UNIVERSITY OF GLASGOW

The Molecular and Genetic Evolution of Foot-and-Mouth Disease Virus

by

Grace Logan

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
College of Medical, Veterinary and Life Sciences
Institute of Biodiversity, Animal Health and Comparative Medicine

January 2017

Abstract

Foot-and-mouth disease virus (FMDV) (Family: *Picornaviridae*, Genus: *Aphthovirus*) is a significant global pathogen with extensive economic impact. FMDV has a low fidelity RNA-dependent RNA polymerase and lacks proof reading capability. This coupled with its relatively short generation time and large population sizes means it exists in a swarm of genetically closely related variants. The reservoir of diversity contained within this mutant spectrum allows the virus to adapt rapidly to new environments. Much of the previous work looking at virus evolution has focused on the consensus level genetic sequence. The advent of next generation sequencing (NGS) technologies enables evolutionary studies of the entire viral swarm. This PhD project uses NGS technologies to interrogate the swarm structure by investigating factors affecting the viral swarm and the dynamics of variants within it. Furthermore, this work shows how analysis of the swarm can reveal fundamental information about virus biology.

A PCR-free NGS methodology was developed to create deep sequencing data sets of all genomes present within an FMDV viral swarm. The elimination of the PCR step results in less errors being introduced in the sequencing process thereby improving the resolution and reliability of the identification of low level variants. This optimised method was then used to define and compare the FMDV swarms of several wildtype isolates. This revealed differences in swarm structure from isolate to isolate and produced evidence of within swarm selection. Not all proteins known to be under selection at the consensus level were also under selection within the swarm. The diversity of viruses within the swarm was found to be dependent upon the host from which a virus was sampled, with African buffalo potentially able to maintain multiple infections. Subconsensus variants in these mixed samples had mutations at positions previously associated with immune escape. Investigation of the evolution of swarm structure when adapting to new cell type *in vitro* indicated that two distinct population structures can exist relative to the existence of adaptive pressure. These two population structures have different distributions of variable nucleotides but comparative total levels of variation (as measured by Shannon's entropy). Deep sequencing of the virus swarm enabled the discovery of conserved novel stem loop structures, which were hypothesized to be required for packaging of the virus genome. Mutating these sites produced a virus with decreased packaging efficiency.

This thesis includes novel analysis techniques for considering the viral swarm. It demonstrates how investigating the diversity in the swarm can help us to understand virus molecular biology, its evolution and the limits upon this. Understanding viral evolution at this scale has the capacity to improve our fundamental understanding of the biology and evolution of FMDV which can in turn inform vaccine design and disease control strategies.

Contents

List of Tables

List of Figures

Declaration of Authorship

I, GRACE LOGAN, declare that this thesis titled, 'THE MOLECULAR AND GENETIC EVOLUTION OF FOOT AND MOUTH DISEASE VIRUS' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"An expert is a person who has made all the mistakes that can be made in a very narrow field."

Niels Bohr

Acknowledgements

Diolch Eleanor am ysgrifennu'r prosiect a credu fod dwi'n ddigon dda I wneud e ustus. Dwi'n gobeithio dwi 'di dechrau ateb rhai o'r cwestiynnau ti di 'ofyn ond os dwi heb at yr lleiaf roedd e'n daith diddorol! Toby, thank you for adopting me even when you had no input in the hiring process. Hopefully I wasn't too awful an addition (even if everything I did was not always awesome). Thanks Don for hiring someone who answered 'I'm afraid I don't know' to at least five of your questions in the interview. Thank you for being endlessly efficiently reading my work and pushing me to get my first paper published. Thanks to Dan for always making me feel welcome in Glasgow and for the discussions in the pub that lead me to reluctantly do some statistics. Thank you to Louise Matthews, Richard Orton, Paul Johnson and Richard Reeve for your input over the years.

Thank you to the picornavirus structure/molecular biology group members past and present for having me in your lab, for answering my questions, for sharing your reagents and leaving unicorn covered post-its on my desk on those days when that was necessary. Thank you Joe and Stephen for being human calculators and Sarah for being a font of useful knowledge. Thank you Terry for all of your endlessly practical advice and thanks to Nick for his 24hr bioinformatic facebook chat helpline. Thanks to Clare for help with everything LaTex. Thank you to the WRL for the provision of samples, advice and letting me use their equipment. Thanks in particular to Valerie, Britta, Jemma and Claudia for all of the time you have taken to help me, I truly appreciate it.

Jenny, Caroline and Katy: thank you not for keeping me sane, but for putting my insanity into perspective. Thank you to Ali and Stu for the golf breaks. Thanks to all those in cake club for making Thursdays awesome and for making me fat enough to need to attend fit club. Thanks to fit club for making exercise hilarious. Thanks to those who ran and visited the bar for many a great evening.

Thank you to my parents for the brains I inherited from you. For encouraging me to be a bit nerdy, for choosing good schools and taking me to science museums. For encouraging rational research even though my results sometimes seemed irrational. For being there with enthusiasm and support each long essay you had to proof read. Thanks also to my sister for doing so many degrees that only a PhD would out compete her. Immie and Maddy, thank you for reminding me there is life outside science! Aan my Ui, dankie dat jy so ondersteunend was die afgelope tyd. Dankie dat jy my rustig gehou het en my gehelp het om tyd af te vat. Dankie vir die bier en die blomme, dit het my aangemoedig toe ek dit die nodigste gehad het. Dankie vir die motivering.

Abbreviations

3'UTR	**Three prime Unt**ranslated **R**egion
5'UTR	**Five prime Unt**ranslated **R**egion
A	**A**denine
ATP	**A**denosine **T**ri **P**hosphate
BAM	**B**inary sequence **A**lignment **Ma**p
BFA	**B**ovine **F**oetal **A**orta
BHK	**B**aby **H**amster **K**idney
BRBV	**B**ovine **R**hinitus **B** **V**irus
BTY	**B**ovine **T**hyroid
C	**C**ytosine
$CaCl_2$	**Ca**lcium **Ch**loride
cDNA	**C**omplementary **D**eoxyribo**n**ucleic **A**cid
CCS	**C**ircular **C**onsensus **S**equence
cre	**C**is-acting **R**eplication **E**lement
CSU	**C**entral **S**ervices **U**nit
DEFRA	**D**epartment for **E**nvironment, **F**ood and **R**ural **A**ffairs
DNA	**D**eoxyribo**n**ucleic **A**cid
DPC	**D**ays **P**ost **C**ontact
dsDNA	**D**ouble **S**tranded **D**eoxyribo**n**ucleic **A**cid
DWV	**D**eformed **W**ing **V**irus
EDTA	**E**thylene**d**iaminete**t**raacetic **A**cid
eIF4G	**E**ukaryotic translation **I**nitiation **F**actor **4G**
ELISA	**E**nzyme-linked **i**mmuno**s**orbent **a**ssay.
EMCV	**E**ncephalo**m**yo**c**arditis **V**irus

ERAV	**E**quine **R**hinitus **A** **V**irus
EtBr	**Et**hidium **Br**omide
EU	**E**uropean **U**nion
FMD	**F**oot-and-**m**outh **d**isease
FMDV	**F**oot-and-**m**outh **d**isease **v**irus
FRF	**F**ront **R**ight **F**oot
G	**G**uanine
gDNA	**G**enomic **D**eoxyribo**n**ucleic **A**cid
HCV	**H**epatitis **C** **V**irus
HIV	**H**uman **I**mmunodeficiency **V**irus
HPI	**H**ours **P**ost **I**nfection
ICP	**I**nfectious **C**opy **P**lasmid
IF	**I**mmuno**f**luorescence
IRES	**I**nternal **r**ibosome **e**ntry **s**ite
KCl	Potassium Chloride
KNP	**K**ruger **N**ational **P**ark
$\mathbf{L}^{pro}$	**L**eader **p**rotease
M	**M**olar
$\mathbf{MgSO_4}$	Magnesium Sulphate
MHC	**M**ajor **H**istocompatability **C**omplex
MLD	**M**aximum **L**adder **D**istance
MOI	**M**ultiplicity **o**f **I**nfection
NaCl	Sodium Chloride
$\mathbf{NaH_2PO}$	Sodium Hydrogen Phosphate
$\mathbf{nfH_2O}$	**N**uclease **F**ree Water (H_2O)
NGS	**N**ext **G**eneration **S**equencing
OIE	**O**ffice **I**nternational des **É**pizooties
ORF	**O**pen **r**eading **f**rame
PB	**P**ro**b**ang
PBS	**P**hosphate **B**uffered **S**aline
PBSa	**P**hosphate **B**uffered **S**aline (calcium and magnesium free)
PBS-BSA	**P**hosphate **B**uffered **S**aline

	containing **B**ovine **S**erum **A**lbumin
PCI	**P**henol **C**hloroform **I**soamyl alcohol
PCR	**P**olymerase **C**hain **R**eaction
PFU	**P**laque **F**orming **U**nits
PPS	**P**utative **P**ackaging **S**ignal
PV	**P**olio**v**irus
qPCR	**Q**uantitative **P**olymerase **C**hain **R**eaction
RdRP	**R**NA **d**ependant **R**NA **p**olymerase
RGD	Arginine-Glycine-Aspartic Acid
RNA	**R**ibo**n**ucleic **A**cid
rRNA	**R**ibosomal **R**ibo**n**ucleic **A**cid
RT	**R**everse **T**ranscription
RT-PCR	**R**everse **T**ranscription with **P**olymerase **C**hain **R**eaction
SAM	**S**equence **A**lignment **M**ap
SAT	**S**outhern **A**frican **T**erritories
SBL	**S**equencing **b**y **L**igation
SBS	**S**equencing **b**y **S**ynthesis
S-fragment	**S**hort **fragment**
SL1	**S**tem **L**oop one
SL2	**S**tem **L**oop two
SMRT	**S**ingle **M**olecule **R**eal **T**ime
SNP	**S**ingle **N**ucleotide **P**olymorphism
T	**T**hymine
TB	**T**uberculosis
TCR	**T** **c**ell **r**eceptor
TGAC	**T**he **G**enome **A**nalysis **C**entre
TPI	**T**he **P**irbright **I**nstitute
TPB	**T**ryptose **P**hosphate **B**roth
TRIS-HCL	**TRIS**(hydroxymethyl)aminomethane **H**ydro**ch**l**o**ride
U	**U**racil
USD	**U**nited **S**tates **D**ollar

VESV	**V**esicular **e**xanthema of **S**wine **V**irus
VGM	**V**iral **G**rowth **M**edia
VP	**V**iral **P**rotein
VPg	**V**iral **P**rotein **G**enome-linked
VSV	**V**esicular **S**tomatitis **V**irus
WGS	**W**hole **G**enome **S**equence
WRLFMD	**W**orld **R**eference **L**aboratory for **f**oot-and-**m**outh **D**isease
ZMW	**Z**ero **M**ode **W**avelength

Chapter 1

Introduction

This thesis investigates the genetic and molecular evolution of foot-and-mouth disease virus (FMDV). This introductory chapter describes what is known about FMDV and the disease it causes; foot-and-mouth disease (FMD). A summary of the current global distribution and predicted financial burden is included followed by a description of the virus, its life cycle and the clinical signs it causes. A description of transmission, current diagnostic techniques and vaccine design is also made. The latter section of this chapter discusses viral evolution. This describes how viruses exists in a diverse swarm of genetically related variants. The creation of this diversity is discussed along with the concepts of the sequence space the swarm exists within and effect of the mutant spectrum on fitness. The current molecular techniques available to consider these concepts are summarised. Finally the objectives of this thesis are outlined.

1.1 Foot-and-mouth disease

Foot-and-mouth disease virus is a global pathogen with an extensive economic impact. The virus is a member of the family *Picornaviridae* within the genus *Aphthovirus*. There are seven serotypes of FMDV. Serotype O and A were the first to be assigned in 1922 by Vallée and Carré. They were named after their country of origin. Vallée O was named after the northern French region of Oise and is now known as type O. Vallée A was named after Allemagne (Germany) and is now referred to as type A [1]. Simultaneously Waldman and Trautwein identified three serotypes; Waldman A, B and C. Waldman A and B were then found to be the same serotypes as Vallée O and Vallée A and the French nomenclature was retained. Waldman type C was novel and is now known as type C [2]. In the 1950s, subsequent to the classification of the European serotypes, Asia

1 was isolated in Pakistan and India [3, 4] and three distinct serotypes were isolated in southern Africa (southern African territories 1-3)(SAT1-3) [5].

1.1.0.1 Global distribution

FMDV is not evenly distributed throughout the globe. This is affected by a number of factors including (but not limited to);

- livestock density and the presence of reservoir hosts
- animal husbandry, movement and trade
- investment in disease control

The global FMDV population can be roughly divided into seven regional pools. Pool 1 covers south-east Asia with spillover into eastern Asia. Pool 2 represents southern Asia. Pool 3 covers EurAsia (including the Middle East). In these three regions type O, type A and Asia 1 are circulating. Pools 4, 5 and 6 cover eastern, western and southern Africa respectively. In pool 4 type O, type A and SAT 1, 2 and 3 are circulating. In pool 5 type O and A circulate with SAT 1 and 2 and in pool 6, southern Africa, only the SATs tend to circulate. Pool seven covers South America and has only type A and type O circulating.

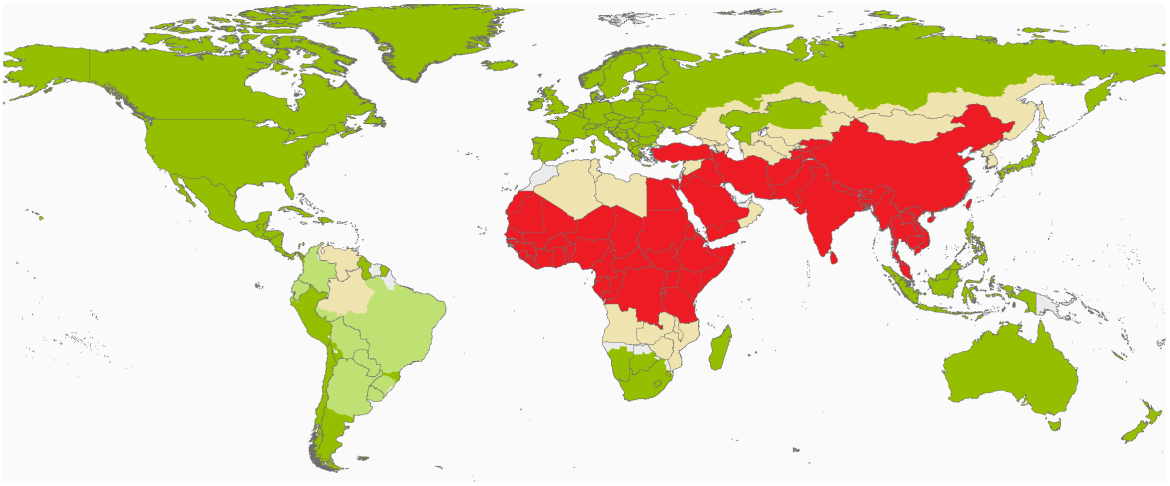


FIGURE 1.1: **Conjectured FMD country status** Countries marked in dark green are OIE member countries and zones recognised as free from FMD without vaccination. Countries marked in light green are OIE member countries or zones recognised as free from FMD with vaccination. All other countries are without an OIE official status for FMD (accurate in Aug 2016)[6]. The conjectured status of FMD in countries without an OIE official status (as compiled by the WRLFMD, Pirbright) is shown in yellow for countries with sporadic outbreaks and in red for those countries/regions predicted to have endemic infection.

A, O and C are the most globally distributed [7] although C has not been isolated since 2004 and may no longer be circulating [8]. As can be seen from the pools, the southern African serotypes mostly circulate in sub-saharan Africa [9] and Asia 1 outbreaks are usually restricted to Asia [10] although both SATs and Asia 1 have made incursions outside these limits.

1.1.0.2 Economic cost

FMDV outbreaks can have an extensive economic impact in both FMD free countries and those endemically infected. Conjectured FMD country status is outlined in Figure 1.1.

Outbreaks in FMD free countries An outbreak in an FMD-free country causes the loss of stock and the loss of many trade rights. Restoration of trade rights requires regaining FMD free status from the World Organisation for Animal Health (OIE). Recommendations set by OIE policy currently suggest quarantine and slaughter in response to an FMD outbreak in an FMD free country. This involves 'stamping out' of all animals infected and those in contact or at risk of infection (i.e. herds at neighbouring premises). This results in a large loss of life amongst susceptible livestock. In the 2001 outbreak in the United Kingdom 2030 infected premises were identified across the country which resulted in the culling of 6 million animals. Only 1.3 million of these were on infected premisses [11]. There is increasing public pressure to adopt policies less costly to livestock lives than 'stamping out'. For bovine tuberculosis (TB) the Department for Environment Food and Rural Affairs (DEFRA) employs a policy of removal of animals only if they have tested positive for the disease [12]. This reduces the impact of loss of life improving animal welfare and decreasing the financial burden of outbreaks but would struggle to be effective with FMDV due to its highly contagious nature. Others have suggested the use of emergency vaccination and the Council of the European Union (EU) has issued a directive which includes the provision and use of vaccination [13, 14]. The loss of stock and the inability to trade has an extensive economic impact. The direct impact of the UK 2001 FMDV outbreak was estimated at 8 billion pounds [15].

Endemic Countries The economic impact of an outbreak in FMD free countries can be clearly documented as described above. However, the continued burden of infection in endemic countries is more challenging to quantify. Rushton *et al* have proposed a framework for assessing the impact of FMD outlined in Fig. 1.2 [16]. Although the mortality of FMD is limited the disease has numerous more constant affects. Knight and Rushton highlight that as well as the immediate loss of production, the long term

productivity can be affected by changes in the structure of the herd due to lower fertility. The cost of managing disease through diagnostics, vaccination and movement control is high but without those controls access to the international markets cannot be achieved. Vaccination and production losses caused by FMD are estimated to cost 5 billion USD per annum globally [17].

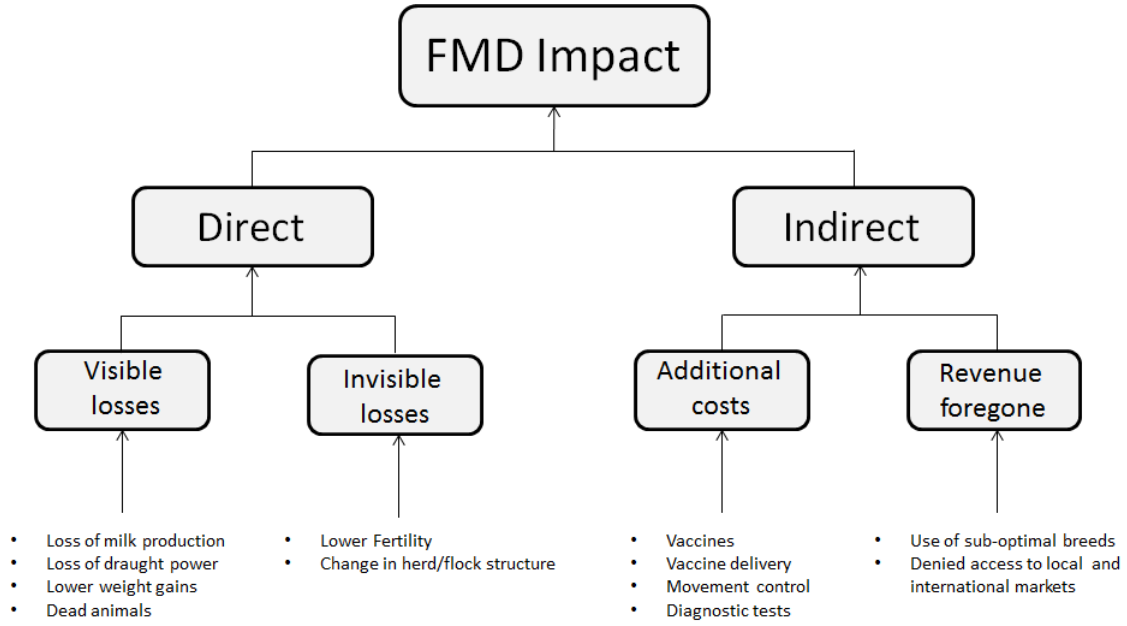


FIGURE 1.2: **Impact of FMD** The Impact of FMD in endemic countries can be broadly split into direct losses (both visible and invisible) and indirect loses such as the cost associated with disease control and revenue that cannot be achieved due to trade restrictions. Figure taken from Knight-Jones *et al* [17]

1.1.1 The disease

Hosts FMD is a disease of cloven hooved animals. It has an extensive host range with the majority of even-toed ungulates (order *Artiodactyla*) able to contract the disease [18]. Species that have been identified with FMD are outlined in Table 1.1.

Order	Family	Number
Artiodactyla	*Bovidae*	24
	Camelidae	4
	Cervidae	10
	Giraffidae	1
	Suidae	4
	Tayasuidae	1
Hydracoidea	*Procaviidae*	1

Insectivora	*Erinaceidae*	2
	Talpidae	1
Lagomorpha	*Leporidae*	1
Probiscidae	*Elephantidae*	2
Rodentia	*Bathyergidae*	2
	Capromyidae	1
	Chinchillidae	1
	Dasyproctidae	1
	Hydrochaeridae	1
	Hystricidae	1
	Muridae	4
	Sciuridae	2

TABLE 1.1: **Species susceptible to FMDV** A list of families within orders able to contract FMDV. Number represents the number of species within that family identified as susceptible [19].

