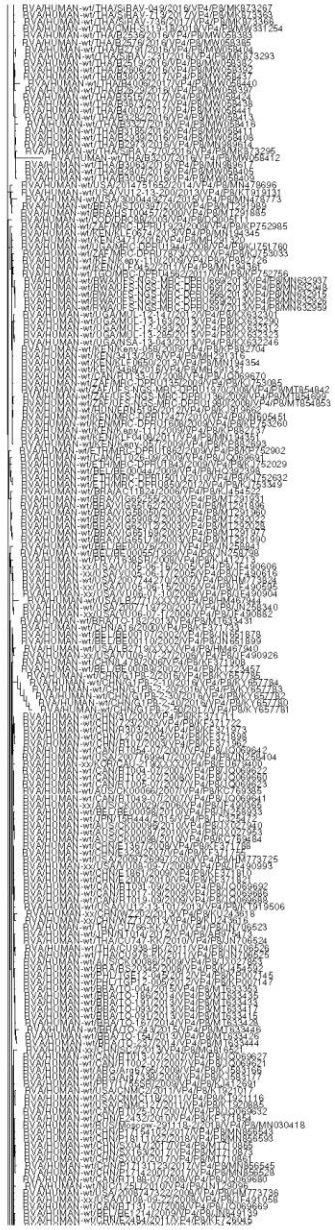


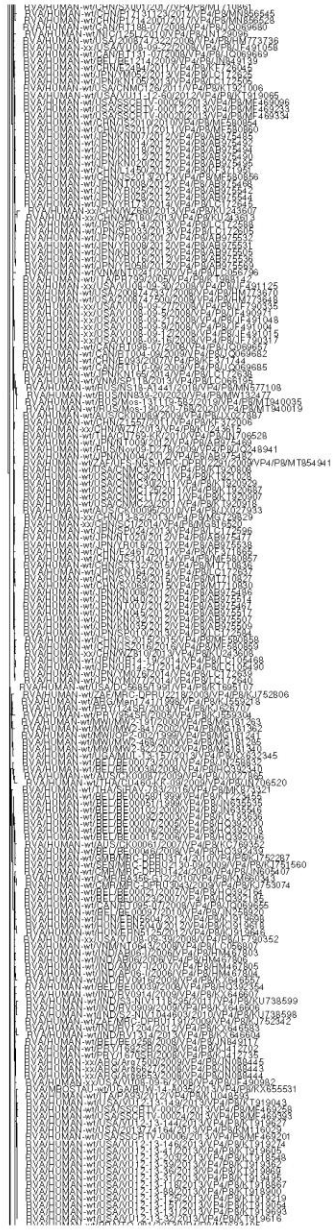


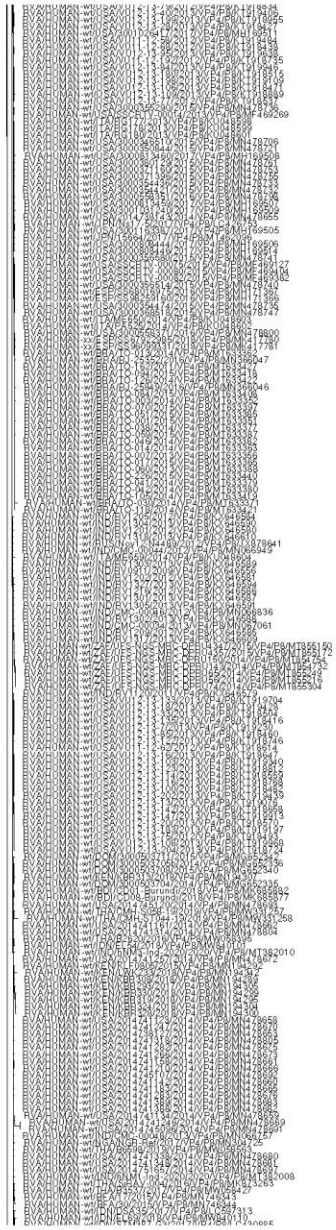