Carrier Animals An animal is considered a persistent carrier if live virus is present in an asymptomatic animal over 28 days post recovery from acute infection [20]. Some animals do not suffer acute infection but are still able to carry the disease. FMD carrier animals have been identified in sheep, buffalo, goats and cattle. Virus recovered from these animals is generally low titre [21] and has been found to persist for up to three years [22].

Site of infection In cattle the tissue that serves as the primary site of replication is contested. Some studies have found it to be the pharynx [23] while other studies have identified bronchiolar epithelium [24]. These differences may be due to the specific infecting virus. Different cell types have a different range and number of receptors on the cell surface which is the distinguishing factor in tropism. Virus particles disseminate from the primary site of replication to oral and podal epithelium potentially via monocytes or macrophages [25]. In pigs, virus is predicted to replicate in the pharynx [26]. From here it passes through local lymph nodes and enters the blood stream to disseminate to mostly podal epithelial tissues. This is where lesions in pigs most commonly occur [27].

Clinical Signs FMD generally causes low mortality although this can be elevated in older, younger or immunocomprimised animals [28]. As well as age and immune status, severity of disease can be affected by the host species or breed. For example, FMD infection in African buffalo produces no clinical signs [29].

Clinical signs of FMD (if apparent), appear from 7-10 days post infection and include fever and vesicular lesions on tongue, snout, teats and feet leading to lameness [30]. These symptoms can result in refusal to feed causing loss of body condition and drop in milk production.

Transmission The most common form of transmission is through direct contact between infected and susceptible animals. Transmission can also occur through indirect contact such as contact between a susceptible animal and infected feed products or equipment [31]. This is due to the presence of virus in saliva, expelled air, milk, urine and faeces of acutely infected animals [32]. In cattle the respiratory tract is generally considered to be the infection route although infection has also been recorded through skin abrasions and mucous membranes [33]. Pigs are much less suceptible to infection via aerosolized virus [34]. In pigs infection is mostly caused by direct contact or ingestion of infected feed.

Diagnostics The clinical signs of FMDV bare some similarity to other animal disease (such as vesicular stomatitis in cattle and pigs). Therefore, clinical diagnosis by veterinarians in the field is supplemented by laboratory diagnosis. Samples from the field can be sent to country reference laboratories or the World Reference Laboratory for FMD (WRLFMD) based at The Pirbright Institute.

The OIE gold standard for diagnosis of FMD is isolation of the virus from primary bovine cell culture (Bovine Thyroid Cells (BTys)) [35]. This is supplemented with viral antigen enzyme-linked immunosorbent assay (ELISA) and real time polymerase chain reaction (RT-PCR) detection [35] which allow for virus serotyping. Sanger sequencing of a section of the capsid coding region of the genome is also performed to provide data for molecular epedemiological studies and allow for more informed vaccine selection [36]. Vaccine selection is further supplemented by vaccine matching, a process that involves virus neutralisation tests of animal serum with a panel of vaccine candidates [37]. Some pen-side tests are also available including field PCR machines (and isothermal amplification processes) and lateral flow devices [38, 39]

Vaccine Currently almost all available vaccines are inactivated virus preparations [40] although some peptide based vaccines are in use in China. There are vaccines designed against single strains and multivalent vaccines designed to protect against multiple strains but no single vaccine exists that protects against all seven serotypes of FMDV. The majority of FMDV vaccines are prepared by growth in cell culture, clarification, inactivation and blending with appropriate adjuvants [41]. Extensive research

continues into the manufacture of a multivalent vaccine appropriate for all serotypes and strains of FMDV. Recently published work has focused on the stabilisation of an empty FMDV capsid [42, 43]. This technique involves the recombinant expression of FMDV capsids. This would negate the need to grown up large quantities of virus for subsequent binary ethyleneimine inactivation [44] and allow the production of vaccine outside of high containment facilities.

1.1.2 Foot-and-mouth disease virus

1.1.2.1 Classification

As previously described FMDV is a *Aphthovirus* within the family *Picornaviridae*. This is a diverse virus family with thirty one different genera further divided into 54 species [45] (Fig. 1.3). There are also currently over 40 unclassified viruses within this family.

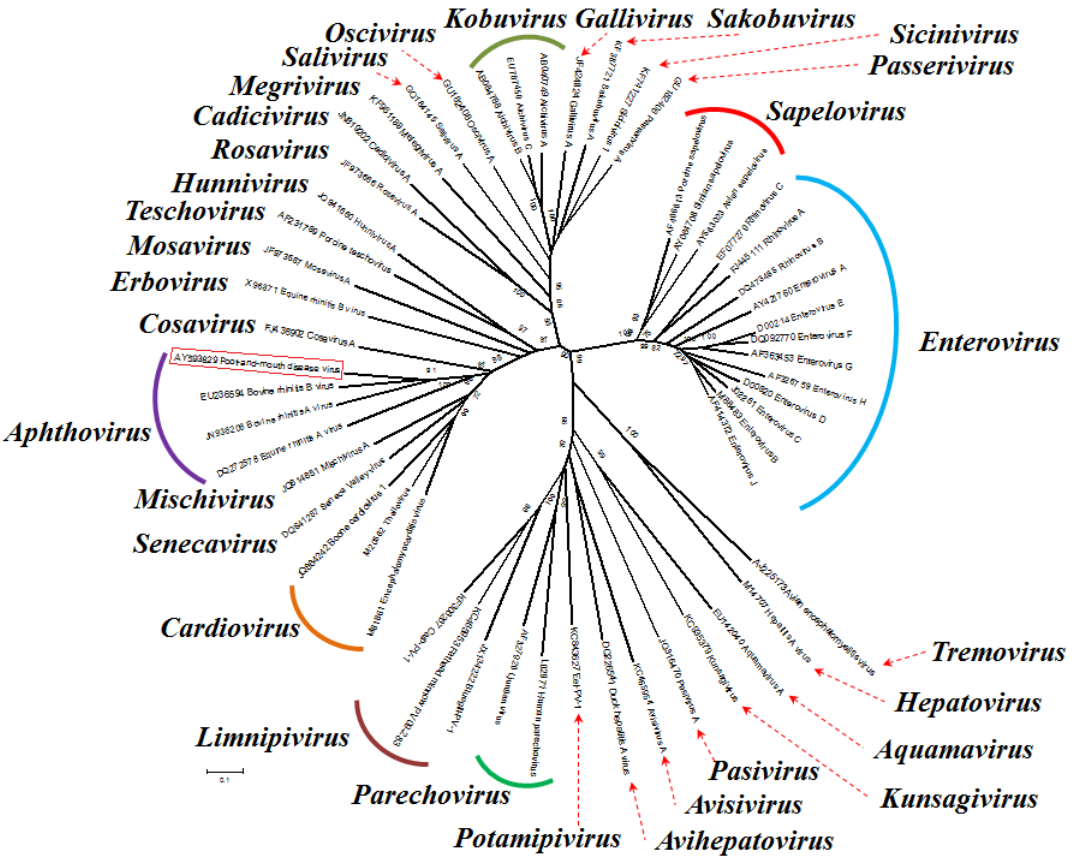


FIGURE 1.3: **Phylogenetic tree of the *Picornaviridae* family** An unrooted neighbour joining tree of 31 genera of the family *Picornaviridae* completed using the sequences of 3D polymerase by Nick Knowles [45]. FMDV is highlighted in red

The genome organisation of the genera of *Picornaviridae* is similar. Further details of each of the proteins within the genome is described in the subsequent section. These similarities have allowed for work completed in some model organisms to be consider potentially indicative of viral function in other viruses within the family (Fig. 1.4).

The most studied virus in this family is poliovirus (PV). Poliovirus is a human enterovirus and is the causative agent of poliomyelitis. There are three serotypes of PV; poliovirus type 1 (PV1), poliovirus type 2 (PV2) and poliovirus type 3 (PV3). Official global eradication of poliovirus type 2 was declared in September 2015. PV3 was last identified in 2012 in Nigeria and it is hoped this serotype has also been eradicated. Poliovirus type 1 is therefore the most highly circulating virus although this strain is currently thought to be restricted to Pakistan and Afghanistan [46]. Due to the extensive cost on human life (or quality of life) a large amount of funding has been provided to research effective vaccines and study this virus. Much of what is known about poliovirus' life cycle is predicted to be similar in other *Picornaviridae* and it is sometimes used as a proxy for FMDV when comparable experiments have not been completed.

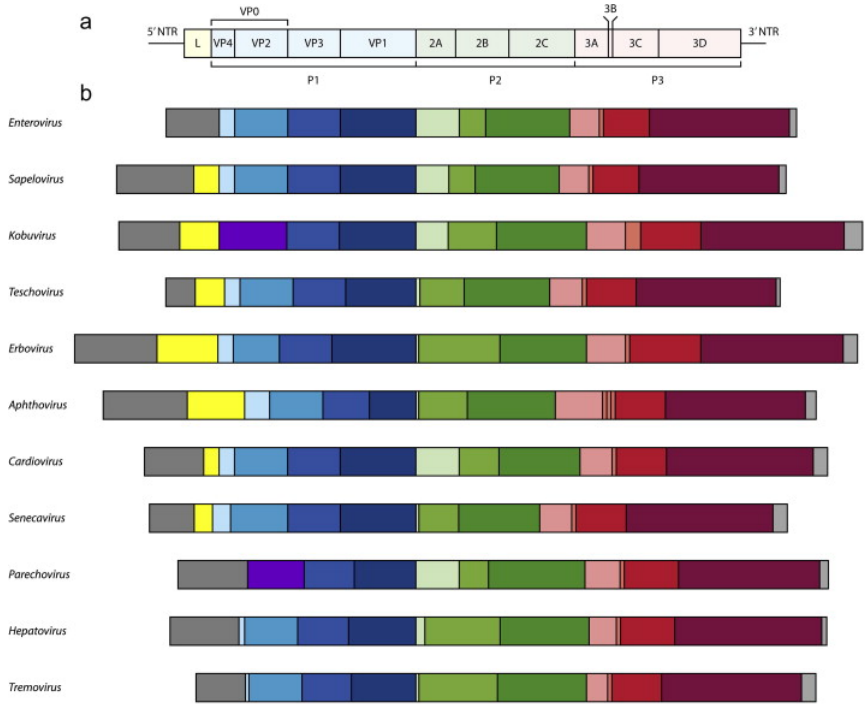


FIGURE 1.4: ***Picornaviridae* genome organisation** Each genome contains a 5 prime untranslated region followed by a Leader protease which can be produced in two isoforms in some family members (Grey/Yellow). The structural capsid proteins follow (shades of blue) with some virus members (Kobuvirus and Parechovirus) lacking a cleavage event completed by the other viruses and thus having only three capsid proteins one of which is a precursor in other viruses (VP0-purple) non structural proteins coded by P2 (green) and P3 (red) are similar although FMDV codes for three copies of 3B compared to 1 in each of the other virses. Figure taken from Lewis-Rogers *et al* [47]

1.1.2.2 Virion

The FMDV virion is approximately 30nm in diameter. The virus produces four capsid proteins (VP4, VP2, VP3 and VP1) coded for by the P1 section of the polyprotein. Proteolytic processing of P1 produces VP1, VP3 and the precursor VP0. A single copy of each of these assembles to form a protomer structure. Five protomeric subunits assemble to form a pentamer and 12 pentamers assemble to form a capsid. A maturation event occurs in which VP0 undergoes RNA dependant autocleavage to form VP2 and VP4 stabilising the capsid [48]. FMDV forms an icosahedral capsid that exhibits 2-, 3- and 5- fold axes of symmetry [49]. The capsid demonstrates pseudo T=3 architecture [50](Fig. 1.5).

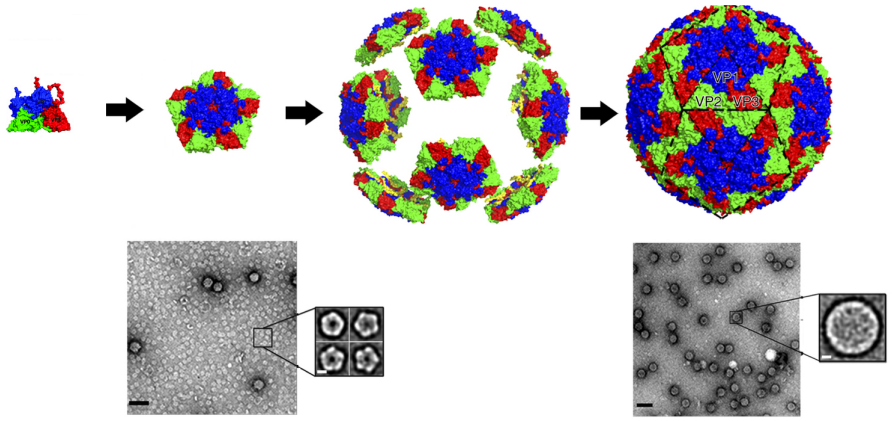


FIGURE 1.5: **FMDV capsid** Surface representation of atomic models of an FMDV capsid protomer. Five of these produce a pentamer, twelve pentamers form an intact virion. Capsid subunits are colour coded; VP1 is blue, VP2 is green, VP3 is red and VP4 is yellow. Negative-stain EM images are shown of dissociated pentamers and intact inactivated FMDV O1Manisa (O1M) capsids. Figure adapted from Kotecha *et al* [43, 51]

VP4 is internal to the capsid and has been predicted to interact with the RNA [28]. The other three capsid proteins (VP1, 2 and 3) are external. VP2, VP3 and VP1 are involved in antigenicity and cell entry [52–54]. VP1 has an extended loop that protrudes from the capsid surface. This loop, termed the G-H loop, contains an Arg-Gly-Asp motif (RGD motif) that mediates the binding of virus to integrin receptors on the cell surface [55]. Recombination and mutation within this region can allow access to cells with different cell surface receptors. This often involves the appearance of positively charged amino acids allowing the utilisation of negatively charged proteoglycan receptors such as heparin sulphate [56, 57]. The FMDV pentamer is acid labile and dissociates when exposed to a pH below 6.5 [58].

1.1.2.3 Genome organisation

FMDV is a single stranded RNA virus. The genome is approximately 8500 nucleotides in length [59]. The genome consists of a 5 prime un-translated region (5'UTR) a single open reading frame (ORF) and a 3 prime un-transalted region (3'UTR). Covalently bound to the 5' end of the genome is a viral encoded protein (VPg or 3B)[60]. Within the 5'UTR there are a number secondary structure elements: the short fragment (S-fragment), poly-c tract, RNA pseudoknots, the *cis*-acting replication element (*cre*) and internal ribosome entry site (IRES). The *cre* and IRES have been found to be important for translation and replication as described later. Although studies have been completed to consider their function, the exact role of the S-fragment [61, 62], poly-c tract [63, 64] and RNA pseudoknots [65] remain unknown. The 3'UTR also contains secondary structure elements. There are two stem loops directly following the ORF called stem loop 1 (SL1) and stem loop 2 (SL2) which are involved in viral replication. The 3' end of the genome consists of a poly-A tail (Fig. 1.6).

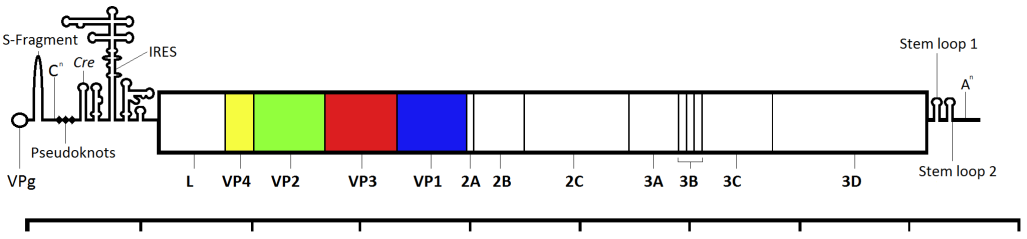


FIGURE 1.6: **FMDV genome** Schematic of the FMDV genome showing the secondary structure elements of the 5'UTR, the proteins coded for by the single ORF and the components of the 3'UTR.

1.1.2.4 Open reading frame

The FMDV genome encodes for a single open reading frame that is translated into a single polyprotein.

The N-terminus of polyprotein codes for a papain-like proteinase called leader protease (L^{Pro}). This protein has two start codons 84 nucleotides apart leaving it able to create two separate forms of this protein: Lab and Lb. Both proteins facilitate autocatalytic cleavage from the polyprotein (L^{Pro}/P1 cleavage site)[66]. Both forms of L^{Pro} also function to cleave host factors. For example, L^{Pro} acts as a viral deubiquitinase that negatively regulates the Type 1 interferon pathway [67] and cleaves a number of other host factors [68].

P1/2A, the section of the polyprotein downstream from L^{Pro}, contains the structural proteins VP1-4 (termed in some literature 1A-1D) and non-structural protein 2A. The structural elements undergo proteolytic processing to form the viral capsid as described previously.

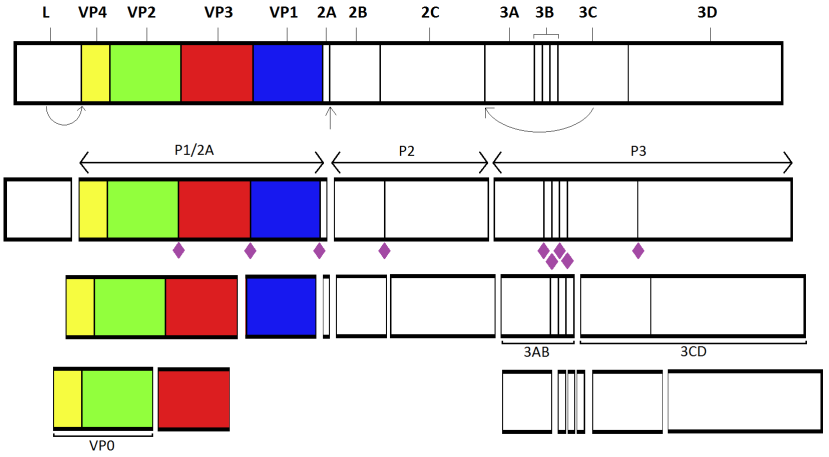


FIGURE 1.7: **FMDV Polyprotein** The leader protease separates itself from the polyprotein via autocatalytic cleavage. P3 is separated from the polyprotein via 3C protease cleavage. P1/2A is separated from P2 via ribosomal slippage. Subsequent processing is completed by 3C protease cleavage (purple diamond). The cleavage mechanism of VP0 into VP4 and VP2 is predicted to be RNA dependant autocleavage.

The subsequent section of the polyprotein, P2, contains non-structural proteins 2B and 2C. P2 is separated from P1 via ribosomal skipping. This is caused by the hydrolysis of a peptidyl-tRNA ester linkage formed during translation promoted by a sequence contained in the C-terminus of 2A [69]. The function of 2B in FMDV in not well classified. Work has found that it may function as a viroporin localised to the endoplasmic reticulum and acting to disrupt cell membranes. It has also been implicated in the formation of autophagosomes [70]. FMDV non-structural protein 2C has been associated with numerous functions. It has a predicted amphipathic helix at the N-terminus that may be associated with membrane binding [71]. It has ATPase activity [72] and has also been found to bind beclin, a key component of the cellular autophagy pathway [73]. It also has helicase domains [74].

The P3 coding region includes 3A, 3C and 3D. 3A is a membrane-associated protein that has been implicated in influencing host range [75–78]. There are three copies of 3B (or VPg), a feature unique to FMDV among the picornaviruses. VPg acts as a primer for initiation of RNA replication. 3C is a virally encoded protease. 3C processes P1-2A to produces the structural proteins (VP0, VP3 and VP4) as well as processing of P2 and P3 to produce the non-structural proteins [79] as shown in Fig. 1.7. 3D is an RNA

dependent RNA polymerase (RdRP). It has limited copying fidelity and determination of its crystal structure has shown no evidence of a proofreading domain [80, 81].

1.1.2.5 Life cycle

FMDV viral particles bind to cell surface receptors to trigger cell entry. In the case of the majority of wildtype FMDV strains the RDG motif on the GH loop of VP1 interacts directly with integrin receptors on the cell surface. Virions are then trafficked to endosomal compartments. In the case of integrin binding this is mediated by clathrin coated vesicles. These early endosomes have low internal pH [82]. The lower pH causes protonation of residues at the 2-fold axis of the virion, resulting in electrostatic repulsion of the capsid subunits from one another [83]. This exposes the hydrophobic internal capsid protein VP4 which interacts with the membrane of the early endosome to create a pore that allows for the release of viral RNA into the cytosol [84].

Translation can begin immediately via use of cellular machinery [85]. The majority of cellular mRNA translation is dependant on a 5' terminal cap. The 5' terminal of the mRNA is scanned by the ribosome until the 5' terminal cap is identified, this initial interaction then prompts recruitment of the rest of the cellular translational machinery. FMDV lacks this terminal cap structure. Instead, translation initiation relies upon the internal ribosomal entry site. A number of cellular proteins are essential for viral translation. These cellular proteins interact with the IRES [86, 87]. One such protein is translation initiation factor 4G (eIF4G). L^{Pro} cleaves eIF4G halting cellular translation. The C-terminal cleavage product binds with the IRES and 40S ribosomal subunit prompting ribosome assembly and viral RNA translation [88]. The virus shuts off host translation and transcription by cleavage of cellular proteins (for example eIF4G as described above) [89].

The mechanism that produces the switch from translation to replication is unknown. Some work has suggested that the accumulation of the precursor to the RdRP (3CD) triggers reduced host factor binding to the IRES [90]. Replication occurs in replication complexes formed from cellular membranes. These membranes are thought to be derived from the endoplasmic reticulum or golgi apparatus [91, 92]. The *cre* may act as a template for binding protein-linked dinucleotides to VPg creating VPGpUpU at the 5'end of the genome [93]. VPgpUpU acts as a primer for replication. Some work suggests the genome circularises bringing the VPgpUpU and poly-A tail together [94]. The generation of the VPgpUpU may also be *cre* independent at the 3'end of the genome [95]. FMDV RdRP 3D catalyses the elongation of the negative strand to generate double stranded RNA. This double strand is then unwound.

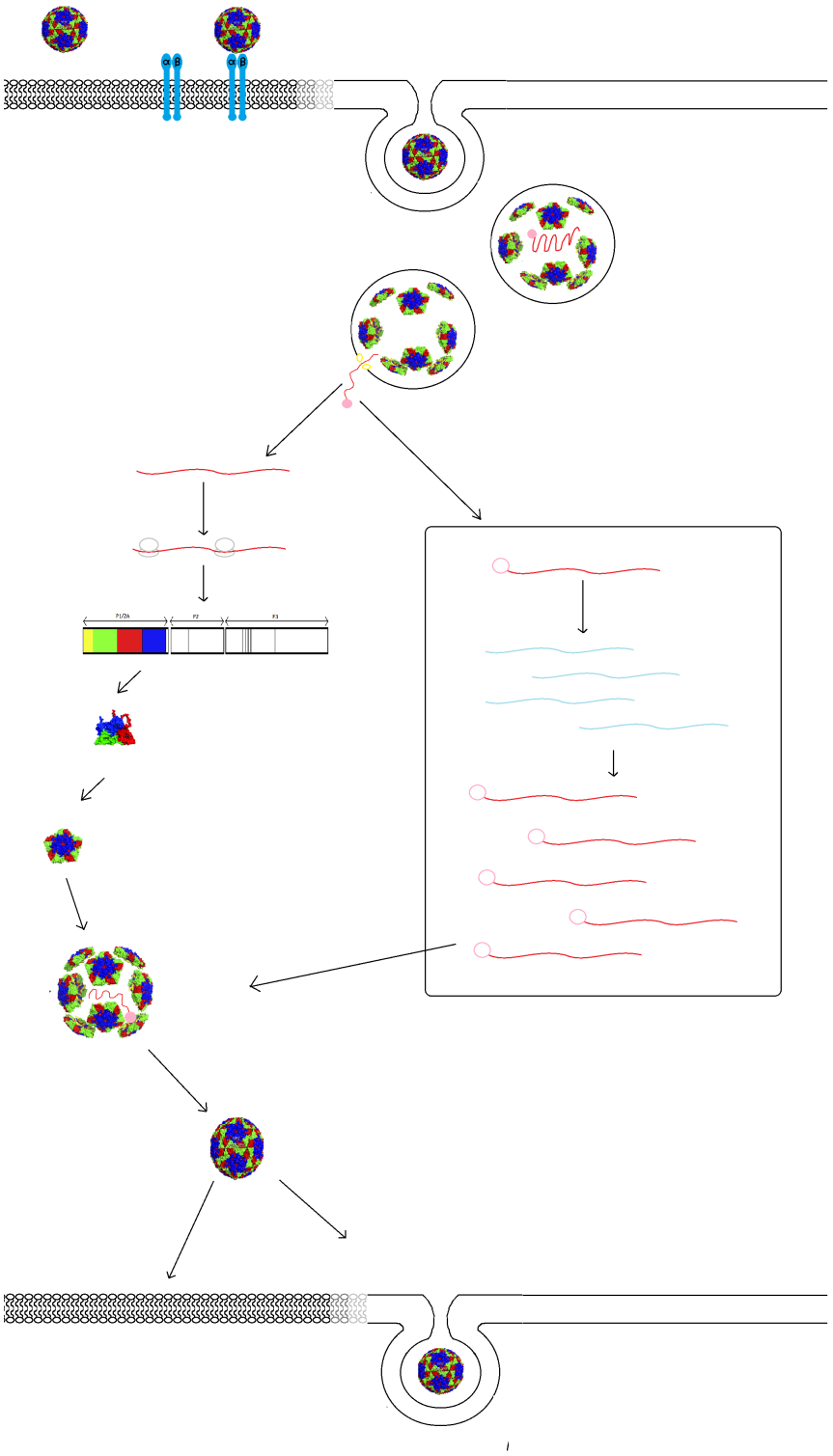


FIGURE 1.8: **FMDV Lifecycle** A cartoon of known elements of the FMDV life cycle [96]. FMDV virions bind to receptors on the cell surface facilitating cell entry. Acidic early endosomes cause capsid dissociation and VP4 forms a pore in the endosomal membrane allowing RNA to exit. CAP-independant translation utilising cellular machinery begins immediately and subsequent replication. RNA is packaged and the virions leave the cell via cellular lysis of prelytic release via cytoplasmic blebs.

The exact process of this is unknown although 2C has been implicated due to its helicase motifs [97]. Priming and elongation of a positive sense strand can then occur. A study in poliovirus has shown that a single negative sense RNA strand can act as a template for multiple positive strands although this work has not been replicated in FMDV [98].

RNA encapsidation is not well characterised. Studies in poliovirus have suggested that only newly formed RNA strands linked to a VPg are encapsidated leading to the suggestion that replication and encapsidation could in some way be linked [98, 99]. Whether RNA is inserted into a preformed capsid, as seen in other enteroviruses, or the capsid forms around the RNA utilising RNA packaging signals has yet to be discovered.

FMDV egress from the cell is generally accepted as being a results of cell lysis [100] although some evidence suggests there is some prelytic release of FMDV virions via cytoplasmic blebs [101, 102].

FMDV has a relatively short replication cycle *in vitro* with formed virions appear 4-5 hours post infection (hpi) [103].

1.2 Viral Evolution

A large number of RNA viruses are of global importance in both human and animal biology. Recent epidemics such as the 2014 Ebola epidemic has highlighted the importance of fully understanding these viruses and how they evolve [104]. Rapid rate of evolution complicates control of RNA viruses as it can allow them to become resistant to treatments [105].

Most RNA viruses are believed to exist as a swarm of genetically related variants. Due to their high mutation rates and short generation times they have extreme evolutionary dynamics with large population sizes exhibiting extensive antigenic variation and genetic diversity making traditional population genetics challenging [106, 107]. A viral population's ability to adapt to new environments and survive evolutionary bottlenecks is partially a consequence of the genetic diversity that is present in a viral swarm [108–110]. This has been clearly shown in field studies and experimental analysis. For example, deformed wing virus (DWV), originally a diverse viral population with relatively low pathogenicity amongst honey bees (*Apis mellifera*), overcame an evolutionary bottle neck to survive in *Varroa destructor*. This new vector improved transmission to *A. mellifera* and prevalence. Although resulting in an immediate decrease in diversity, the original diversity allowed its successful evolution and subsequent increased pathogenesis and proliferation [111]. Similarly, decreased diversity in *Polioviridae* has been shown to

decrease pathogenicity and decrease ability to survive in different tissue environments [112].

Viral evolution is largely attributed to genetic drift and recombination within these populations or swarms [113, 114]. Synonymous point mutations are suggested to occur across the genome and be fixed through positive selection [115]. Understanding how diversity allows viruses to adapt is critical to studying the spread of virus and controlling disease. For example for disease control, virus transmission tracing of FMDV in the 2001 epidemic [115], and the use of entropy to study viral evolutionary bottlenecks [116] and in the design of more multivalent vaccines, seen in the use of ancestral strains of HIV for vaccine design to increase cross-reactivity to more modern genetic variants [117].

1.2.1 How diversity is created

Genetic diversity within viral swarms can be attributed to a myriad of factors such as mutations rate, recombination, selective pressure, population size and speed of replication [118–121].

1.2.1.1 Viral mutation rate

RNA dependant RNA polymerases High mutation rates within viral species are considered to be due to their error prone polymerases and lack of proofreading [122]. Viral polymerases tend to exhibit low fidelity and a high rate of misincorperation. This can be affected by a number of things including RNA structure [123]. Secondary structure can affect the polymerases progression and some regions of code have been shown to cause a higher rate of polymerase slippage [124]. Viral RNA polymerase mutation rates are be predicted to be 10^{-4} mutations per nucleotide copied, much greater than the majority of DNA viruses [107, 125–128].

Measuring mutation rate There are numerous difficulties associated with accurate measurement of mutation rate. Haydon *et al* [129] succinctly summarised common issues in the prediction of mutation rates as;

- not all virus particles being equally infectious
- not all viral particles leaving a host at the same rate
- a limited number of hosts spreading the infection (not all host being infectious, even if they are infected)

- and biased selection in purification processes

Apparent mutation rates can also appear higher due to sequencing errors and may vary dependent on current selective pressures or periods of replication versus periods of dormancy [130]. It may therefore be inaccurate to assume a constant mutation rate for a virus. Further to the potential for varying mutation rates dependent on environment, it has been suggested that there is variation in mutation rate across the genome with less than 40% of sites within the capsids being conserved compared to highly conserved regions within 3B and D [131].

Error Threshold RNA viruses are predicted to live at the maximum error rate they can achieve and survive [119, 132]. The point at which the number of mutations that occur becomes lethal is referred to as the error threshold or error catastrophe [133]. Experiments have been completed to increase mutation rates in RNA viruses through mutagen introduction resulting in viral extinction due to reaching this error threshold [134–136]. Even minor increases in the mutation frequency in poliovirus (using ribavirin) reduce the percentage of infectious genome sufficiently to cause viral extinction [137].

1.2.1.2 Recombination

Recombination is an important factor in the creation of genetic diversity. Evidence of recombination has been found in FMDV [138, 139] although it's direct impact on the diversity of a viral swarm has not been investigated. Recombination has the capacity to have an extensive effect on viral evolution. For example, HIV has a higher rate of recombination than mutation [140].

1.2.1.3 Selective pressure

The RdRP introduces mutations; selective pressures determine which mutations outcompete others and thus remain in the population in higher number. Selective pressure has an effect on genetic diversity from an environmental to cellular level. Viral success is dependent upon a viruses ability to leave a host, its transmission capability, environmental stability, entry into a host, evasion of host immune system, entry into a cell and within cell replication efficiency. All of these selective pressures contribute towards the genetic diversity that persists within a viral population.

Mutation does not always result in a beneficial or adaptive outcome, viral evolution is often associated with fitness trade-offs and epistatic interactions as reviewed by Wargo

and Kurath [141]. It has, however, been repeatedly demonstrated that viral population diversity can provide a selective advantage by allowing virus to adjust to changes in environment. It should be considered that an adaptive advantage in one environment could have a negative effect in another. This non-constant nature of selective pressures on viruses has a great effect on genetic diversity.

1.2.1.4 Population size

Population size and replication speed also affect genetic diversity. A greater population size and higher replication rate results in the potential for a higher mutation rate and greater overall genetic diversity. Large population size allows for large numbers of variants to be present however a bottleneck transmission event can result in the fixation of a mutation and its propagation in the virus population regardless of its prevalence in the original large population. A negative correlation has been shown between the rate of nucleotide substitution and the size of genome regardless of population size [142]. This correlation promotes the view that a high level of mutation imposes limits on genome size unless error-correction mechanisms are in place such as seen in Coronavirus.

1.2.1.5 Replication strategy

Genetic diversity can be affected by viral replication strategy. For example, the number of positive and negative RNA strands present and being replicated has an overall effect on the level of genetic diversity [143].

1.2.2 How diversity is compared

1.2.2.1 Sequence space

A viral swarm's sequence space contains all possible sequences and can be represented geometrically where genetic similarity is represented by distance on the diagram. An example of this is seen in Fig. 1.9 [144]. Each round of replication a virus undergoes produces progeny that differ from the starting virus. Each round produces a more complex mutant spectrum with variants more genetically diverse (represented at a further distance from) the starting virus.

The space a sequence can cover is not infinite. The mutant genomes that arise will only continue to exist if they are able to replicate and persist. Detrimental or non-functional mutants will be removed from the population by purifying selection. The selective

pressure on this virus swarm is however complicated. A mutation on the genome that causes a minor fitness cost may be within a genome that is otherwise fit. This can cause low fitness mutants to appear in a swarm at a higher proportion than might otherwise be predicted [145].

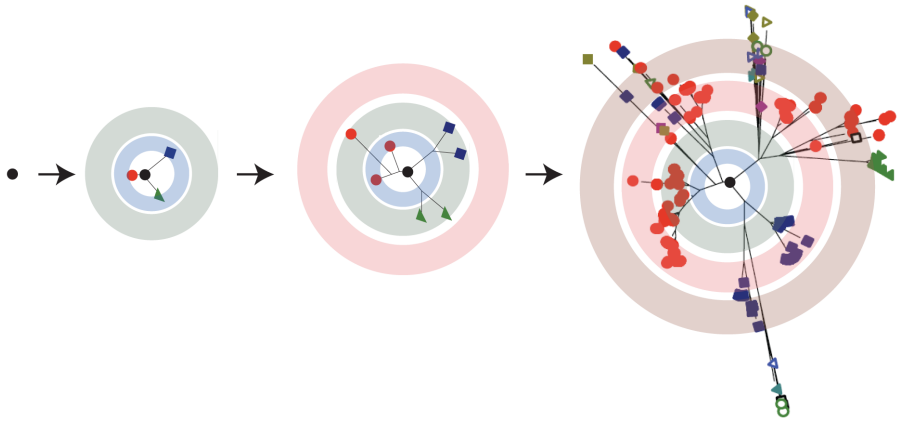


FIGURE 1.9: **Sequence space expands through replication** A geometric representation of a viral swarm's sequence space. Each round of replication produces a larger mutant spectrum in the progeny that is genetically diverse from the originally infecting virus. Figure taken from Lauring *et al* [144].

1.2.2.2 Fitness

The term fitness has been used to mean a variety of things within population dynamics and genetics [146]. Barker defines five different forms of fitness; phenotypic, genotypic, adaptedness, adaptability and durability [147]. In short this can be summarised as an organisms ability to survive and reproduce in their environment [146]. RNA replication rate can be used as an indicator of this but fails to consider numerous elements in viral persistence such as:

- the ability of the virus to transmit to a new host
- host range
- antigentic variation/immune escape

This becomes even more complicated when considering the virus population as a whole. The replicative efficiency or antigenicity/tropism/transmissibility of viruses within the swarm may be variable. Different viruses will be more or less fit than the progeny virus depending on the genomic mutations between them.

1.2.2.3 Survival of the flattest

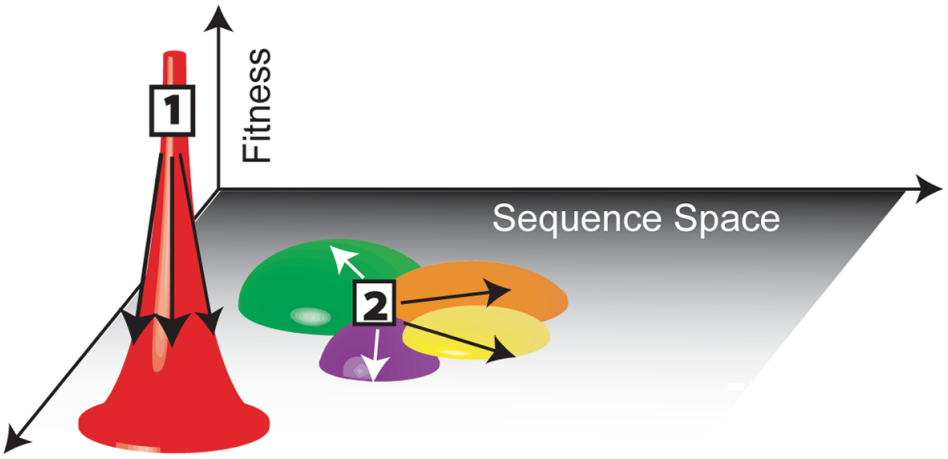


FIGURE 1.10: **A graphical representation of survival of the flattest** Population one is high fitness but any mutation will lead to a dramatic drop in fitness. In population two a mutation will have a lesser effect on fitness making it more mutationally robust. Figure taken from Lauring *et al* [144].

In Darwinian evolution there is a concept of survival of the fittest. In simplest terms, with replication as an indicator of fitness, the fastest replicator would be the most likely to survive. However, in RNA viruses this concept does not hold. Due to high mutation rate viruses that are most able to overcome deleterious mutations are the most likely to survive. Therefore it is potentially the slower replicators that are the more fit. With a high replication rate and low fidelity polymerase a virus swarm would be very diverse but many of viruses are less fit than the progenitor. In comparison low replication rate and a low fidelity polymerase produces a less diverse population with less low fitness variants. On average, the population replicating more slowly is more fit, it takes up a flat fitness landscape Fig. 1.10 [144]. This concept has been shown to be accurate via computer modelling [148] and in lab experiments [149]

1.2.2.4 Eigen's quasi-species

The concept of a quasi-species was first described in Eigen, McCaskill and Schuster's 1988 paper 'Molecular Quasi-species' [150]. In this paper they offered an alternative to Darwin's survival of the fittest;

'...not a single fittest type but an optimal ensemble - a master sequence together with its frequent mutants - will survive. Such stationary mutant distributions, usually distinguished by a unique consensus sequence, were named quasi-species...'

In a quasi-species natural selection acts on the the swarm of mutants in which a virus exists. This means the population evolves as a single unit. Research has considered if an RNA viral swarm conforms to this population structure. This would require experimentation to shown that selection acts on the population as a whole. To date it has been demonstrated that RNA viruses exists as a swarm of genetically related variants and changing the swarm structure (through high fidelity polymerases or the introduction of mutagens) does affect the pathogenicity and fitness of the infection. This does not categorically provide evidence of natural selection acting on the virus population as a unit. It merely shows the extensive effect populations size, polymerase fidelity and genetic bottlenecks can have on the successful progression of infection.

Until work has been completed that indicates natural selection acts on the swarm as a unit rather than individual viruses within it the term quasi-species may not be appropriate.

1.2.2.5 Bottlenecks

A population bottleneck is any event that causes a decrease in the size of a population. For example, in FMDV, transmission from one host to another has been found to be a population bottleneck or transmission from one site of replication to another [116]. The smaller population that overcomes a population bottleneck can be less diverse. Furthermore, smaller populations face a higher level of genetic drift which can result in lower level variants being lost, decreasing the diversity further [151].

1.2.3 Tools to create data sets

The advent of high throughput sequencing has allowed the more efficient and precise exploration of genetic diversity both within and between samples. The majority of past work has focussed on VP1 due to technological and financial limitations but new sequencing platforms allow for full genome sequencing and subsequent analysis.

1.2.4 High throughput sequencing

Research and investment in high through put sequence has been extensive in the last twenty years. The associated decrease in costs has made high through put sequencing a cost effective process in a research laboratory setting. Illumina estimate that the cost per gigabase has decreased from approximately one hundred million dollars in the year 2000 to 1-10 dollars on the HiSeqX in 2014. High throughput sequencing technologies can be broadly split into two categories; short read sequencing and long read sequencing.

1.2.4.1 Short read sequencing

Short read sequencing technologies fall into two categories; sequencing by synthesis (SBS) and sequencing by ligation (SBL). Both of these technologies involve the clonal amplification of the DNA fragment of interest. DNA amplification can be bead based (and beads are subsequently attached to a solid surface) and solid state. Amplification involves fragmentation of the DNA of interest and addition of an adaptor sequence that allows the fragmented DNA to bind to either a bead or solid surface (such as the Illumina flow cell). These bound fragments are then amplified on the bead or flow cell. The complete genomics technology (designed by the Beijing Genomics Institute) has a slightly different amplification involving the creation of what they term DNA nanoballs. This involves the iterative ligation, circularisation and cleavage of DNA fragments to create circular templates with adaptors. Rolling circle amplification of these fragments creates billions of nanoballs. These nanoballs are then distributed on a solid surface in a manner similar to bead base amplification technologies.

SBS technologies include The Roche 454 platform, Ion Torrent and Illumina HiSEQ, MiSEQ and NextSeq platforms. Illumina SBS is explained in detail below. SBL is the technology used by SOLiD and Complete Genomics. In this process the DNA fragments hybridises to a fluorescent probe. This is then ligated to an oligo and imaged. The fluorescence of the probe identifies the base to which it is complementary bound.

1.2.4.2 Long read sequencing

The development of long read sequencing technologies has allowed for the establishment of linkage information, identification of repetitive elements within genomes and also investigation of copy number alterations or variations in DNA structure. There are two main techniques for long read sequencing; single molecule real time (SMRT) sequencing and synthetic approaches. Synthetic approaches rely upon the existing short read technologies outlined above to synthetically construct long reads in silico. Examples of this include the Illumina long-read sequencing platform and 10x genomic instruments such as GemCode and Chromium. Synthetic read technologies do not actually generate long reads but build large fragments using computational assembly from DNA barcodes. SMRT sequencing technologies do create long reads. Two main technologies have been established in this area. The PacBio SMRT sequencer has a flow cell with millions of picolitre wells. Each of these wells has a transparent bottom with a polymerase molecule attached to it. The circularised DNA strand of interest passes through the polymerase through zero mode waveguides (ZMW). The dNTP that is incorporate is visualised with a laser and camera system focused on the polymerase. The laser excites and the camera

records the colour and duration of the light emitted as a tagged nucleotide is incorporated. The flurophore is subsequently cleaved to allow it do diffuse away from the censor. The circular nature of the DNA strand used allows the same template to be sequenced several time. This is information is then used to create a circular consensus sequence (CCS). The Oxford Nanopore works differently that all other technologies listed previously. Rather than monitoring the incorporation of hybridisation of a nucleotide it directly detects the compositions of a single stranded DNA molecule. Motor proteins pass the strand through a protein pore which has a current passed through it. The temporal trace of signals produced represents a shift in the voltage which is specific to each potential kmer.