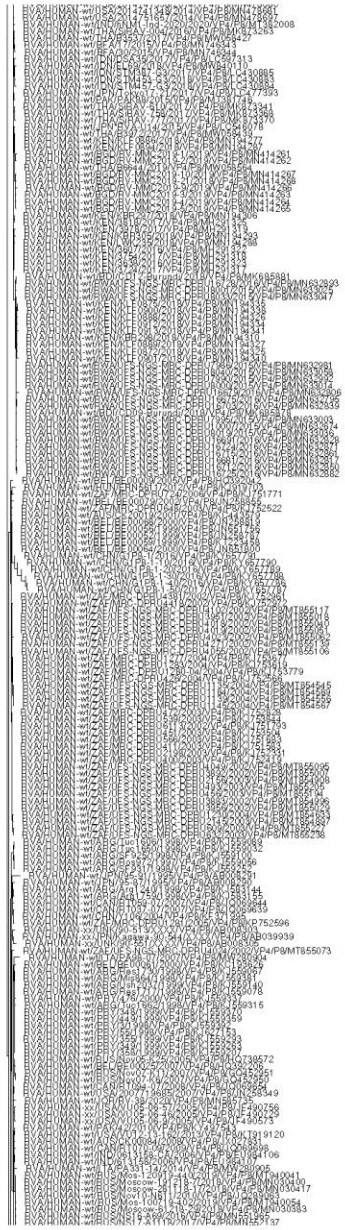


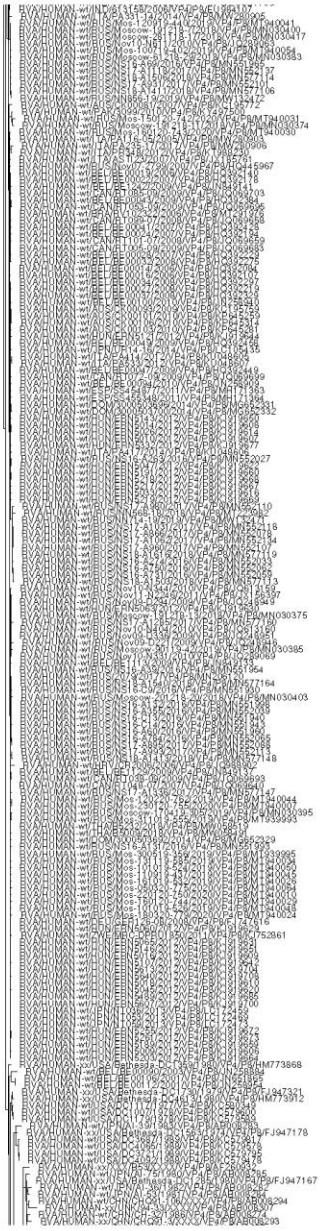