These long read sequencing technologies offer a lot of interesting information for WGS experiments such as those included in this thesis. The ability to determine linkage and successfully determine repetitive regions have many potential uses. However, they are still in their relative infancy and have high error rates. Further development will allow for high speed sequencing of whole genome allowing for the identification of genotypes with in the viral swarm. Until then a mix of both long and short read sequencing may offer the best option [152].

1.2.4.3 Illumina sequencing by synthesis

Illumina have a number of sequencing platforms, including the MiSeq, HiSeq and MiniSeq, all of which use the same sequencing technology: sequencing by synthesis (SBS).

Library preparation Creation of a library of DNA fragments ready for sequencing is an important step of the sequencing process. Illumina sell several kits that can be used for library preparation one of which is the Nextera XT Libary preparation kit. The first step of Nextera XT library preparation is the simultaneous fragmentation of DNA and addition of adapters. This is achieved with the use of transposons that insert themselves into the double stranded DNA approximately 200bp apart (Fig. 1.11A). Primers are then added that are complementary to the adapters and 12 cycles of high fidelity PCR are completed to add sequencer binding sites, indexes and regions complementary to the flow cell (Fig. 1.11B).

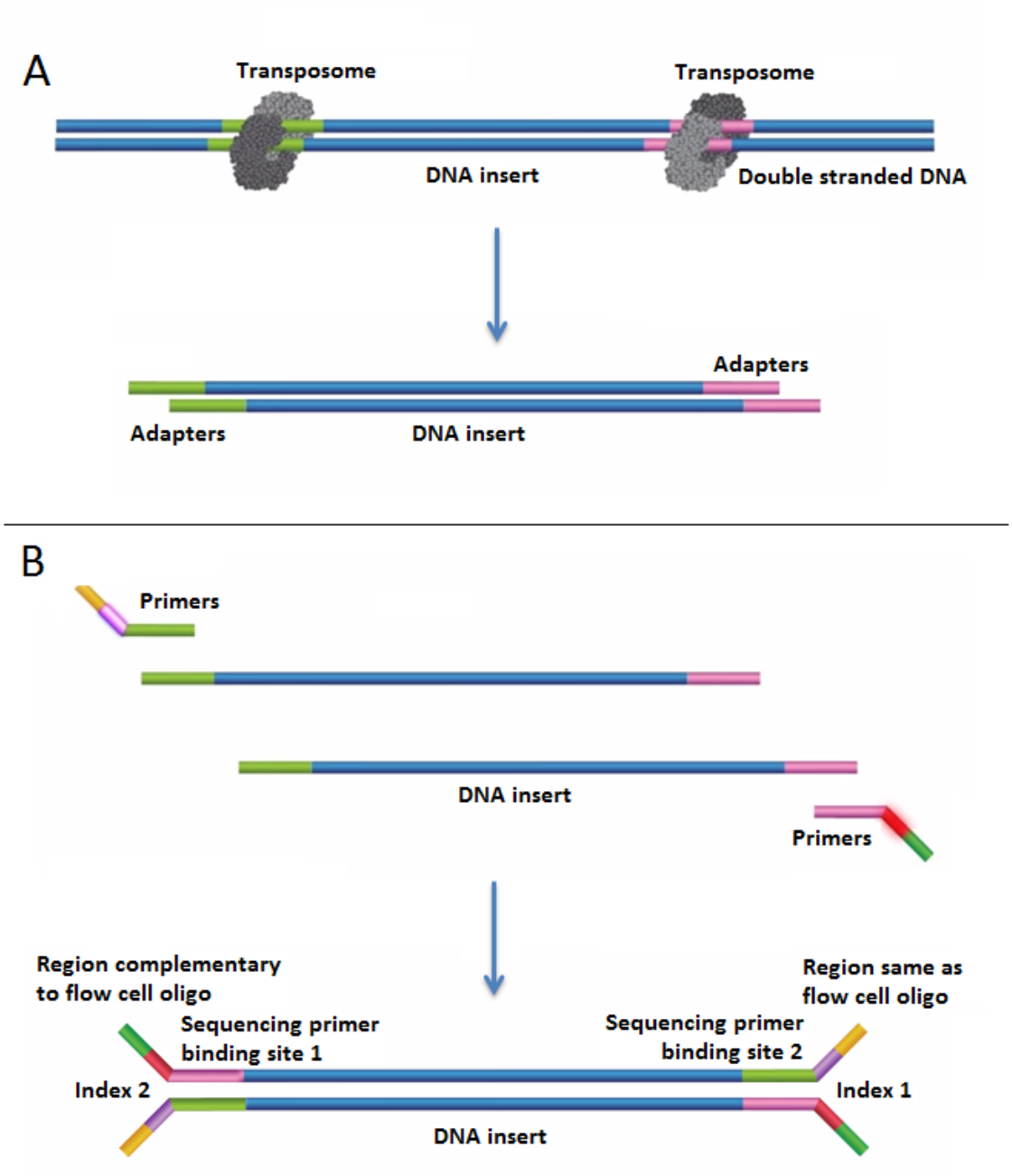


FIGURE 1.11: **Illumina NexteraXT library preparation** A) Transposons are added to double stranded DNA facilitating the fragmentation of the DNA and addition of adaptor sequences. B Primers are added, complementary to the adaptors, and reduced cycle amplification is preformed producing a double stranded DNA fragment with primer binding sites, indexes and regions complementary to the flow cell. Figure adapted from Illumina literature [153].

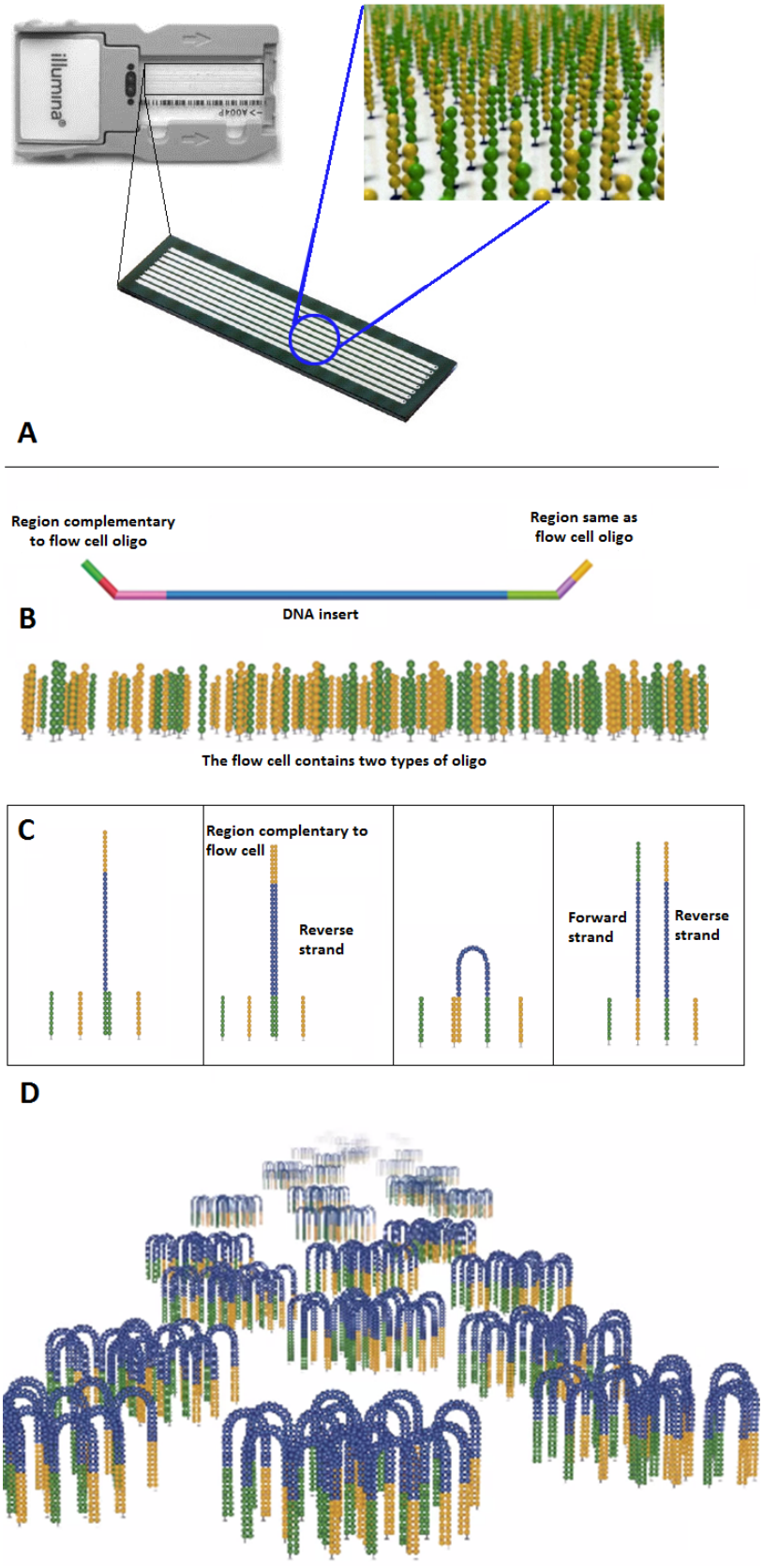


FIGURE 1.12: **Illumina cluster formation** A) The Illumina flow cell is a glass slide with several lanes coated with two types of oligo. B) The oligos are complementary to regions on the end of the DNA fragments C) Fragments undergo bridge amplification D) This process forms multiple clusters of DNA all representing the same original fragment. Figure adapted from Illumina literature [153]

Clustering The flow cell is a glass slide with multiple lanes coated with two different oligos (Fig. 1.12A). These oligos are complementary to regions on the DNA fragments (Fig. 1.12B). The end of the fragments hybridises with the flow cell. Polymerase is added and a reverse strand is created. This double stranded region is the denatured and the original strand is washed away. Bridge amplification is then completed, where stands are clonaly amplified. The strand folds and hybridises to the other oligo, polymerase generates a complementary strand, the bridge is denatured and two copies of the strand are left tethered to the flow cell (Fig. 1.12C). This is repeated until a cluster of fragments exists, each cluster representing only one original fragment of input DNA (Fig. 1.12D). This process occurs simultaneously for each fragment. The reverse strand is cleaved and washed off and the 3' end of the forward strand is protected to ensure it does not bind to the flow cell again.

Sequencing The process of SBS starts with the addition of a sequencing primer (Fig. 1.13A). A mix of all four fluorescently labelled nucleotide with reversible terminators are added (Fig. 1.13B).

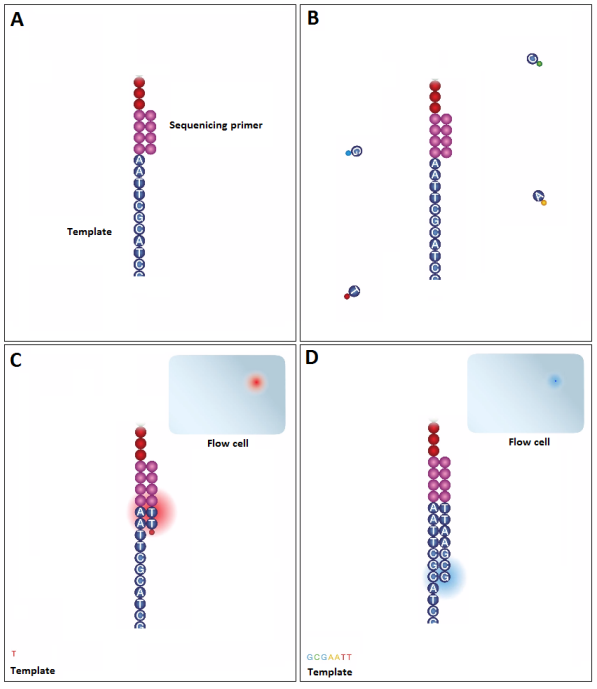


FIGURE 1.13: **Illumina sequencing by synthesis** A) Addition of a sequencing primer complementary to the strand. B) Addition of all four florescently tagged bases simultaneously. C) The complementary base is incorporated and identified by fluorescent excitation. D) This process is repeated until the length of DNA has been read. Figure adapted from Illumina literature [153]

The complementary base binds and the terminator prevents any further bases binding. This means only one nucleotide is added to each strand per cycle. The flow cell is then excited by a light source and the unique signal from each base allows the base that has been incorporated for that cycle for each cluster to be identified (Fig. 1.13C). The reversible terminator is then cleaved and the process repeated to read the rest of the DNA strand (Fig. 1.13D). The number of cycles represents the length of the read. The read product is then washed away and the associated index is read. This allows for each read to be pooled together with all other fragments originating from the same index. Using different indexes during the library preparation step means up to 96 samples can be multiplexed on to the flow cell and the data can bioinformatically sorted and attributed to the contributing sample. The 3' end of the sample is now de-protected and a new bridge forms. Index two is read at this point and the read product of that is washed away. Polymerase is added to produce a double stranded bridge, the strands are unwound and the forward strand washed away. The termini of the reverse strand is capped to ensure it does not re-bind to the flow cell. Read two sequencing primers are added and the sequencing of the reverse strand of each cluster begins.

Bioinformatic analysis Each read produced by the sequencer is sorted by the index associated to it, providing files with hundreds of thousands of reads attributed to each sample muliplexed on a run. Each position in each read is associated with a quality score which represents the probability the base has been correctly identified. Decisions can be made to discard reads that are of a low quality. Otherwise, these reads are built together to create a sequence by aligning over lapping region (de novo assembly) or aligned to an already published relevant genome (Fig. 1.14).

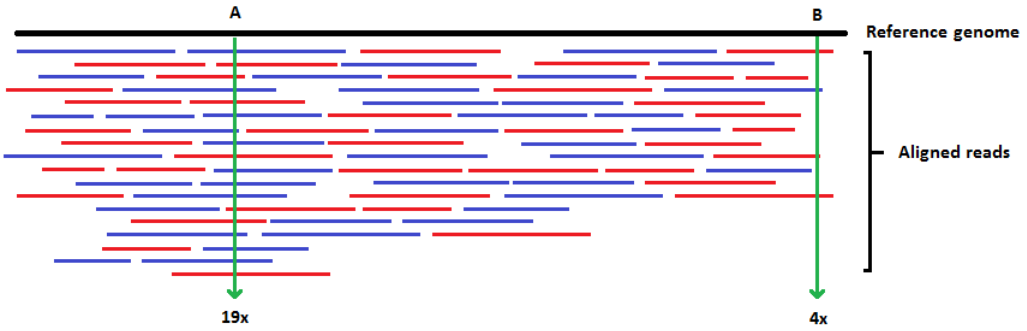


FIGURE 1.14: **NGS read coverage** Forward (red) and reverse (blue) reads are orientated and aligned to a reference genome. The coverage of each genome positions is the number of reads that represent that region. Position A is represented in 19 reads therefore its coverage is 19. Positions B is represented in 4 meaning its coverage is 4.

If a large number of reads has been produced per sample then several reads will overlap the same region of the genome, this increased coverage has seen more detailed information gathered and lower frequency variants detected .

1.2.4.4 NGS Limitations

The limitations to the methodology associated with these high-throughput machines include error accumulation. Errors accumulate throughout the sequencing process particularly in amplification and reverse transcription steps. If PCR is used to amplify a sample artifacts can be incorporated through physical errors such as artificial recombination or nucleotide misincorperation or via skewing of amplification and subsequent obscuring of true sampling. While errors might not be important for the assembly of consensus sequences, they have a significant impact in experiments where individual reads are being interrogated. It is clearly necessary to overcome these issues and so far efforts have been based on correcting for errors through mathematical models or tagging samples with an ID that can be traced throughout the PCR process [143, 154]. The ability to reach the required quantity of sample to begin library preparation (1ng of cDNA for the Nextera XT Kit) without amplification steps would be greatly beneficial.

There are also some problems relating to variety of the samples multiplexed. Similar sequences in close proximity on a flow cell can cause inaccurate reads by the machinery. This is an issue that has been acknowledged by the manufactures and as such each machine is able to provide a quality score. Sadly as yet there is no scale on which to compare this between platforms as reviewed by Beerwinkel et al [155].

Method	Adapter type	Amplification?	Separation	Sequencing chemistry	Approximate read length (bases)†	Approximate maximum amount of data per run†
Roche 454*	Adapters	Emulsion PCR	Microbeads and 'picotitre' plate	Pyrosequencing	400–700	700 Mb
SOLiD	Adapters	Emulsion PCR	Beads on glass slide	Ligation	50–75	20 Gb
Illumina*	Adapters	Bridge amplification *in situ*	Glass slide hybridization	Reversible terminators	25–500	600 Gb
Helicos	Poly(A) adapter	No amplification	Flow-cell hybridization	Reversible terminators	25–55	35 Gb
PacBio	Hairpin adapters	Linear amplification	Captured by DNA polymerase in microcell	Fluorescently labelled dNTPs	1000	Not available
Ion Torrent*	Adapters	Emulsion PCR	Ion Spheres and high-density array	Detection of released H^+	35–400	1 Gb

FIGURE 1.15: **Summary of NGS platforms and their capabilities** Different sequencing platforms vary on their adapter type, amplification method, sample separation and sequencing chemistry. This can affect the length of the read produced and the amount of data collected from a sequencing run. Figure taken from Radford *et al* [156].

There are a number of variations between platforms (Table 1.15). Roche GS FLX 454 offers slightly longer reads but Illuminas MiSeq and HiSeq enable ultra-deep coverage. Considering the prediction of diversity is reliant on distribution of reads along the

genome rather than length, the latter offers a greater service in this research area. It should be noted that this platform is relatively new to the market and concrete protocols have not yet been established. As such read coverage has been shown to differ by orders of magnitude [155].

Error correction methods Due to the challenges associated with distinguishing sample preparation introduced error from low level variants significant work in recent years has been focused on producing methodologies that limit or identify errors so they can be accounted for in downstream analysis pipelines.

CirSeq Circular-sequencing is a technique established by the Andio Lab [157, 158]. It focuses on the circularisation by ligation of small fragments of the genome or gene of interest. These circular fragments are then used in subsequent sample preparation and sequencing steps. The circular nature of the genomic fragment results in products of the RT reaction that contain several repeats of the same sequence (Fig. 1.16).

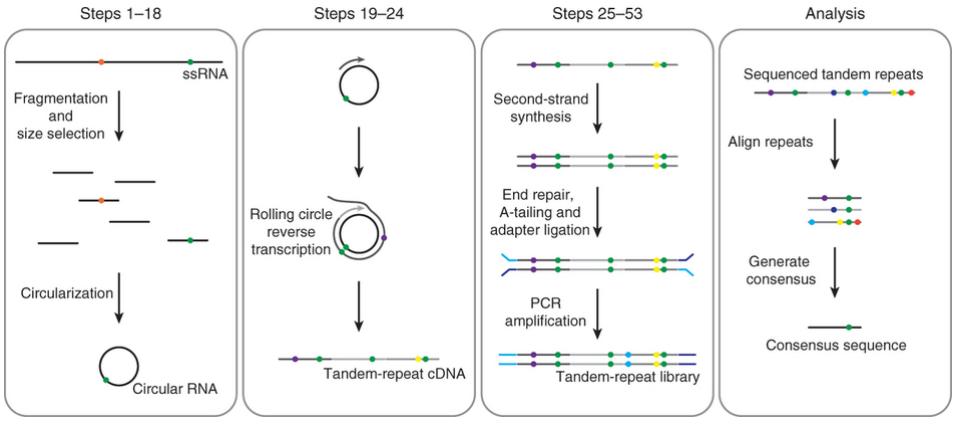


FIGURE 1.16: **Schematic of CirSeq methodology** A schematic of CirSeq methodology taken from Acevedo and Andino's paper 'Library preparation for highly accurate population sequencing of RNA viruses' [157]. Colour dots represent true genetic variants (green) and process introduced errors (all other colours).

For example, if the circularised fragment is 200nt long and the read length is 600nt (assuming efficient reverse transcription) a read can include three tandem repeats of the same region. These tandem repeats are then aligned and only mutations that appear in all three repeats are accepted. This works to identify and remove RT, PCR and sequencing introduced errors. This allows the method to have a greater resolution and confidently identify low level variants. A direct comparison was made between CirSeq and the sequencing methodolgy used in this thesis (Appendix A). Although CirSeq offers

an excellent error correction methodology and thus improved resolution of minority variants the cost associated with the process both in time and financial reagents is extensive. For this method to work each sample must be run individually on the sequencer without multiplexing. This was not viable due to reagent costs. Consequently, this method was not used in this work.

Barcoding Barcoding is the process of individually tagging RNA or DNA molecules with a unique ID or barcode. This methodology has numerous uses in NGS. It has been used to consider questions such as how diversity is transmitted from host to host. For example, previous studies have used barcoding to understand how virus populations are transmitted between individuals [159].

Barcoding not only allows for the tracking of variants but also for the correction of errors and biases. As each sequence is individually identifiable, repeats can be grouped together and a consensus sequence created as with CircSeq described above [157, 158]. This results in the irradiation of sample preparation induced errors and sequencing errors. The distribution of each unique barcode can also be considered to address PCR re-sampling errors [154, 160]. All of this combined allows for a better resolution of low level variants. Therefore studies using this technique have been completed to consider low level drug resistant variants [154, 160, 161].

The barcoding methodology relies upon the unique identifying sequencing to be sufficiently different to be recognisable. This has its own challenges. PCR or sequencing errors can result in false barcodes otherwise known as offspring primer IDs [162, 163]. Barcode design is therefore complex. Barcodes can be designed to account for the mutation spectrum created by a sequencer. For example, the Illumina MiSeq most commonly produces substitution errors, so barcodes must be designed that will not resemble one another after one substitution [164]. Increasing the length of the barcode can aid in this although this is clearly limited by read length. To ensure the barcodes include sufficient redundancy to be distinguished from background noise error correction codes such as Hamming codes are used in their design [162]. Although clever design can improve these issues the number of unique barcodes that can be used is limited by their similarity to each other and distinguishing them from background noise [163].

This limitation in barcode number and the complexity of its use meant using it in this study was not viable although future work to consider the transmission of genetic diversity would benefit from this methodology.

1.3 Objectives of PhD

This introductory chapter describes how virus survival and evolution is greatly affected by the viral swarm. The focus of this thesis is to investigate the swarm of viruses in which foot-and-mouth disease virus exists.

All current sequencing methodologies for RNA viruses involved an RT step (to produce cDNA from the RNA genome) and subsequent PCR to amplify the cDNA of interest. The PCR has the capacity to introduce errors that can be misconstrued as variation in deep sequencing giving an incorrect view of the viral swarm. Chapter 2 describes the design of a method to sequence the swarm without PCR. This method focuses on the production of data sets that represents the entire swarm rather than solely the consensus/majority genome. This method does not use PCR during sample preparation reducing the errors produced and improving the resolution of low level variants. This method is validated against multiple poly-adenylated viruses and has the capacity to be used in both a high-throughput consensus sequencing environment and in deep sequencing methodologies such as the experiments included within this thesis.

FMDV has previously been shown to exist as a swarm of genetically related variants but there is no detailed understanding of the dynamics of this swarm. Chapter 4 makes use of the newly designed sequencing method to define the parameters of the swarm and consider how it can be compared from isolate to isolate. This work looks at how diversity indices can be used to describe variation within a swarm similarly to how they indicate variation in an ecological population. This work shows to what extent the viral swarm varies between different isolates of FMDV.

Having defined the parameters of the swarm, work was completed to consider what determines the swarm structure. Chapter 5 considers the effect the host can have on the viral swarm. Previous work has suggest that the SAT viruses are highly variable in comparison to their European counterparts. However, this work shows that this variability may not be a product of the virus itself but the host from which it is isolated. This indicates that evaluating the swarm can offer information about the environment in which a viruses exists. This work is also important from a diagnostic perspective as some subconsensus level variants identified were of antigenic importance.

Chapter 6 looks at how the swarm changes during adaptation to a new host. FMDV is a highly adaptable virus with a broad host range. Consensus level changes and different phenotypic features are investigated. Virus swarm structure was considered under the adaptive pressure of a new host cell line. This work suggests in periods of adaptation or no adaptation limitations on swarm variation may result in predictable dynamics

of variable positions. This indicated how useful interrogating the swarm can be in understanding the pressures a virus may or may not be under.

Chapter 7 shows how an understanding of the swarm can help to investigate fundamental questions about the virus life cycle. Investigation of the swarm that is successfully packaged is used to identify genome regions required for viral encapsidation. This work starts to bring clarity to a long unanswered question of packaging in the *Picornaviridae*. Furthermore, the methodology used for this experiment would be highly transferable to other viruses with unknown packaging constrains.

Chapter 8 contains a summary of what was investigated in this thesis and a discussion of what this new information tells us about FMDV evolution and adaptation. The novel insight this work produces is commented upon as well as how these finding add to the current body of knowledge. Future work that could be completed to add clarity to un answered questions is also included.

Chapter 2

PCR-free sample preparation for sequencing

2.1 Introduction

Next generation sequencing (NGS) protocols for RNA viruses usually involve a reverse transcription (RT) reaction to produce a cDNA template from the RNA genome and subsequent rounds of polymerase chain reaction (PCR) to amplify the cDNA. PCR allows for the specific amplification of a genome or gene of interest. This results in a higher percentage of the final sequencing data representing the region of interest [165]. In some samples the quantity of genetic material may be low and thus the ability to specifically amplify it can be extremely beneficial. However, as next generation sequencing technologies improve the requirement for input material is decreasing and thus this will become less of an issue.

Both RT and PCR enzymes are error prone. The extent of the error rate introduced by RT, PCR and sequencing is thoroughly summarised by Orton *et al* [166]. Even with high fidelity enzymes there is an inherent rate of misincorperation in both process meaning the strand copied may not be identical to the template. This can be magnified when an error is introduced in an early round of PCR and subsequently copied numerous times. These so called 'jackpot mutations' can then appear to be relatively high level variants within a population. Due to the high level of error introduced in sample preparation analysis of subsequent data requires error correction. This often relies upon an error threshold cut-off. Variants below this cut-off will not be considered as they could be a result of error introduced in sample preparation. Single nucleotide polymorphisms (SNPs) represented at a high rate and overall distributions of variants can be tracked

and PCR NGS sequencing techniques have been highly successful in numerous studies [116, 167–169]. However, these methodologies often miss low level variants.

PCR methodologies are also limited by the fragment length they can produces. Multiple PCR reactions may be needed to cover a genome. The placement of primers can result in under represented regions where the primer sits.

In an effort to decrease errors incorporated and avoid jackpot mutations a sample preparation method was designed that did not include a PCR amplification step. The yield from the method described was sufficient to allow sequencing on the Illumina Miseq and was not limited by fragment length [170].

2.2 Published Method

The following work has been published (Fig. 2.1)[170] and is included in its entirety below.

FIGURE 2.1: Methodology paper (BMC Genomics). The method outlined in this chapter has been published in BMC genomics in 2014 [170].

2.2.1 Abstract

Background Next-Generation Sequencing is revolutionizing molecular epidemiology by providing new approaches to undertake whole genome sequencing (WGS) in diagnostic settings for a variety of human and veterinary pathogens. Previous sequencing protocols have been subject to biases such as those encountered during PCR amplification and cell culture, or are restricted by the need for large quantities of starting material.

We describe here a simple and robust methodology for the generation of whole genome sequences on the Illumina MiSeq. This protocol is specific for foot-and-mouth disease virus or other polyadenylated RNA viruses and circumvents both the use of PCR and the requirement for large amounts of initial template.

Results The protocol was successfully validated using five FMDV positive clinical samples from the 2001 epidemic in the United Kingdom, as well as a panel of representative viruses from all seven serotypes. In addition, this protocol was successfully used to recover 94% of an FMDV genome that had previously been identified as cell culture negative. Genome sequences from three other non-FMDV polyadenylated RNA viruses (EMCV, ERAV, VESV) were also obtained with minor protocol amendments. We calculated that a minimum coverage depth of 22 reads was required to produce an accurate consensus sequence for FMDV O. This was achieved in 5 FMDV/O/UKG isolates and the type O FMDV from the serotype panel with the exception of the 5' genomic termini and area immediately flanking the poly(C) region.

Conclusions We have developed a universal WGS method for FMDV and other polyadenylated RNA viruses. This method works successfully from a limited quantity of starting material and eliminates the requirement for genome-specific PCR amplification. This protocol has the potential to generate consensus-level sequences within a routine high-throughput diagnostic environment.

2.2.2 Background

Foot-and-mouth disease has been associated with severe productivity losses in cloven-hoofed animals characterised by vesicular lesions of the feet, tongue, snout and teats as well as fever and lameness [171]. The disease has a serious impact upon food security, rural income and significant economic consequences for any country harbouring the virus [172]. An integral part of any viral disease control strategy is the epidemiological tracing of virus transmission together with conventional field investigations. For RNA viruses with high evolutionary rates, this is routinely achieved with the application of molecular and phylogenetic methods [167, 168, 173] one example being the global tracing of foot-and-mouth disease virus (FMDV) [174]. Next-generation sequencing platforms offer much promise as rapid, cost-effective, and high-throughput methods for the generation of viral genome sequences. Recovering whole genome consensus level sequences of viruses provides important information for outbreak epidemiology and pathogen identification [175–178].

The positive-sense single-stranded RNA genome of FMDV is comprised of a single long open reading frame. This encodes a polyprotein which is flanked by 5' and 3' untranslated regions of approximately 1200 nt and 95 nt, respectively, terminating in a poly (A) tail. The 5' UTR contains highly structured RNA which is involved in both replication and translation. Approximately 300-370 nt from the 5' end of the genome lies a homopolymeric cytidylic acid [poly(C)] tract of 100-170 nt [179]. The genome sequence upstream of the poly(C) tract is known as the S fragment and that downstream as the L fragment.

Previously, tracing and monitoring of the trans-boundary movements of FMDV has been successfully achieved using consensus sequences of the VP1 region [9, 180, 181]. However, over shorter epidemic time scales, where viral populations have not substantially diverged, VP1 sequencing cannot provide the required resolution to discriminate between viruses in field samples collected from neighbouring farms within outbreak clusters. At this scale, WGS at the consensus level has proven to be a powerful tool for the reconstruction of transmission trees [182].

Previous strategies for viral WGS include PCR and Sanger sequencing methods or microarray approaches [182, 183]. Commonly, these processes have limited throughput and are both resource and labour-intensive with biased outputs that may not reflect the true diversity within samples [184, 185]. Furthermore, such methodologies have been subject to errors incumbent within the nature of the protocol i.e. those protocols reliant upon DNA amplification generate biased datasets from which it is difficult to make firm conclusions [186]. Such strategies have also been dependent upon a priori knowledge of virus sequences for primer design and are limited by potential inter and intra-sample sequence variation [187].

This study describes the optimisation of a robust, high-throughput protocol for WGS of all seven serotypes of FMDV excluding the 5' genomic termini and poly(C) tract. It does not use PCR amplification prior to the sequencing steps and overcomes the requirement for large starting quantities of template nucleic acid, which has previously limited the suitability of some NGS technologies for processing viral field isolates [188–190]. This protocol, with minor changes, was also applied to other polyadenylated RNA viruses.

2.2.3 Methods

2.2.3.1 Virus specimens

The protocol was initially developed and validated using an FMDV field isolate (O/UKG/35/2001) submitted to the FAO World Reference Laboratory for FMD (WRLFMD,

Family	Genus	Species	Serotype	Isolate	Passage history
Picornaviridae	*Aphthovirus*	Foot-and-mouth disease virus	O	UKG/1734/2001	10% epith. susp.
				UKG/1450/2001	10% epith. susp.
				UKG/14597/2001	10% epith. susp.
				UKG/1558/2001	10% epith. susp.
				UKG/4998/2001	10% epith. susp.
				UKG/1485/2001	10% epith. susp.
				UKG/35/2001	10% epith. susp.
				TUR/11/2013	BTy2
			A	TUR/12/2013	BTy2
			C	KEN/1/2004	BTy2
			Asia 1	TUR/13/2013	BTy2
			SAT 1	TAN/22/2012	BTy2
			SAT 2	TAN/5/2012	BTy2
			SAT 3	ZIM/6/91	BTy2
		Equine rhinitis A virus	1	D1305-03, dromedary, Dubai, 2003	Vero2
	Cardiovirus	Encephalomyocarditis virus	1	VR-129B, chimpanzee, Florida, 1944.	BHK3
Caliciviridae	Vesivirus	Vesicular exanthema of swine virus	B34	B1-34, pig, California, 1934.	PK5, IB-RS5

TABLE 2.1: **Viruses used in development and validation of the non-amplification protocol BTy**: Primary Bovine Thyroid ; **PK**: Pig kidney epithelial cells; **BHK**: Baby Hamster Kidney; **IB-RS**: instituto Brazilia Renal Swine; Numbers denote passage number.

Pirbright, UK) during the 2001 FMD outbreak in the United Kingdom. It was further validated with a panel of other samples originating from this outbreak as well as with a panel of viruses representing all FMDV serotypes. The protocol was also validated with other representative polyadenylated RNA viruses. The details of all viruses used in the study are described in Table 2.1. Where appropriate, viruses were cultured for one replication cycle in bovine thyroid cells (BTy) as described previously [191]. Dilutions between 1x10^8 to 1x10^6 viral copies/μl of O/UKG/35/2001 were made with viral cell culture supernatant in virus negative suspensions of bovine epithelium to mimic real clinical samples with different viral loads.

2.2.3.2 RNA extraction and FMDV-specific RT-qPCR

	Primer Name	Primer Sequence
RT-qPCR	Callahan 3DF [192]	ACTGGGTTTTACAAACCTGTGA
	Callahan 3DR [192]	GCGAGTCCTGCCACGGAA
	Callahan 3DP [192]	TCCTTTGCACGCCGTGGGAC
First-strand synthesis	UKFMD Rev6 [115]	GGCGGCCGCTTTTTTTTTTTTTT
	NK72 [193]	GAAGGGCCCAGGGTTGGACTC
	UKFMD UKG 4926R	AAGTCCTTCCCGTCGGGGT
	EMC-2B65R [194]	TCGGCAGTAGGGTTTGAG
	ERAV-2A22R [195]	GGGTTGCTCTCAACATCTCCAGCCAATTT
	Vesi-3D1R	CKNGTNGGYTTNARNCC
	Vesi-3D2R	TANCANCCRTCRTCNCCRTANGT

TABLE 2.2: **Primers and probes used in quantitation and WGS of FMDV and other RNA viruses** International Union of Pure and Applied Chemistry (IUPAC) nucleotide ambiguity codes: N: G or T or A or C; K: G or T; Y: T or C; R: G or A

Total RNA was extracted from 460 μl of cell culture virus isolate or original suspension [consisting of 10% tissue suspensions generated in M25 phosphate buffer (35 mM $Na_2HPO_4.2H_2O$; 5.7 mM KH_2PO_4; pH 7.6; made in-house)] using RNeasy MiniKit (Qiagen) according to manufacturers instructions. Total RNA was eluted in 50 μl of nuclease-free water and quantified using the Qubit RNA High Sensitivity (HS) Assay Kit (Life Technologies). FMDV-specific RNA was detected using an FMDV-specific real-time RT-qPCR as described previously (Table 2.2) [192] and quantified using an RNA standard derived from O/UKG/35/2001.

2.2.3.3 gDNA depletion

Genomic DNA (gDNA) was depleted from extracted total RNA samples through the activity of rDNase1 using the DNA-free DNAse kit (Life Technologies). Briefly, 10 μg of extracted nucleic acid in a 50 μl volume was combined with 5μl of DNase Buffer and 1 μl of rRDNase1 (2 U), and incubated at 37 C for 30 min. Inactivation agent was added as per manufacturers protocol and the sample was incubated for a further 2 min at room temperature with periodic mixing. The samples were then centrifuged at 17,000 xg for 2 min and the DNAse-treated supernatant was retained for subsequent processing.

2.2.3.4 cDNA synthesis

First-strand cDNA synthesis (reverse transcription) was performed using Superscript III First-Strand Synthesis System (Life Technologies) according to the manufacturers protocol. Briefly, 10 μl of DNase-treated total RNA was combined with oligonucleotide primers (Rev6 (2 μM), NK72 (2 μM) or FMDV-4926R (2 μM)) depending on the application of the protocol, random hexamers (50 ng/μl: Life technologies), dNTPS (10 mM: Life Technologies) and nuclease-free water (Life Technologies) (Table 2.2). Reactions were incubated at 65°C for 5 min and cooled on ice for 5 min. A second reagent mix was added containing SuperScript III enzyme (200 U: Life Technologies), RNaseOUT (40 U: Life Technologies), 0.1 M dTT (life Technologies) and 25 mM $MgCl_2$, before incubating at 50°C for 50 min. A final incubation with RNase H (2 U: Life Technologies) was then performed at 37°C for 20 min.

Second-strand synthesis was performed using NEB Second Strand Synthesis kit (NEB) as per manufacturers instructions using 20 μl of cDNA. The resulting dsDNA was purified using Illustra GFX DNA/gel clean-up kit (GE) as per manufacturers instructions and samples eluted in 30μl of nuclease-free water. Double-stranded cDNA samples were then quantified using the Qubit dsDNA High Sensitivity (HS) Qubit kit (Life Technologies) after which samples were adjusted to 0.2 ng/μl using nuclease-free water where appropriate prior to library preparation.

2.2.3.5 Illumina library preparation

One nanogram of each dsDNA sample was used to prepare sequencing libraries using the Nextera XT DNA Sample Preparation Kit (Illumina) according to manufacturers instructions. Libraries were sequenced on a MiSeq using 300 cycle version 2 reagent cartridges (Illumina) to produce paired end reads of approximately 150 bp each.

Sequence data analysis Consensus sequences were attained using a complete published sequence as a template or, where a closely related template was not available, a de novo assembly. Sequence read quality was monitored with FastQC [196] prior to Sickle [197] trimming all bases with a q score of <30. For de novo trimmed Fastq files were processed using Velvet v1.2.10 [198] with an optimum Kmer length determined by Velvet-Optimiser. A minimum contig length of 1000 was included in L fragment analysis. A BLAST search with the contigs confirmed viral origin [199]. Final contig assemblies were completed manually in BioEdit [200]. Alignments between MiSeq data and appropriate reference genome (from publication or de novo assembly) were completed using Bowtie2.1.0 [201] and SAM/BAM processing carried out using Samtools [202]. Alignments were visually checked using Tablet [203]. Coverage data and graphs were generated using Bedtools [204] with final graphical output produced using Prism v6 (GraphPad).

Read coverage required to obtain an accurate consensus sequence A sorted alignment file (.sam) of FMDV O/UKG/35/2001was generated using Bowtie2.1.0 [205] and Samtools [206]. A bespoke python script that truncated the samtools mpileup output format (Appendix D) was used to simulate files with varying levels of coverage. A consensus sequence was generated from each of these files using mpileup (Samtools). The consensus sequences created were compared in BioEdit [200] and their sequence identities recorded. This was completed for all FMDV type O isolates with a sufficient number of reads and the mean was calculated.

2.2.4 Results

Protocol accuracy: calculations of minimum coverage required for accurate consensus Next-generation sequencing analysis provided large numbers of short read sequences that were assembled and aligned in order to determine a consensus sequence. To define how much redundancy was required for accurate reconstruction of consensus level sequences, we determined the minimum read coverage required to obtain a robust consensus from the protocol described. Analysis was completed on all FMDV type O samples with sufficient coverage (Fig. 2.2). From this a mean was calculated showing a minimum coverage of 22 reads was required to obtain an accurate consensus sequence in this instance.

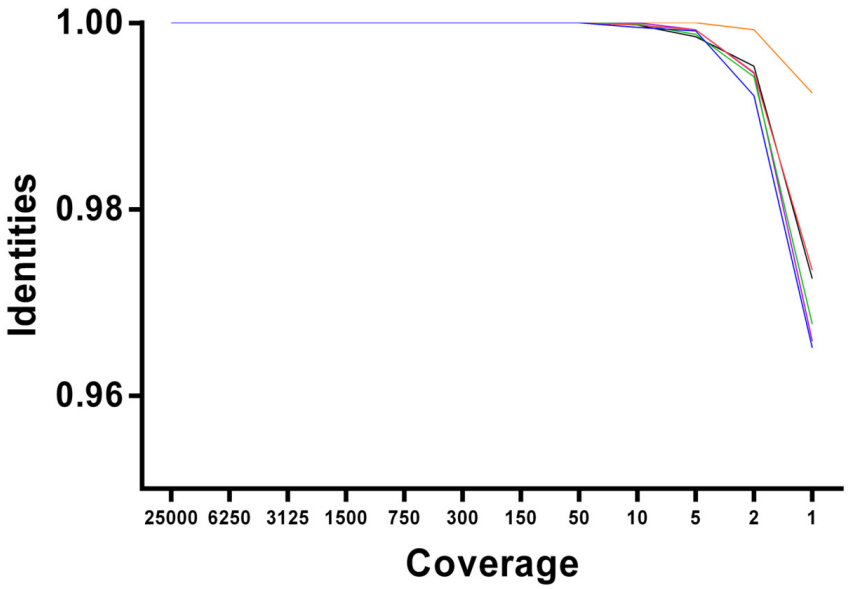


FIGURE 2.2: Read coverage required to obtain an accurate consensus sequence. The consensus sequence resulting from varying levels of coverage was assessed for accuracy. Isolates O/UKG/1450/2001 (blue), O/UKG/1558/2001 (green), O/UKG/1734/2001 (purple), O/UKG/4998/2001 (orange) and O/UKG/14597/2001 (red) alongside the type O exemplar from the serotype panel (black) were analysed. Points on the graph represent a comparison of the identities (scored on the y axis) of a consensus made with total reads and a consensus made with limited read coverage (detailed on the x axis). On average, an identity score of one was maintained up to (and including) a coverage limit of 22 reads. Below this level of coverage, the accuracy of the identities of the compared consensus sequences decreased i.e. consensus sequences made with a depth of 22x reads were identical to the consensus. Sequences created with less than 22x coverage depth were not identical, and therefore considered less accurate.