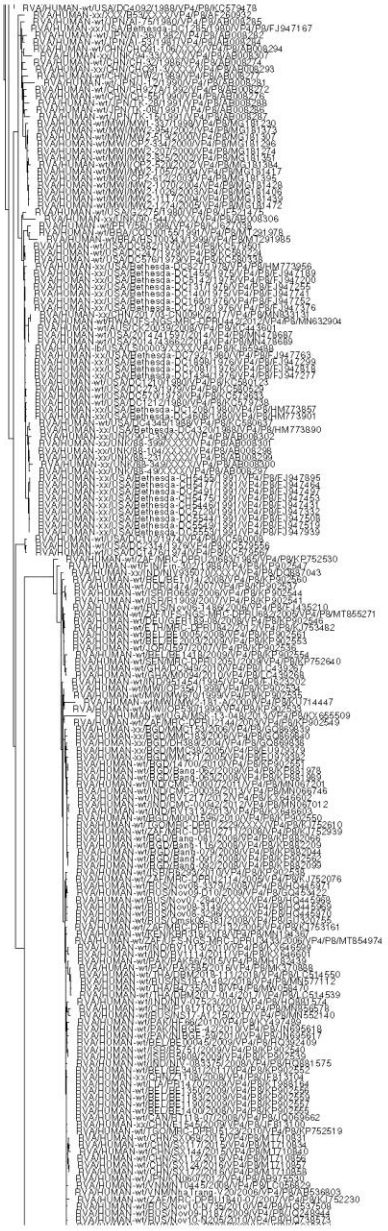












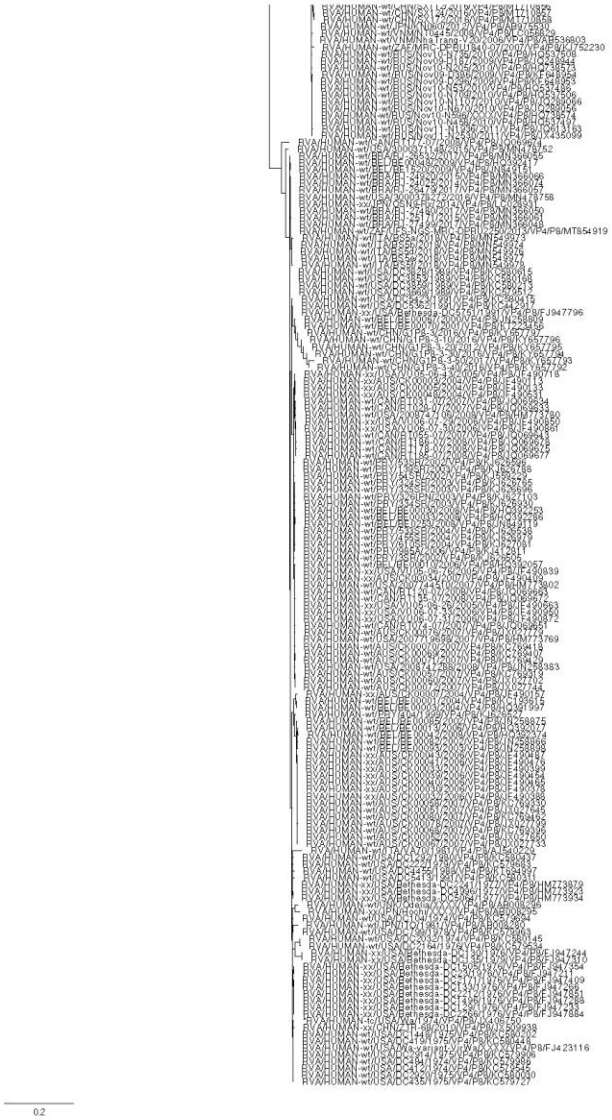