Analytical sensitivity of WGS protocol: consensus sequence was obtained to 1x10^7 virus genome copies The protocol workflow was optimised and tested using a single FMDV O/UKG/35/2001 isolate. Initially, the sensitivity of the protocol in the presence of gDNA (i.e. no rDNase1 treatment) was tested against viral dilutions spanning 1x10^8, 1x10^7 and 1x10^6 RNA copies/μl. The total number of Illumina reads in all five samples ranged between 2.5x10^6 and 1.2x10^6 (Table 2.3). Consensus genome sequences (8176 nucleotides in length) were created from alignments of these reads at each dilution. A decreasing percentage of viral reads correlated with decreasing viral load (17.94%, 14.41%, 1.83%, 0.05% and 0.01% respectively). Consensus sequences were found to be identical in all cases both between individual samples and the reference sequence (data not shown). For this isolate, whole genome sequence was attained (excluding the 5' termini) for 1x10^8 and 1x10^7 genomes copies/μl, however, below this level, coverage was incomplete. Coverage was increased in regions adjacent to primer binding sites and was lowest in the S-fragment (genome positions nt 1376), notably in regions immediately adjacent to the poly(C) tract. The 3' genomic termini were obtained in the

cell culture neat virus sample (1x10^8 copies/μl) with only 2 bases missing at the 5' termini. In order to gain accurate consensus our analysis shows that for type O we needed a minimum viral read depth of 22. By this criterion accurate consensus sequences were generated for >98.1% of the genome, down to 1x10^7 copies/μl. Below this threshold (i.e. <1x10_7 copies/μl) we observed a rapid drop-off in the coverage depth of genome sequences with average coverage across the genome dropping from 639 (1x10^7) to 18 (1x10^6) (Table 2.3). Furthermore both genomic termini, notably the 5' end, were also lost with decreasing viral load.

gDNA depletion increases proportion of reads attributed to virus genome We investigated the impact of genomic DNA (gDNA) depletion by rDNase1 treatment upon the final library complexity. Removal of gDNA was confirmed by Qubit measurement before and after treatment (data not shown). Although the majority of DNA in the sample was eliminated it should be noted that some residual DNA remained in the sample. Samples that had not been subjected to rDNase1 treatment contained increased total number of reads, compared to samples that had been treated with rDNase1 (average: 1.9x10^6 vs. 3.8x10^5 reads, respectively). However, a higher percentage of reads aligned with the reference template for gDNA depleted samples compared to untreated samples (Table 2.3).

Validation of protocol on field samples of FMDV and reproducibility Five field samples submitted to the UK FMD National Reference Laboratory (Pirbright, UK) during the UK 2001 outbreak were tested using the sequencing protocol for UKG specific viruses as described above. Virus load in all samples was quantified by real-time RT-qPCR (Table 2.3). Four of five samples (O/UKG/1450/2001, O/UKG/1558/2001, O/UKG/1734/2001 and O/UKG/14597/2001) contained between 1.8x10^8-5.0x10^8 copies /μl. The remaining sample (O/UKG/4998/2001) was of lower viral loads with 1.01x10^7 copies/μl, respectively. The number of viral reads per sample varied between 1x10^6 (sample O/UKG/1450/2001) and 1x10^4 (O/UKG/4998/2001), potentially reflecting differences in viral load. Reads were trimmed and aligned to a reference sequence FMDV O/UKG/35/2001 (AJ539141). All samples exhibited increased coverage at primer specific sites (Fig. 2.3) and decreased coverage at the sites adjacent to the FMDV poly(C) tract and at the 5 termini of the S fragment. Samples with viral load >1x10^8 copies/l exhibited >69% of reads aligning to the reference template. The sample with the lowest viral load, O/UKG/4998/2001, resulted in 67.5% of reads aligning to the template. Complete genome sequences (excluding genomic termini) were obtained for all samples. Isolate O/UKG/1450/2001, which exhibited the highest viral load and total numbers of reads, generated a coverage depth >22 across 99.72% of the genome.

Sample ID	Serotype	Dnase Treatment	Viral Load (cp/μl)	Total No. Reads	Total Viral Reads	Percentage Viral Reads	Mean Coverage Across Genome	Percentage Consensus > Depth 22
UKG/35/2001	FMDV-O	N	4.47 x10^8	1.21x10^6	2.17x10^5	17.94	3965	99.28
UKG/35/2001	FMDV-O	N	1.65 x10^8	1.77x10^6	2.55x10^5	14.41	4641	99.3
UKG/35/2001	FMDV-O	N	3.98 x10^7	1.92x10^6	3.51x10^4	1.83	639	98.12
UKG/35/2001	FMDV-O	N	7.94 x10^6	2.08x10^6	1x10^3	0.05	18	38.35
UKG/35/2001	FMDV-O	N	1.35 x10^6	2.47x10^6	1.75x10^2	0.01	3	0
UKG/35/2001	FMDV-O	Y	4.47 x10^8	4.63x10^5	1.19x10^5	25.83	2178	99.36
UKG/35/2001	FMDV-O	Y	1.65 x10^8	1.76x10^5	4.11x10^4	23.37	743	98.29
UKG/35/2001	FMDV-O	Y	3.98 x10^7	3.29x10^5	8.29x10^3	2.52	149	93.71
UKG/35/2001	FMDV-O	Y	7.94 x10^6	4.62x10^5	1.07x10^3	0.23	19	35.71
UKG/35/2001	FMDV-O	Y	1.35 x10^6	3.73x10^5	1.11x10^2	0.03	2	0
UKG/1734/2001	FMDV-O	Y	2.89 x10^8	5.14 x10^5	4.12 x10^5	80.12	6961	99.46
UKG/1450/2001	FMDV-O	Y	4.95 x10^8	1.23 x10^6	1.10 x10^6	88.97	18362	99.72
UKG/14597/2001	FMDV-O	Y	1.77 x10^8	2.94 x10^5	2.03 x10^5	69.02	3557	97.67
UKG/1558/2001	FMDV-O	Y	4.39 x10^8	6.11 x10^5	5.27 x10^5	86.29	9391	99.68
UKG/4998/2001	FMDV-O	Y	1.01 x10^7	2.97 x10^4	2.01 x10^4	67.489	352	80.55
TUR/11/2013	FMDV-O	Y	2.22 x10^9	1.29 x10^6	8.22 x10^5	63.92	14848	99.57
TUR/12/2013	FMDV-A	Y	7.06 x10^8	1.18 x10^6	5.51 x10^5	46.49	10011	-
KEN/1/2004	FMDV-C	Y	4.41 x10^8	1.17 x10^6	4.61 x10^5	39.45	8049	-
TUR/13/2013	FMDV-Asia 1	Y	2.03 x10^9	1.69 x10^6	9.04 x10^5	53.61	10241	-
TAN/22/2012	FMDV-SAT 1	Y	1.14 x10^9	1.43 x10^6	7.26 x10^5	50.9	13185	-
TAN/5/2012	FMDV-SAT 2	Y	1.35 x10^9	1.18 x10^6	5.35 x10^5	45.48	9724	-
ZIM/6/91	FMDV-SAT 3	Y	1.47 x10^9	2.70 x10^6	1.36 x10^5	50.21	2453	-
VR-129B	EMCV-1	Y	-	2.63 x10^6	2.12 x10^6	80.34	31208	-
D1305-03	ERAV-1	Y	-	3.78 x10^4	2.68 x10^4	70.98	409	-
B1-34	VESV-B34	Y	-	4.77 x10^5	6.84 x10^4	14.34	1112	-
ISR/2/2013	FMDV-O	Y	4.50 x10^6	16 x10^4	1.05 x10^3	6.53	18	-

TABLE 2.3: **Library complexity of all samples run whilst optimising the protocol for whole genome sequencing** N = no; Y = yes; cp = copies. Different factors of library complexity including total number of reads, number of viral reads, coverage and mean coverage depth across the genome (percentage consensus depth indicates areas in which depth is over 22).

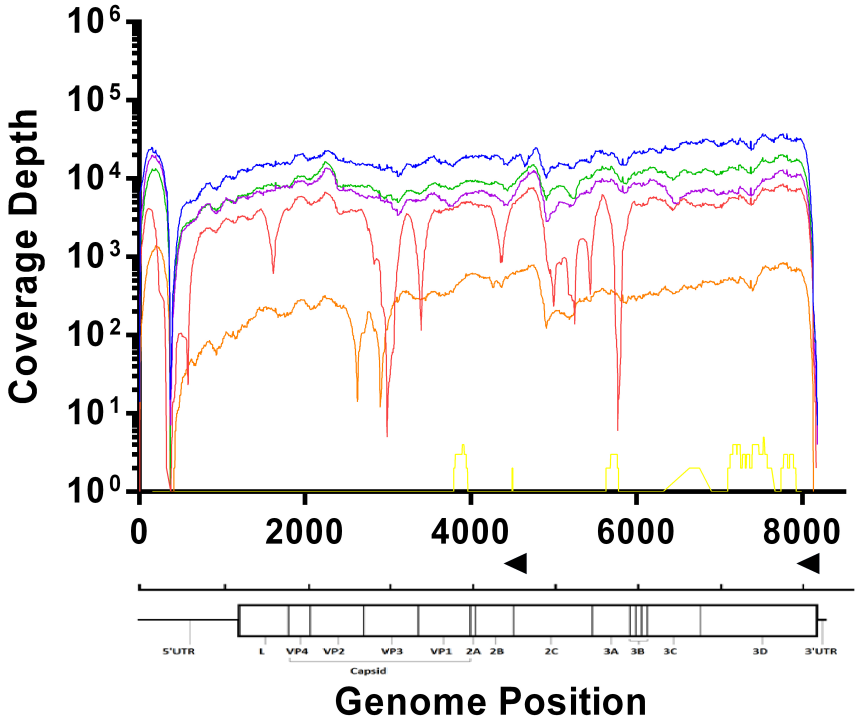


FIGURE 2.3: **Application of protocol to field isolates from 2001**. Coverage of between 100010,000x was achieved for 4/6 UKG 2011 isolates (O/UKG/1450/2001 (blue), O/UKG/1558/2001 (green), O/UKG/1734/2001 (purple) and O/UKG/14597/2001 (red)) with a drop in coverage at the poly(C) tract (375 bp position). O/UKG/4998/2001 (orange) showed lower coverage of between 10-100x. Primer locations are shown as black arrowheads above the genome illustration.

For the five samples that generated a whole genome sequence, the coverage across the L fragment was even, peaking in regions of reverse transcription primer binding (Fig. 2.3). All genome sequences have been submitted to GenBank (KM257061-KM257065). To evaluate reproducibility, one isolate (O/UKG/35/2001) was sequenced 15 separate times. Analysis was completed on each of these 15 repeats and no changes in the consensus sequence produced were observed.

Application to cell culture negative FMDV A diagnostic virus sample O/ISR/2 /2013, submitted to the WRLFMD in 2013, was sequenced using the whole genome sequencing protocol. The virus could not be isolated in cell culture, but FMDV RNA was detected with diagnostic real-time reverse transcription-quantitative PCR (RT-qPCR) and quantified as $4.5\text{x}10^6$ copies/μl (Table 2.3). The majority of the genome sequence was generated [(94.10%), with an average coverage depth of 18] with several gaps evident across the genome length (Fig. 2.4).

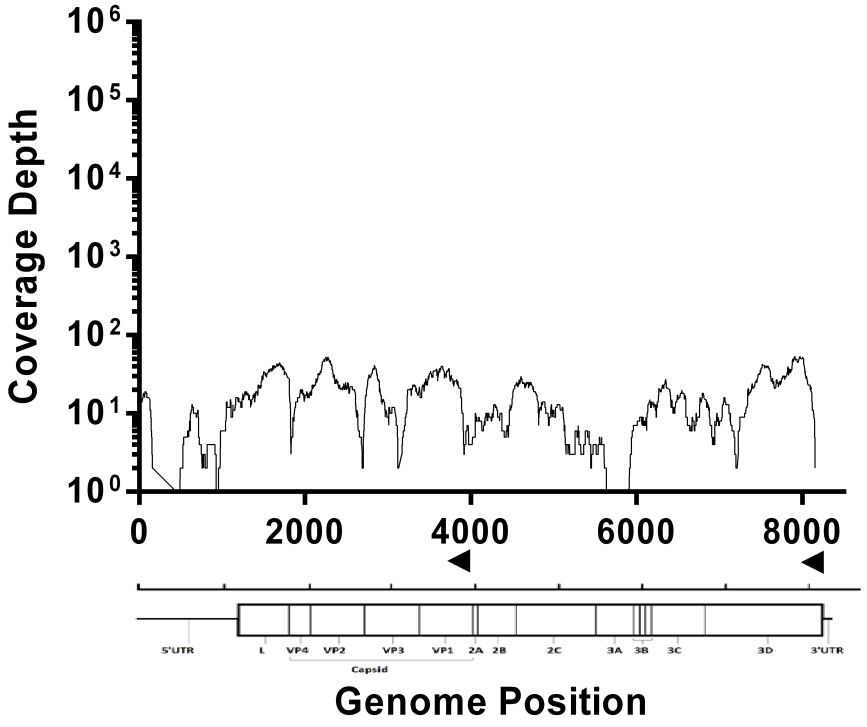


FIGURE 2.4: **Genome coverage profile for FMDV/O/ISR/2/2013** The Israel 2013 isolate of FMDV O was negative when tested in cell culture in IB-RS-2 and BTy cells. This protocol provided coverage of above 10x for the majority of the genome although full genome consensus was not acquired. The expected dip in coverage at the poly(C) was observed. Primer locations are shown as black arrowheads above the genome illustration.

Pan-FMDV application of WGS protocol After validation with FMDV UKG field samples the protocol was used to determine whole genome sequences for a panel of RNA viruses representing each of the seven FMDV serotypes (Fig. 2.5). In order to optimise the protocol we replaced the type O specific primer 4926R with a pan-FMDV primer NK-72 designed to bind a region conserved between all seven FMDV serotypes (Table 2.2. The panel had a viral load >1x10^8 copies/μl. De-novo assemblies were completed to provide a consensus against which all reads were aligned. All viruses gave similar depth of coverage (approx. 1x10^4) and exhibited comparable library complexity with the exception of SAT 3 whose depth of coverage was reduced (average coverage: 1x10^3) (Table 2.3). The 5' genomic termini was also missing from all panel viruses ranging from 9 bases of A and Asia1 to 15, 17, 22 and 27 for SAT 2, SAT 1, SAT 3 and O respectively (accession numbers KM268895-901).

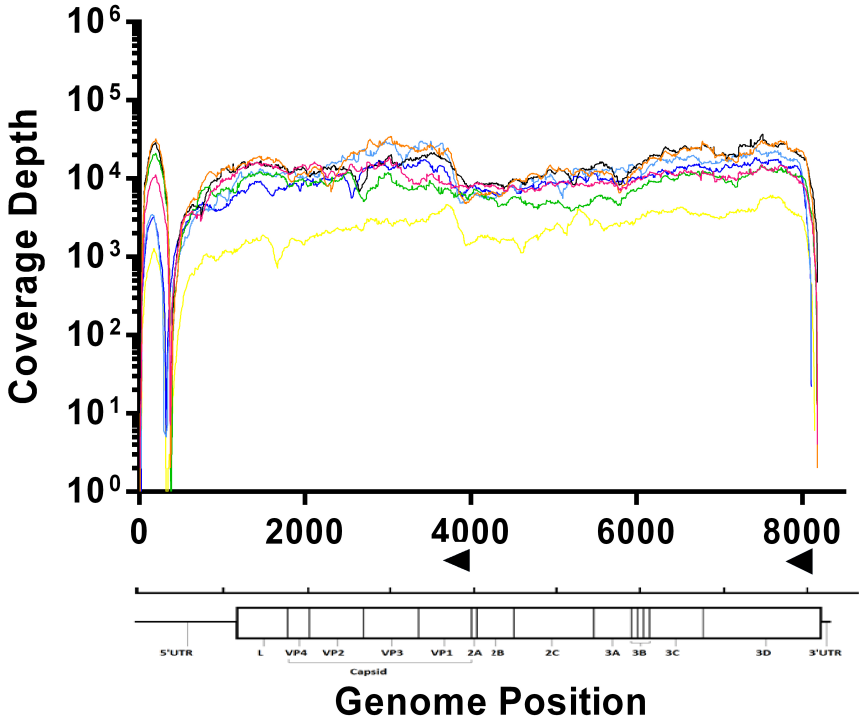


FIGURE 2.5: **Genome coverage profiles for FMDV serotype panel** Sequence data coverage at each position along the genome is shown for serotype O (black), A (pink), Asia 1 (orange), C (green), SAT 1 (light blue), SAT 2 (blue), and SAT 3 (yellow). The majority of the coverage is above 1000x. In all viruses tested, a poly(C) tract within the FMDV genome at 375 bp was associated in a reduction in coverage. The coverage depth observed for SAT 3 was lower than other serotypes. Primer locations are shown as black arrowheads above the genome illustration.

Application to non-FMDV RNA viruses In order to demonstrate the suitability of this method to attain whole genome sequence from other poly(A) tailed RNA viruses, we tested the protocol upon three different viruses including encephalomyocarditis virus 1 (EMCV-1) equine rhinitis A virus 1 (ERAV-1) and vesicular exanthema of swine virus B34 (VESV-B34)(Fig. 2.6). For all three viruses, first-strand cDNA synthesis was performed using the 3 oligo-dT primer Rev 6 and sequence-specific primers replacing the pan-FMDV specific NK72 (Table 2.2). The complete genome sequence, apart from the poly(C) tract was determined for EMCV-1 ATCC VR-129B (KM269482). The complete genome sequence, apart from 100+ nt at the 5' end of the genome was determined for ERAV-1 D1305-03 (KM269483). Similarly, the majority of the calicivirus VESV-B34 genome was determined apart from six nt at the 5' end of the genome (KM269481).

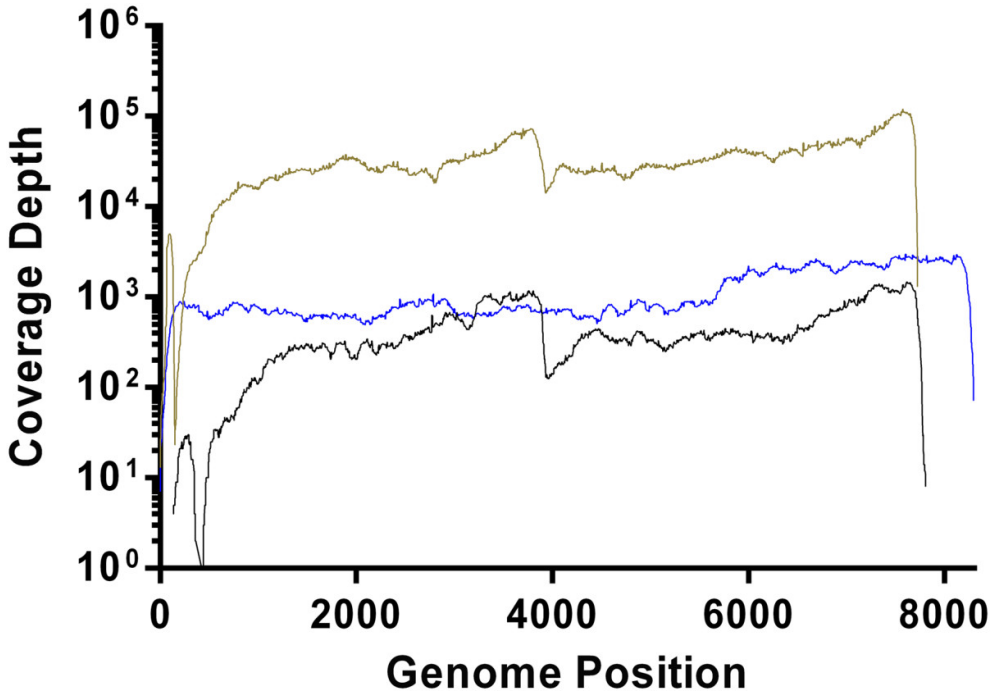


FIGURE 2.6: **Genome coverage profiles for three non-FMDV panel of viruses.** Coverage of 10,000 was achieved for the majority of the EMCV-1 genome (olive). Peaks in coverage can be observed at the location of sequence specific primers used in the RT reaction (4000 bp and 8000 bp). A dip in coverage was evident at the poly(C) tract. The ERAV-1 genome showed between 10x and 100x coverage with visible peaks in coverage at the specific primer sites (4000 bp and 8000 bp) (black). Approximately 100x coverage of the majority of the VESV-B34 genome was achieved (blue).

2.2.5 Discussion

We have described a novel sample preparation method incorporating minimal amplification for the accurate sequencing of RNA viruses to a consensus level, using an Illumina MiSeq. This protocol is an affordable and reproducible method to generate whole genome sequences for FMDV and other RNA viruses, which could be adapted to routine high-throughput diagnostic laboratory workflows. The protocol was validated using FMDV type O (Fig. 2.3 and shown to be applicable to all other serotypes of FMDV (types A, C, Asia 1, SAT 1, SAT 2 and SAT 3) (Fig. 2.5) as well as other picornaviruses (EMCV-1 and ERAV-1) and a calicivirus (VESV-B34) (Fig. 2.6).

We have shown that the protocol is able to produce whole genome sequences from samples with viral loads as low as $1x10^7$ virus RNA copies per μl. Further validation was performed with five samples submitted during the UK 2001 FMDV outbreak. The generation of five genomes from these samples, without PCR amplification or virus culture, demonstrated the potential for this method to investigate larger outbreak sample sets in a high-throughput, diagnostic setting, such as the UK 2001 FMDV outbreak.