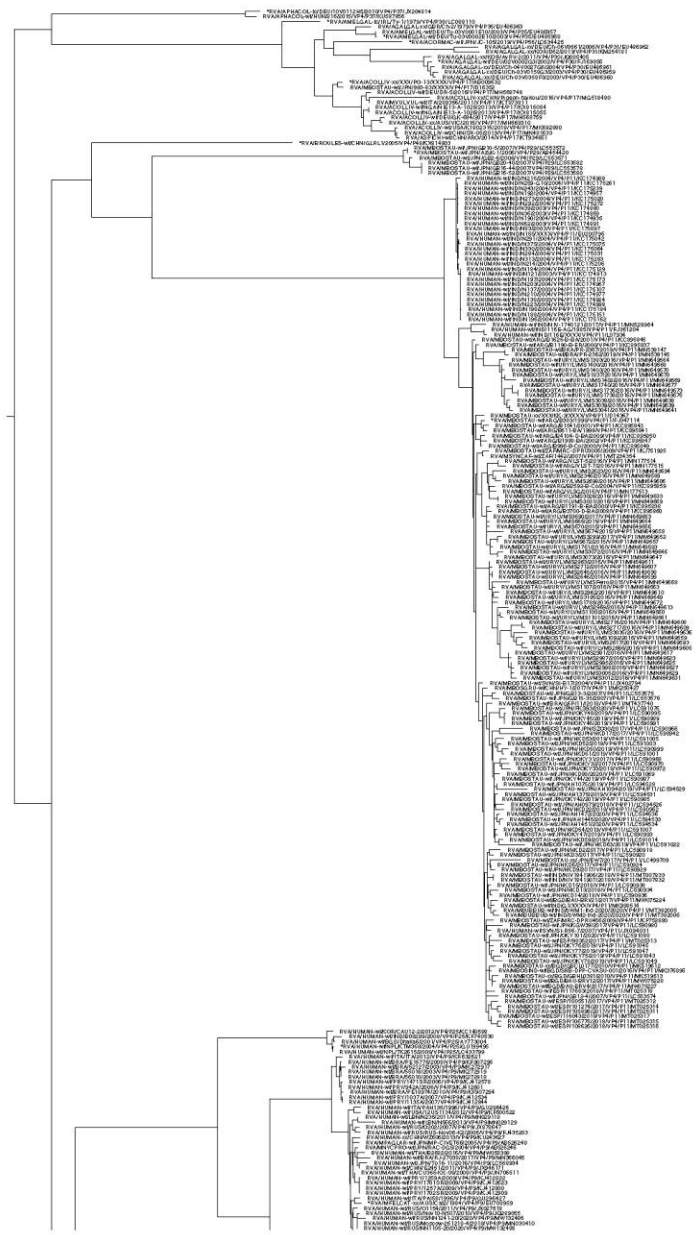
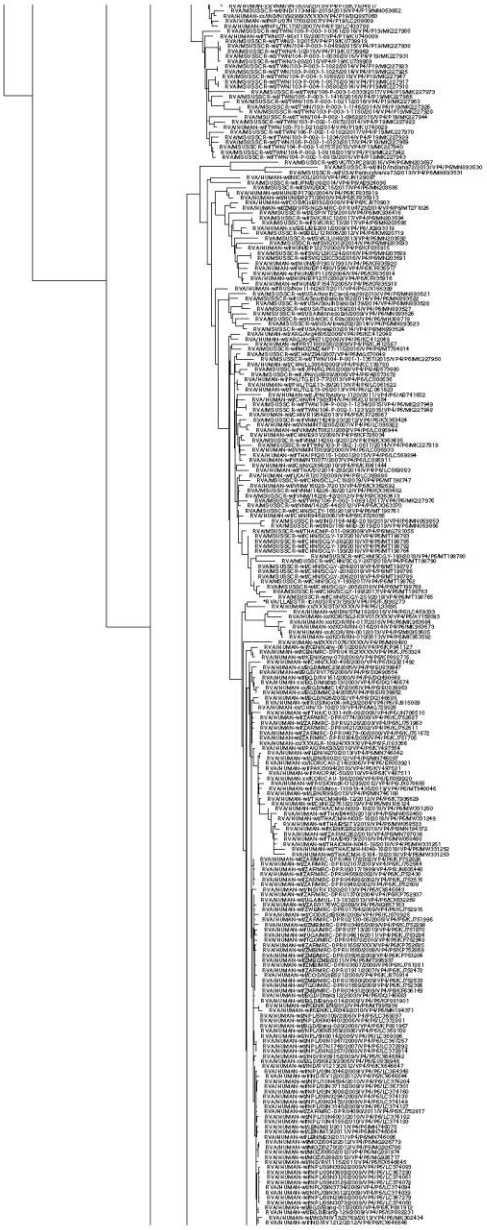
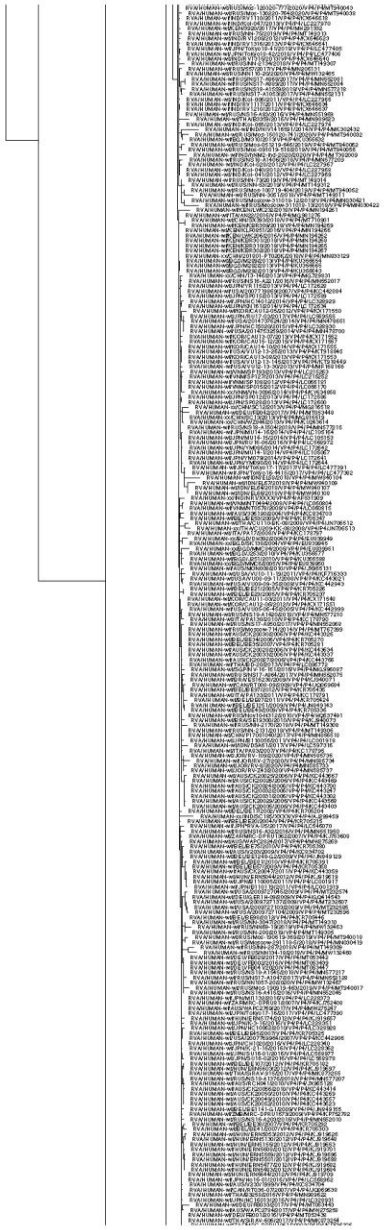


Figure S1: **VP4 Nucleotide Phylogenetic Tree with Sequence IDs.**







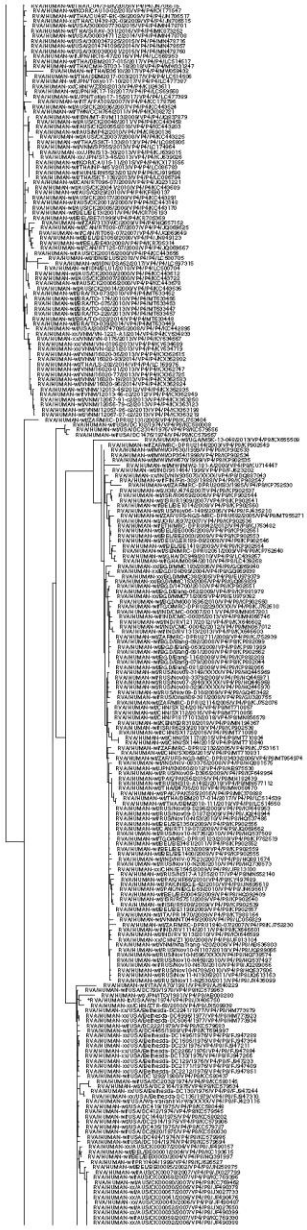


Figure S2: **VP4 Amino Acid Phylogenetic Tree with Sequence IDs.**

CURRICULUM VITAE

Lizel Mendoza

510-331-9200

mendl21@wfu.edu

EDUCATION

Wake Forest University, MS — Winston-Salem, NC, USA

Biology — August 2021 – Present

University of Arizona, BS — Tucson, AZ, USA

Natural Resources: Emphasis on Conservation Biology, Major — August 2015 – December 2016

Pima Community College, AS — Tucson, AZ, USA

Associates of Science: Anthropology — January 2014 - May 2015

RESEARCH EXPERIENCE

Wake Forest University, Department of Biology — Winston-Salem, NC, USA

Master's Student - Dr. James Pease Laboratory — August 2021 – Present

- Apply bioinformatic tools to model and align thousands of protein and nucleotide sequences.
- Incorporate Python scripts and terminal commands to analyze the genetic diversity of the strains.
- Participates in creating and managing a comprehensive internal rotavirus sequence metadatabase.
- Mentors undergraduates interested in virus evolution and computational-based research.
- Attends weekly laboratory meetings to discuss evolutionary and genomic-related research.

University of Arizona — Tucson, AZ, USA

Research Assistant — June 2016 – December 2016

- Wrote a research proposal and paper on, *Undergraduate Directed Research: Measuring the Genetic Diversity of a Small, Isolated Mexican Red-Bellied Squirrel Population*
- Excised tissue and extracted DNA from ear and skin samples taken from small, isolated Mexican red-bellied squirrel population in Florida
- Microsatellites (or STRs) are amplified using PCR in order to determine the amount of genetic diversity conserved from the original population in Mexico through the

small introduction and subsequent bottleneck
- Incorporated DNA quantification and fragment analysis to examine the purity of samples
- Analysis of heterozygosity levels by looking at the allele frequency distinguished by these STRs using GeneMapper.

TEACHING EXPERIENCE

Wake Forest University, Department of Biology — Winston-Salem, NC, USA

Teacher's Assistant — August 2021 – Present

- Create weekly lesson plans that reinforce lecture topics and build critical and applied thinking, observation, data analysis, and technical communication skills.
- Assign and evaluate assignments focused on reading, annotating, and discussing primary literature and experimental design and preparedness.
- Attend weekly TA meetings to prepare for upcoming labs and discuss student issues.
- Provide additional mentorship during office hours.