PCR processes have previously been shown to be error prone [173] and thus eradication of this step has the opportunity to improve the quality of the data. Our protocol differs from previous studies in the literature through inclusion of sequence specific primers, as opposed to random priming at the first strand cDNA stage [207, 208]. This decision was made with the intent of maximising coverage, across the whole genome, specifically for FMDV; although it is possible that primer induced bias could be introduced into sequences through use of sequence specific primers.

We have also demonstrated the effectiveness of adapting this method for WGS of other RNA viruses (Fig. 2.6). We foresee this protocol being practicable for unknown positive sense polyadenylated viruses through use of random primers and, where appropriate, an oligo-dT primer.

The specificity previously provided by PCR has been replaced with reduction of host DNA and the optional use of specific primers in the reverse transcription reaction. Instead of enriching viral RNA we depleted host genomic DNA. We did not target ribosomal RNA in order to keep reagent costs low thus maintaining the suitability of the protocol for high-throughput sample processing. The method described here was capable of generating whole genome sequences of FMDV field isolates with a coverage depth of up to $1x10^4$ (data not shown) that was considered sufficient for the study of minority variants [192], with only a minimal amount of PCR at the library preparation stage. This PCR amplification involved 10 cycles of amplification by a hi-fidelity DNA polymerase, thereby posing minimal risk to biasing the final sequence data [209].

It was evident that in genome sequences generated using this protocol the genomic termini and poly(C) tract exhibited lower coverage depths. The 5 genomic termini were always under-represented within the genomes. This was particularly evident in samples of decreased viral load suggesting that increasing the input RNA of such samples could improve this coverage. Additionally homopolymeric regions, such as the long poly(C) tract of FMDV, have been demonstrated here to cause significant decreases in coverage. With Sanger sequencing, large parts of the genome are often missing or primer derived. For example, twenty seven to fifty nucleotides of the full genome sequences obtained by Sanger sequencing described by Valdazo-Gonzalez et al., [210–212] were primer derived (from the forward and reverse primers to amplify 5' and 3' termini of both the S and the L fragment) and thus the method described here offers a notable improvement on the resolution of these regions. As previously stated, a minimum read depth of coverage required to create an accurate consensus for a type O sequence was on average 22 (Fig. 2.2). Even after implementation of this criterion, consensus sequences were generated with a depth of >22, at more than 80.6% genome positions. This was observed in the

5 UKG field isolates tested and >99.6% for type O virus tested as part of the panel of serotypes (Table 1.).

Such advances in WGS will likely impact fields such as virus evolution, diagnosis, and generation of high/low pathogenicity variants. We have already shown this method can be advantageous in a diagnostic setting with the successful sequencing of 94.1% of the genome of a culture negative field isolate. FMDV reads were successfully identified although the resulting profile exhibits several gaps in the genome sequence suggesting that the RNA was in fact degraded - an observation potentially explaining the inability for this virus to grow successfully in cell culture. For this protocol to be fully functional within a diagnostic environment, it remains to be confirmed whether it is able to correctly identify all viruses or serotypes within mixed samples.

2.2.6 Conclusion

This paper outlines the development of a high-throughput protocol for the generation of whole genome sequences of all seven serotypes of FMDV. With minimal changes applied to priming in the first strand synthesis stage such a strategy can be tailored to other RNA viruses. The application of NGS to virology will prove invaluable to the fields of molecular epidemiology and phylogenetic outbreak tracing. This paper describes a fast, robust and affordable protocol, which is essential to realise this potential.

2.3 Thesis-specific optimisation of the published method

The published method described above [170] was designed to allow high throughput of samples in a diagnostic environment. Most diagnostic samples are amplified in cell culture to provide a large amount of high titre virus. Therefore, minimising loss of sample through processing was not required. Improvements on the molecular method for sample preparation, included below, focused on improving sample yield and ensuring that a comparable quantity of each sample was reliably processed.

2.3.1 Increasing sample yield

2.3.1.1 RNA extraction

In the published protocol, the RNeasy kit (Qiagen) was used to extract RNA from cell lysate. However, numerous studies have found the RNA yield of the TRIzol extraction to

be superior to that of RNeasy when extracted from a variety of sample types [207, 213–215]. Therefore, to increase RNA yield TRIzol was used as an alternative to RNeasy. The time taken to complete a TRIzol extraction is approximately 4x longer than the RNeasy kit. In an attempt to achieve high yield and efficient use of time a kit that extracts from samples in TRIzol was judged as the best methodological start point. DirectZol (Zymo Research, cat. R2072) allows the input of samples in TRIzol and completes the clean up using a column based chemistry. A comparison was made between TRIzol extracted RNA and Direct Zol extracted RNA from the same cell lysate. Extractions were completed as per manufacturers instructions.

	TRIzol	DirectZol
A260/280	1.93	2.15
A260/230	0.96	1.65
ng/μl	40.7	34.0
Time (mins)	60	15

TABLE 2.4: **TRIzol and DirectZol extraction comparison: RNA yield and purity** One cell lysate sample was split in two and the RNA extracted using TRIzol and DirectZol. Nanodrop was used to compare the quantity and purity of extracted RNA.

TRIzol yielded 25% more RNA than DirectZol. However the purity of the DirectZol sample was greater (Table 2.4). For pure RNA a 260/280 ratio of approximately 2 would be expected and a 260/230 ratio of 2-2.2. A low 260/280 ratio suggests a containment such as phenol or residual protein. This is further confirmed by a shifted trough on the Nano drop trace (Fig. 2.7).

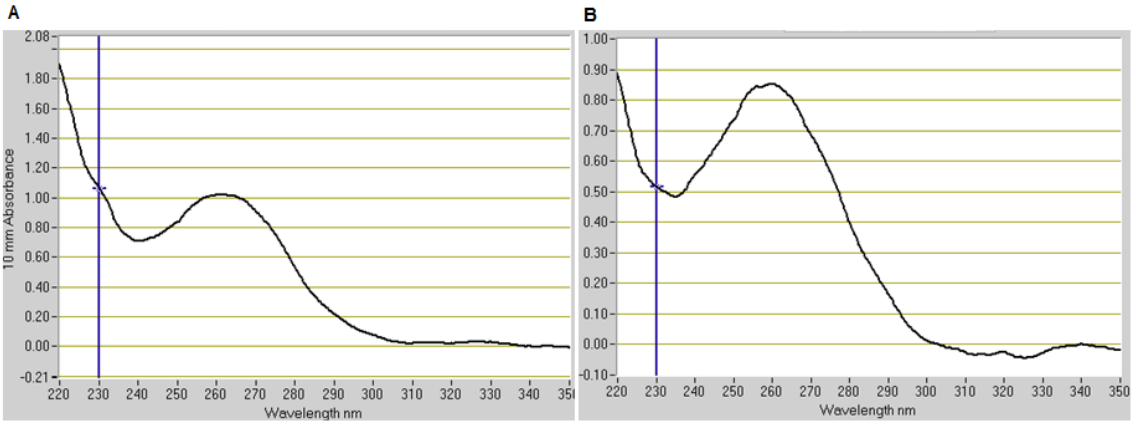


FIGURE 2.7: **TRIzol compared to DirectZol: TRIzol extraction produces higher yield but lower purity** A nanodrop trace of extracted RNA extracted using A) TRIzol and B) DirectZol. Absorbance (y-axis) compared to wavelength (x-axis) is shown.

This can be indicative of phenol contamination [216]. Phenol is capable of degrading the tagmentation enzyme in the Nextera XT preparation kit (Illumina) and thus phenol contamination can have a negative affect on sequencing success [217]. From this, Direct-Zol was deemed the best extraction methodology as it offered acceptable RNA yields of high purity in a time efficient manner.

2.3.1.2 DNAse Treatment

Following RNA extraction, samples were DNAse treated as stated in the published method [170]. If the quantity of RNA in the sample was above the recommended input for the DNA-free DNA removal kit (ThermoFisher, cat. AM1906) samples were diluted and split into several reactions and DNase treated separately. This allowed for the transference of all the original sample rather than only the quantity that was present in one reaction.

2.3.1.3 Ethanol Precipitation

Having split the sample into several DNAse reactions, an ethanol precipitation step was included to consolidate the output of the DNAse reactions into the smallest volume possible that provided the correct concentration for the subsequent RT step. RNA from the DNase reactions was pooled into quantities of up to 400μl, the maximum volume to complete an ethanol precipitation in at 1.5mL eppendorf tube. If several precipitations were completed for one sample all were re-suspended in the same 10μl of nuclease free water (nfH_2O).

2.3.1.4 Reverse Transcription

The optimum input for the reverse transcription reaction using the OneStep RT-PCR kit from Qiagen (cat. 210210) is 1pg-5μg. If the output of the ethanol precipitation was a higher concentration than required for the RT step it was diluted and several RT reactions preformed. Again this allowed for all of the original sample to be processed. The reverse transcription method from Acevedo *et al*'s 2014 paper was used [157] to allow for direct comparison between the CircSeq method and the method optimised from Logan *et al* [170].

2.3.1.5 Second strand synthesis

The second strand synthesis reaction was completed as described in Acevedo *et al*'s 2014 paper [157]. As before, multiple reactions were completed where appropriate to allow for all the sample to be used.

2.3.1.6 Clean-up of double-stranded DNA

In the published method clean up of double-stranded DNA was completed using GFX clean-up kit (GE Lifesciences, cat. 28-9034-70). However, the loss of material during this clean-up process was extensive. To ensure high yield and high purity this step was instead replaced with a phenol-chloroform-isoamyl alcohol (PCI) extraction.

2.3.2 Resolution of this method

Next generation sequencing methodologies and the sample preparation steps utilize error prone enzymes. Therefore distinguishing low frequency variants from process-derived errors in sequencing data can be challenging. By taking into account the error rate of each enzyme used in the sequencing process it is possible to consider above what level a change can be confidently considered a mutation rather than a processing error. This is considered the resolution of the method, the limit of detection of low level variants.

Error rates of enzymes were provided by manufacturers on request and are expressed as frequency of error incorporated, for example the error rate for the reverse transcription enzyme is $\frac{1}{22222}$ and therefore one error would be expected every 22,222 bases processed. The error for the second strand sythesis enzyme is $\frac{1}{111111}$. The error rate for the PCR steps involved in the NexteraXT is $\frac{1}{2500000}$. The error rate for the MiSeq, when using a cutoff of Q30, is $\frac{1}{1000}$.

$$ErrorRate \approx \frac{1}{22222} + \left(\frac{1}{111111}/2\right) + \left(\left(\frac{1}{2500000}/2\right) * 12\right) + \frac{1}{1000} \tag{2.1}$$

The error rate for the second strand synthesis is divided by two due to the double stranded nature of the nucleic acid at this point in the protocol. The Illumina PCR error rate is divided by 2 (for both strands) but multiplied by 12 due to the 12 rounds of PCR. The approximate error rate as calculated by Equation 2.1 is 0.001052. That suggests 1/950 is an introduced error, a threshold of 0.1%. In other words, a mutation would only be distinguished from errors if it was present in the data at at least 0.1%.

There are numerous concerns with this threshold:

- This calculation represents an average situation. For a given positions with a quality score of 30 the probability of an error is 0.001. However, these are random processes and as such some positions will have more errors than predicted and some less.
- The threshold must be considered in the context of genome coverage. A threshold of 0.1 when the coverage is 10,000 would require the variant to be present at least 10 times. However with a coverage of only 1000x the variant would have to be present only once to be considered relevant.
- Errors are more common at some sites than others. GC content and secondary structure can affect enzymes and errors are more common at some positions than others. Furthermore, some mutations are more likely than others. For example the tendency for one base to mutate to another can be different. In RT or PCR reactions transitions are favoured over transversions and in Illumina A to C errors or G to T errors are more common due to the similarity in fluorescence signal produced by these bases.

Due to the complex nature of this issue many approaches have been established in an attempt to bioinformatically assess the probability of low level variants being accurately called [166]. Some sample preparation steps to allow the identification of errors include barcoding and CircSeq the latter of which was tested in this study and is described in Appendix A.

In summary, the resolution of the standard method [170] can be estimated at 0.1%. However any variants at this low level must be considered in the context of coverage and further work needs to be completed to be more confident of the accuracy of this figure.

Some optimisation steps were not used in this thesis but are outlined in Appendix B.

Chapter 3

Methods

3.1 Media and Buffers

3.1.1 Media

- **Maintenance media A**: Glasgows modified eagles medium (Sigma Aldrich, cat. G5154) supplemented with 10% foetal calf serum (Gibco, cat. 16170-78), 5% tryptose phosphate broth, 20mM glutamine (SAFC, cat. 59202C-100mL), penicillin (100 SI units/mL) and streptomycin (100g/mL)(MP, cat. 1670249).
- **Viral growth media (VGM) A**: Glasgows modified eagles medium (Sigma Aldrich, cat. G5154) supplemented with 1% foetal calf serum (Gibco, cat. 16170-78), 5% tryptose phosphate broth, 20mM glutamine, penicillin (100 SI units/mL) and streptomycin (100g/mL)(MP, cat. 1670249).
- **Maintenance media B**: Ham's F12 nutrient mixture (Sigma Aldrich, cat. N4888-500mL) supplemented with 10% foetal calf serum (Gibco, cat. 16170-78), 20mM glutamine (SAFC, cat. 59202C-100mL), penicillin (100 SI units/mL) and streptomycin (100g/mL)(MP, cat. 1670249).
- **VGM B**: Ham's F12 nutrient mixture (Sigma Aldrich, cat. N4888-500mL) supplemented with 1% foetal calf serum (Gibco, cat. 16170-78), 20mM glutamine, penicillin (100 SI units/mL) and streptomycin(100g/mL)(MP, cat. 1670249).
- **Eagles overlay**: Type 1 endotoxin free water containing 8% (w/v) sodium chloride (NaCl), 0.5%(w/v) potassium chloride (KCl), 0.25% (w/v) Magnesium sulphate ($MgSO_4$), 5.65% (w/v) glucose, 0.125% (v/v) ferric nitrate (1%), 0.157% (w/v) Sodium dihydrogen orthophosphate (NaH_2PO_4), 0.365 % (w/v) glutamine, 0.33 % (w/v) calcium chloride ($CaCl_2$), 3.438% (w/v) sodium hydrogen carbonate

($NaHCo_3$), 18.75% (v/v) phenol red (1%), 0.053% (w/v) L-arginine, 0.03% (w/v) L-cystine, 0.026% (w/v) L-histadine, 0.066 % (w/v) L-isoleucine, 0.066 % (w/v) L-leucine, 0.092% (w/v) L-lysine, 0.042% (w/v) L-phenylalanine, 0.06% (w/v) L-threonine, 0.01% (w/v) L-tryptophan, 0.045% (w/v) L-tyrosine, 0.059% (w/v) L-valine, 0.019% (w/v) L-methionine, 0.005% (w/v) L-inositol, 5% (v/v) vitamin (stock concentrate) and 0.1% (v/v) sodium hydroxide.

3.1.2 Buffers

- **Phosphate buffered saline (PBS)**: Type 1 endotoxin free water containing 0.964% (w/v) Dulbecco's phosphate buffered saline (DPBS) (Sigma, D5773-50L) and 0.0132% (w/v) calcium chloride.
- **Calcium and magnesium free PBS (PBSa)**: Type 1 endotoxin free water containing 0.955% (w/v) PBS'a' powder (Sigma, 56064C-50L) and 0.04% (v/v) 1M stabilised Hydrochloric Acid.
- **Tryptose phosphate broth (TPB)**: Deionised water containing 2.95% (w/v) tryptose phosphate broth. Provided by the central services unit (CSU) at the Pirbright Insitute (TPI)
- **4% paraformaldahyde**: PBS (as previously described) containing 4% (w/v) paraformaldahyde powder (Sigma-Aldrich, cat. P6148-1kg). pH adjusted (pH 6.9) and filtered.
- **Phosphate buffered saline containing bovine serum albumin (PBS-BSA)**:PBS (as previously described) containing 0.5% (w/v) bovine serum albumin (Sigma-Aldrich, cat. 05479-50g) and 0.05% (w/v) sodium azide (Sigma-Aldrich, cat. 438456-5g)
- **Tris/Borate/EDTA (TBE)**: Deionised water containing 10% (v/v) 10x Ultra-Pure TBE Buffer (Invitrogen, cat. 15581044)
- **0.1% Triton**: PBS containing 0.1% (v/v) Triton X-100 (Sigma-Aldrich, cat. T9284-100mL)

3.2 Cell lines used

- **Baby hamster kidney cells (BHK) 21** were cultivated in maintenance media A. They were received from the CSU at TPI.

- **Bovine foetal aortal cells (BFA)** were cultivated in maintenance media B. Originally purchased from the European collection of Authenticated Cell cultures (ECACC). Received on the 14/05/12 at passage 14.
- **Bovine thyroid cells (BTY)** were cultivated in maintenance media A. This is a primary cell line prepared by CSU at TPI.

3.3 Antibodies and stains

- **2C2**: Monocolonal antibody 2C2 which recognises FMDV 3A was a kind gift from E. Brocchi (IZS, Brescia, Italy).
- **Alexa fluro-conjugated secondary antibody**: Alexa Fluor-conjugated secondary antibodies were acquired from Invitrogen (cat. A11029).
- **TO-PRO®-3**:The nuclear and chromosome counterstain TO-PRO®-3 was attained from ThermoFisher Scientific (cat. T3605).
- **Crystal violet stain**: PBS containing 0.5% (w/v) crystal violet, 10% (v/v) ethanol (100%), 13.88% (v/v) and formaldahyde (36%).
- **Methylblue Stain**: PBS containing 3.7 % (v/v) formaldahyde, 0.1% (w/v) methylblue and 10% (v/v) 100% ethanol.

3.4 Primers and Probes

Primers and probes used within this thesis are outlined within Table 3.1.

3.5 Protocols

3.5.1 Cell Passage

Media was removed from the flask and cells were washed with once with PBSa. Trypsin-EDTA (0.25%) (ThermoFisher, cat.25300-54) was added (2, 3 or 5 mL for T25, T75 and T175 flask respectively) and the flask was incubated at 37°C fo 3 minutes. Detached cells were washed from the flask with 10mL of appropriate maintenance media and the entire contents of the flask transferred to a universal. Cells were spun 1000rpm for three minutes and supernatant removed. Cells were re-suspended in 1mL of maintenance media using a 1000μL pippette. This 1mL of cell suspension was further diluted with

Purpose	Primer Name	Primer Sequence	Chapter
RT-qPCR	Callahan 3DF [192]	ACTGGGTTTTACAAACCTGTGA	2,7
	Callahan 3DR [192]	GCGAGTCCTGCCACGGAA	2,7
	Callahan 3DP [192]	TCCTTTGCACGCCGTGGGAC	2,7
First-strand synthesis and sequencing	UKFMD Rev6 [115]	GGCGGCCGCTTTTTTTTTTTTTT	2
	NK72 [193]	GAAGGGCCCAGGGTTGGACTC	2
	UKFMD UKG 4926R	AAGTCCTTCCCGTCGGGGT	2
	EMC-2B65R [194]	TCGGCAGTAGGGTTTGAG	2
	ERAV-2A22R [195]	GGGTTGCTCTCAACATCTCCAGCCAATTT	2
	Vesi-3D1R	CKNGTNGGYTTNARNCC	2
	Vesi-3D2R	TANCANCCRTCRTCNCCRTANGT	2
	Oligo Bridge	ATATGCGATCGCCCTTRCGCCCCYTTTCAATTTTTTTTTTTTTTTTTT	B
	Circ Primer 1.4	CGCGGTCCCGTGAGTCCAG	B
	Circ Primer 2.1	TCCAGGCTACAGATCACTTTACCTGC	B
	O-1C272F	TBGCRGGNCTYGCCCAGTACTAC	B
	EUR-2B52R	GACATGTCCTCCTGCATCTGGTTGAT	B

TABLE 3.1: **Primers and probes used in thesis** International Union of Pure and Applied Chemistry (IUPAC) nucleotide ambiguity codes: N: G or T or A or C; K: G or T; Y: T or C; R: G or A

9mL of maintenance media. An appropriate proportion of this re-suspension was used to re-seed a new flask(s).

3.5.2 Cloning

Cloning of GeneArt fragments into an FMDV infectious copy plasmid (ICP) was completed in Chapter 7.

A cloning strategy was designed to systematically knock out regions containing putative packaging signals (PPS) when identified by bioinformatics analysis. One region containing one quarter of the PPS was amended (PPS2, 3 and 4). Each PPS was disrupted both in structure and sequence of the loop. This was completed by editing the nucleotide sequence without affecting the amino acid sequence in order to create replication viable virus. The fragment was ordered from GeneArt (ThermoFisher Scientific).

GeneArt strings were cloned into a TOPO vector (ThermoScientific, cat. K2800J10). The backbone (TOPO) vector was linearised via restriction digest.

3.5.2.1 Restriction enzyme digest

Digests were completed in a 20μL volume reaction as per manufacturers instructions. Enzymes were supplied by New England Biolabs.

Linearised backbone visualised via gel electrophoresis and UV illumination.

Electrophoresis of DNA was completed as described below. Linearised backbone was extracted from agarose gel.