EMPLOYMENT

Roche Molecular Diagnostics — Pleasanton, CA, USA

Scientist I — November 2017 – June 2021

- Trained and certified to run, troubleshoot, and operate the cobas® 6800 and 8800 systems in association with the cobas® LIAT PCR platform.
- Assists in developing nucleic-based assays for commercial use on LIAT including Flu A and B (FABA), Strep A (SASA), Flu A, B and, RSV (FRTA) and, SARS-CoV-2 and Flu A/B (SCFA).
- Analyzes and interprets experimental data using internal software analysis tools, Excel, and JMP
- Produces protocols and other technical documents in compliance with SOP's and presents findings at internal meetings.
- Formulates and quality checks mastermixes, reagents and, buffers for various studies and experiments.

POSTERS/PRESENTATIONS

University of Arizona — Tucson, AZ, USA
OLLI Green Valley Fall 2016 Lecture Series — November 2016
Introduction to Conservation Genetics and Research on Small, Isolated Mexican Red-Bellied Squirrel Population

University of Arizona — Tucson, AZ, USA
Annual EEB Undergraduate Research Poster Session — April 2016
Contrasting Described and Projected Species Richness to Explain Diversification Patterns Among Animal Phyla

LABORATORY EXPERIENCE

The University of Arizona, Department of Geosciences Tucson, AZ, USA

Laboratory Assistant – Dr. Andrew Cohen's Laboratory August 2016 – December 2016

- Conducted phytolith extractions from sediment core samples drilled and shipped from Lake Tanganyika, Africa.
- Interpreted the variation and recorded key diagnostic characteristics between layers of these lake sediment samples that date back as far as 400 million years.
- Prepared and used various chemicals to create solutions and buffers.
- Attended weekly laboratory meetings to discuss recent and related papers to paleolimnology and paleoclimate and to stay up-to-date on the various research projects happening within the lab.

The University of Arizona, School of Natural Resources and the Environment. Tucson, AZ, USA

Research/Laboratory Assistant – Dr. John Koprowski's Lab. June 2016 – December 2016

- Excised tissue from deceased squirrels and prepared them for DNA extraction.
- Extracted DNA using various instruments of different squirrel species.
- Conducted PCR to amplify microsatellite loci using 24 unique primers.
- Supplemented gel electrophoresis to determine the effectiveness of each primer in each PCR sample.
- Prepared samples for DNA Fragment Analysis and Genomic DNA Quantification.
- Analyzed alleles at multiple loci to determine heterozygosity levels using GeneMapper.

University of Arizona, Arizona Research Labs Tucson, AZ, USA

Research/Laboratory Assistant – Dr. John Wien's Laboratory January 2016 – May 2016

- Used statistical software (R and Excel) to quantify organismal distributions among phyla.
- Applied phylogeographic techniques to analyze population genetic patterns across all known and living lizard species in order.
- Optimized DNA extraction protocols on grasshoppers.
- Prepared solutions, buffers, and reagents for molecular genetic research.

The University of Arizona, Department of Ecology and Evolutionary Biology Tucson, AZ, USA

Research/Laboratory Aide – Dr. Katrina Dlugosch's Lab February 2015 – December 2015

- Performed polymerase chain reactions (PCRs).
- Extracted DNA and administered purifications on hundreds of weevil samples.
- Quantified and measured DNA using a Qubit fluorometer and gel electrophoresis.
- Entered and managed data in Excel.
- Provided supplemental training to other students.

TECHNICAL SKILLS

Computational: Microsoft Office, Google Suite, GeneMapper, R, ArcGIS, JMP, Roche Software, Linux, Python, BLAST, Bash, Aliview, Seaview, earth system modeling using ESMF and related software.

Wet Laboratory: Nucleic acid extraction, quantification, and purification, gel electrophoresis, staining, dilution assays, cell culture, PCR, applying biochemical techniques to identify bacteria species.

Fieldwork: Assembling camera traps, mist-netting traps, radio telemetry, live-trapping small mammals, depicting different eye-shine to identify nocturnal animals, executing basic forest mensuration procedures, sight identifying various southwestern plant families and bird species.