3.5.2.2 Gel extraction of DNA

Gel extractions were completed using Illustra GFX DNA and Gel Band Purification kit (GFX, cat. 28-9034-70) as per manufacturers instructions. In short, gels were visualised on a UV lightbox (365nm)(BioRad) and required bands were excised with a scalpel. Agarose bands were placed in a DNAse-free 1.5mL tube. The weight of each bands was determined. To each 10mg of gel 10μL of capture buffer type three was added. The sample was heated to 60°C for 15 to 30 minutes until the gel band was completely dissolved. Samples were mixed by inversion. Gel/capture buffer mix was added to a GFX MicroSpin column and collection tube and incubated at room temperature for 1 minute. The assembled columns were then spun at 16000xg for 30 seconds. Flow through was discarded. 500μL of wash buffer was added and the assembled columns

were spun for a further 30 seconds at 16000xg. The collection tube and flow-through was discarded. The column was placed in a new collection tube and 10-50μL of nfH_2O was added. The column was incubated at room temperature for one minute before being spun at 16000xg for 1 minute to recover the purified DNA.

3.5.2.3 Dephosphorylation with shrimp alkaline phosphatase treatment

Dephosphorylation improves ligation efficiency by preventing the vector from ligating back to itself. 7μL of linearised gel extracted backbone was combined with 0.9μL of 10x reaction buffer and 1μL of shrimp alkaline phosphatase (NEB, cat. M0371S). The reaction was incubated at 37°C for 30 minutes before being heat inactivated at 65°C for 15 minutes.

Dephosphoryalted vector backbone was ligated to the GeneArt strands.

3.5.2.4 DNA Ligation

Ligation was completed using T4 DNA Ligase (NEB, cat. M0202S) and the provided 10x DNA ligase reaction buffer. A vector:insert ration of 1:1, 1:3 and 1:7 in a total volume of 10μL were completed. Ligation were completed overnight at 4°C.

Ligated vector and insert were transformed into *E.coli* One Shot TOP10 chemically competent cells (ThermoScientifc, cat. K2800J10).

3.5.2.5 Transformation of competent cells

A 50μL aliquot of frozen competent cells (One Shot TOP10 chemically competent cells) were thawed on ice for 5 minutes. To the thawed cells 1-5μL of ligation reaction was added. The reaction was tapped gently to mix and incubated on ice for 30 minutes. Reaction was incubated at 42°C for 30 seconds and then placed back on ice. To the reaction 250μL of pre-warmed S.O.C medium was added and the vial was placed in a 37°C shaking incubator for 1 hour at 225rpm. The transformed bacteria were plated on LB/agar plates containing the appropriate antibiotic (e.g. Ampicillin (100μL/ml) or Kanomycin (50μL/ml)). Plates were inverted and incubated at 37°C overnight.

3.5.2.6 Isolation of plasmid DNA

Small scale (Mini-preps) A single colony was picked from an agar plate and used to inoculate a 5mL culture of LB medium containing the appropriate antibiotic. Cultures

were incubated in a 37°C incubator at 300rpm over night. 1mL of this culture was transferred to a 1.5mL microfuge tube and centrifuged for 14000rpm for 1 minute. The supernatant was removed and DNA was extracted using the Qiagen Miniprep Kit (cat. 27106) as per manufacturers instructions.

Large scale (Maxi-preps) A single colony was picked from an agar plate and used to inoculate a 5mL culture of LB medium containing the appropriate antibiotic. Cultures were incubated in a 37°C incubator at 300rpm over night. 200μL of this culture was diluted in 100mL of LB broth and incubated in a 37°C incubator at 300rpm over night. DNA was extracted using the Qiagen Maxiprep Kit (12162) as per manufacturers instructions.

To confirm successful inclusion of the GeneArt insert test digests and agarose gel analysis was completed.

The backbone plasmid (pT7S3) and gene art fragment (previously cloned in to a TOPO vector as described above) were digested with Xma1 and Aflii and the products run on a 1% agarose gel containing ethidium bromide (EtBR). DNA was extracted using the GFX clean-up kit (GE Healthcare Life Sciences, cat.28-9034-70). The backbone was dephosphorylated and vector and insert ligated as previously described. Ligations were transformed into competent cells (TOP10) and colonies produced were minipreped and text digested. Correct plasmids were grown up and Maxipreped.

Maxipreped cultures were confirmed with sequencing on the applied Biosystems 3730 DNA analyser using the BigDye Terminator v3.1 cycle sequencing kit (applied Biosystems) as described below.

3.5.3 Ethanol Precipitation

On ice, 0.1 volumes 3M sodium acetate (pH5.2) was added to the pooled RNA and the sample vortex mixed. 2.2 volumes of ice cold 100% ethanol was added and pipette mixed thoroughly. Samples were then incubated at -20°C for one hour. Samples were pipette mixed and spun (14,000 xg) for ten minutes at 4°C. Supernatant was remove and discarded and 500μl of ice cold 70% ethanol added. Samples were spun (14,000 xg) for five minutes at 4°C. Supernatant was remove and discarded and RNA re-suspended in 10μl of nfH_2O.

3.5.4 Fragmentation

18μL of RNA and 2μL of fragmentation reagent were combined and vortexed. Samples were then incubated at 70°C for seven and a half minutes. Samples were then immediately placed on ice and 2μL of stop solution was added.

3.5.5 Gel electrophoresis

Electrophoresis of DNA was completed using gels of 1% agarose dissolved in TBE containing 0.5μg/mL ethidium bromide. Gels were run at 100v.

3.5.6 Immunofluorescence

Cells were fixed with 4% formaldehyde solution for 40 minutes, PBS washed and permeabilised with 0.1% triton for 15 minutes. After a further PBS wash, block buffer (PBS-BSA) was added for 30 minutes. Primary antibody (2C2) was added for 1hr in block buffer. Cells were washed 5x with PBS and incubated with Alexa Fluor-conjugated secondary antibody for 45 minutes. Cells were washed a further 5x with PBS and ToPro-3 was added for 5 minutes. Stain was removed and cells washed and imaged in sterile water. Cells were imaged on the SpectraMax MiniMax 300 Imaging Cytometer. Total number of cells was calculated using ToPro-3 fluorescence (excitation 625/20nm), infected cells were counted using Alexa 488 fluorescence (excitation 460/20nm).

3.5.7 Phenol Chloroform Extraction

An equal volume of PCI was added and the sample vortexed for 2 minutes. Sample was incubated on ice for 15 minutes before being spun at 13,000rpm for 15 minutes at 4°C. The aqueous phase was collected and 0.1mL sodium acetate was added per mL. Samples were vortexed and 2x volume of chilled 100% ethanol added. Samples were mixed well and incubated at -20°C for thirty minutes. Samples were spun at 13,300rpm for 30 minutes at 4°C and the supernatant was removed. 250μl of 70% ethanol was added and the sample was spun again at 13,300rpm at 4°C for 2 minutes. The supernatant was removed and the pellet air dried before being re-suspended in 10μl of nfH_2O.

3.5.8 Plaque assays

A dilution series was created from the virus to be titred. To create 10^{-1} 100μL of neat virus was combined with 900μL VGM and the sample was pipette mixed. 100μL of 10^{-1}

was combined with 900μL VGM and sample was pipette mixed to create 10^{-2}. This was continued to 10^{-9}. Samples were maintained on ice.

Cells were prepared at 60% confluencey in a six well plate. Three six well plates were prepared per sample.

Eagles Overlay 75mL eagles overlay was supplemented with 1.5mL foetal calf serum, 1 mL of penicillin (100 SI units/mL) and streptomycin(100g/mL) and 5mL of tryptose phosphate broth. Eagles overlay was incubated at 37 degrees until use.

Indibios Agar 25mL of sterile water was combined with 0.6g indubiose A37 Agar (MP, AGA10025). Agar was was microwaved on full power for two minutes until indubiose was completely dissolved. Agar was kept warm at 42 degrees until use.

Procedure Cells were washed with PBS. 100μL of virus dilution was added to relevant wells (Fig. 3.1). Plates were incubated at 37°for 15 minutes. 75mL of eagles overlay mixture (as described above) was combined with 25mL of agar solution. 4mL of Eagles/agar mixture was added to each well. Plates were left at room temperature until agar was set. Plates were incubated at 37°C for 72 hours. 4mL of Crystal Violet stain (as described above) was added to each well. Plates were left at room temperature for at least 4 hours to fix. Excess stain and agar plugs were washed off.

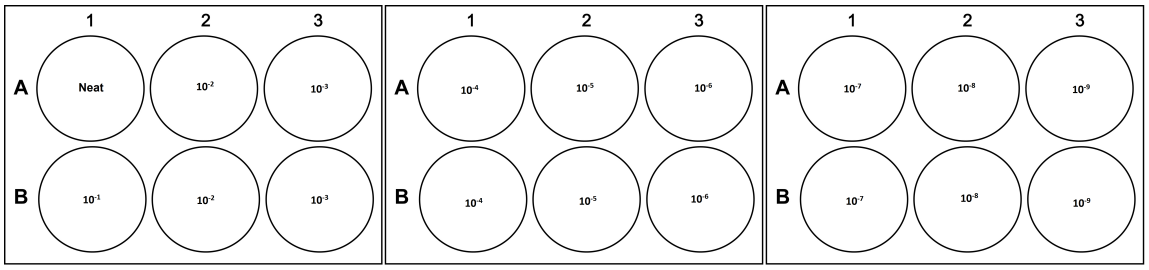


FIGURE 3.1: **Plaque assay plate set up**

Plaques were counted on the dilution that had between 50-100 plaques. Calculations were completed as follows:

If the 10^{-4} wells had 74 and 76 plaques respectively. The average would be calculated as 75 plaques. 100μL of virus was used to innoculate each well so the values are multiplied by 10 to create the PFU/mL: 75x10^5 PFU/mL or 7.5 x10^6 PFU/mL.

3.5.9 qRT-PCR

One-step qRT-PCR was completed using the SuperScript® III Platinum®One-Step qRT-PCR Kit (Invitrogen, cat. 11732030). The Callahan 3D assay was preformed [192]. Primers (3DF and 3DR) and probe (3DP) are detailed in Table 3.1. Reactions of 50μL were set up as follows. On ice, 25μL of 2x reaction mix was combined with 16μL of nfH_2O, 1μL of forward primer (10μM), 1μL reverse primer (10μM) and 1μL fluorescent probe (10μM). To this, 5μL of template was added and 1U of SuperScript®III. Samples were pipette mixed before tubes or plates were sealed and loaded onto the qPCR machine (Mx3005P qPCR System, Agilent Technologies). Samples were heated to 60°C for 30 minutes (cDNA synthesis) before being heated to 95°C for 10 minutes. Samples then underwent fifty cycles of 95°C for 15 seconds and 60 °C for 60 seconds. Data was visualised and analysed using MxPro.

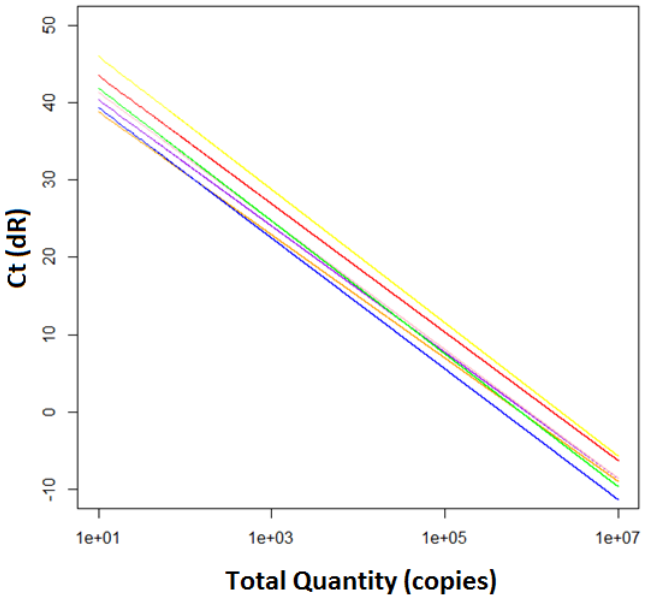


FIGURE 3.2: **qRT-PCR standard curve for each serotype** A standard curve is shown for Asia 1(red), type A(Yellow), type O (purple), type C (orange), SAT1 (pink), SAT 2 (green) and SAT3 (dark blue).

To ensure the assay was comparable for all seven serotypes of FMDV a comparison of RNA standards was made (Fig. 3.2).

3.5.10 Quantification of DNA

The concentration of DNA was determined either by Qubit quantification (as previously described) or optical density readings were taken at 260 and 280nm using a ThermoScientific NanoDrop. Distilled water or elution buffer was used as a reference.

3.5.11 SDS urea page gel

Gel running tank (Biorad, mini-Protean tetra vertical electrophoresis cell), plates and gel casting station were thoroughly cleaned with RNA zap (Thermo Fisher, cat. AM9780), rinsed with distilled water, sprayed with 70% ethanol and wiped clean with kimtech tissue. Glass plates were places in the casting station. 12.5% acrylamide denaturing gel was created by combining 25g of urea, 10mL of TBE, 15.625mL of acrylamide (40%) and nfH_2O up to 50mL. 6.5mL of 12.5% acrylamide denaturing gel (as above) was combined with 37.5μL APS and 3.75μL of TEMED in a 15mL tube. Solution was vortexed and gels poured immediately. Comb was placed and gels were left to set. When set, gels were loaded into the tank and TBE added until the wire was covered. Wells were flushed with TBE. Samples were denatured with 2x denaturing dye (Thermo Fisher, cat. R0641) for 5 minutes at 98°C. 1μL of RNA ladder (Novagen, cat. 69003) in 4μL of nfH_2O with 5μL of 2x denaturing dye was also denatured. Samples were loaded on the gel and run at 300v until the dye front ran off the gel. The gel was then removed and shaken in a solutuon of 20μL of TBE containing 2μL of SYBR gold (Thermo Fisher, S11494).

3.5.12 Sanger Sequencing

Sanger sequencing was completed on a 48 capillary sequencer (ABI3730). Samples were prepared using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, cat. 4337455). High copy or plasmid DNA was diluted to 500ng in a final volume of 10μL in a 0.2mL PCR tube and denatured at 96°C for 1 minute. For PCR products, denaturing was not required and 5-20ng of product for 500-1000bp of product was used. Components of sequencing reaction were combined on ice. A 10μL reaction was made containing 1.88μL of 5x Sequencing buffer, 0.25μL of BigDye® Terminator v3.1, 1.5μL of primer (at 1.6pmol), 5.37μL of nfH_2O and 1μL of template (diluted as described above). Reactions were placed in a thermocycler (Eppendorf) and heated to 96°C for 1 minute. Subsequently 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°for 4 minutes were completed. Ethanol precipitation was completed on samples and they were vacuum dried for 15 minutes to ensure no ethanol was present. Pellets were resuspended in 20μL of Hi-Di™Formamide (Applied Biosystems, cat. 4337455). Samples

were left to fully re-suspend in the dark for 15 minutes before being loaded onto the sequencer.

3.5.13 SpectraMax MiniMax 300 Imaging Cytometer

A 50% confluent 96 well plate of cells was counted on the MiniMax cytometer through transmitted light. This value was used to calculate the viral dilution needed to achieve the required multiplicity of infection (MOI). Media was removed from the plate and virus added. A plate sealer was applied and the plate was place in the plate reader at an incubating temperature of 37°C. Transmitted light images were captured once an hour, ever hour and the number of objects or area covered counted. Data was visualised using GraphPad prism.

3.5.14 TCID^{50}

A ten fold dilution of virus lysate was completed to 10^{-3}. A 2 fold dilution of this was then completed ten times. Virus dilutions were stored on ice. 50μL VGM A was added to each well of a 96 well plate and incubated at 37°C. A confluent T175 flask of BHK21 cells was tripsinized and re-suspended in 10mL of maintenance media A. This cell suspension was counted using a haemocytometer and diluted to 1x10^6 cells/mL. 50 μL of virus was added to each well (already containing the warmed VGM) as shown in Fig 3.3. No virus was added to the last column of wells to act as a no virus control. To each well 50μL of cell suspension was added and plates were incubated at 37°C for 72 hours.

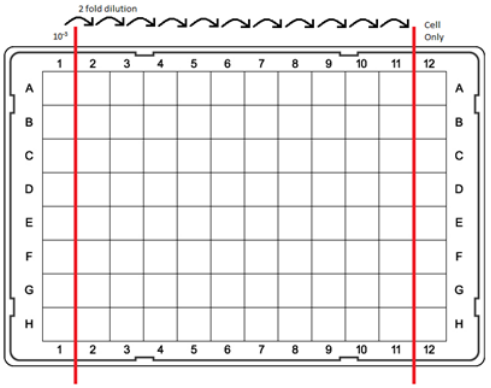


FIGURE 3.3: **TCID50 assay plate set up**

Media was removed from the plate and it was submerged in 1% citric acid. The plate was then submerged in 2% citric acid for 10 minutes. 50μ of methylblue stain was

added to each well and plate was left to stain overnight. Stain was removed and plates visually inspected with the number of cells with 100% CPE in each dilution counted and recorded.

Total number of cells with 100% CPE was counted and this value was divided by the number of wells per dilution (8). A correction factor of 0.5 was subtracted. This value was multiplied by the dilution interval (0.3) and added to the highest dilution where all 8 wells showed complete CPE. This was multiplied by 20 to produce TCID50/mL.

3.5.15 Growth of viral stocks

Bovine Thyroid Cells cells were maintained in maintenance media A at 37°C. Prior to infection with FMDV, the cell monolayer was washed with serum free medium. 2.25ml of serum free medium was added to the 0.25mL aliquot of virus (in glycerol or cell suspension). 2.5mL of this solution was added to each flask. The flasks were incubated at 37°C for 10 minutes to allow the virus to adsorb to the cell monolayer before 20mL of maintenance media A was added and the flasks were incubated overnight at 37°C. Flasks were frozen when full CPE was observed. Flask were thawed and contents transferred to a falcon tube. Sample was spun for 3 minutes at 1000rpm to clarify. Clarified supernatant was removed and aliquoted and stored at -80°C.

3.5.16 Viral Passage

T25 flasks of BHK or BFA cells were maintained in appropriate maintenance at 37°C. Prior to infection with FMDV, the cell monolayer was washed with PBS. Clarified supernatant (from the innoculum or previous passage) was combined with appropriate VGM (1:1) and 1mL of this was added to the cells. The flasks were incubated at 37°C for 10 minutes to allow the virus to adsorb to the cell monolayer before 4mL of maintenance media was added and the flasks were incubated overnight at 37°C. Flasks were frozen when full CPE was observed. Flask were thawed and contents transferred to a falcon tube. Sample was spun for 3 minutes at 1000rpm to clarify. Clarified supernatant was removed and aliquoted and stored at -80°C.

3.5.17 Virus purification

Appropriate cells were infected with clarified cell lysate and maintained at 37°C in VGM. At complete CPE, cultures were freeze-thawed and pelleted by low speed centrifugation. The supernatant was precipitated by the addition of sufficient ammonium sulphate to

produce a 50% saturated solution. Once all ammonium sulphate was dissolved and in suspension, the precipitate was pelleted at 600 rpm at 4°C for 1 hour. After centrifugation, the pellet was resuspended in PBS (pH 7.4) and Igepal added to a 1% final volume. Virus was then purified by initial sedimentation through a 30% sucrose cushion (w/v in PBS), by centrifugation at 28,000 rpm at 12°C for 2.5 hrs. The subsequent pellet was resuspended, as before, and further purification achieved by sedimentation through a 15-45% sucrose gradient (w/v in PBS) by centrifugation at 28,000 rpm at 12°C for 2.5 hrs. The virus was located by measuring the 260 nm absorbance of gradient fractions.

3.5.18 Virus recovery

Plasmids were transfected into 80% confluent BHK cells using mRNA Trans-IT as per manufacturers instructions (Mirus, cat. MIR 2225).

3.6 Analysis Methods

3.6.1 Statistical Analysis

Statistical analysis and visualisation of data in a graphical format was completed using GraphPad Prism6 or R i386 3.2.5.

3.6.2 Bioinformatics analysis pipeline

The bioinformatics pipeline is outlined in Chapter 2[170]. In short, MiSeq reads were downloaded from Illumina Basespace onto a Linux server. Samples are unzipped using gunzip:

```
$ gunzip Read1.fastq.gz
```

As outlined previously [170] reads were trimmed using Sickle using a quality score of 30:

```
$ sickle pe -f Read1.fastq -r Read2.fastq -t sanger -o Read1sk.fastq -p Read2sk
  .fastq -s GL_sing.fastq -q 30 -l 100 -n
```

with -t representing the quality score format, -q representing the quality cut off, -l representing the length the read must be to be accepted and -n deeming no reads with n in will be accepted.

Trimmed reads were then used in *De novo* assemblies with Velvet (using Velvet optimiser to determine Kmer length). An excess of reads input into Velvet optimiser resulted in a

limited number of contigs. Therefore, only a subsample of reads were used in this step. Reads were randomly sampled using SubSampleFastq. The program was cloned from github;

```
$ git clone https://github.com/dylanstorey/SubSampleFastq
```

and used to create a random sample of 20000 of the original reads;

```
$ RandomSubFq Read1sk.fastq Read2sk.fastq -w 20000
```

The two random samples from read one and read two were then shuffled together using a perl script provided with Velvet.

```
$ shuffleSequences_fastq.pl Read1skran.fastq Read2skran.fastq shuffled.fastq
```

Velvet optimiser was then used to run Velvet. Velvet optimiser is a program that runs Velvet with a variety of different Kmer lengths to find the optimum length for contig assembly.

```
$ VelvetOptimiser.pl  -s 11 -e 41 -x 2 -f "-shortPaired -fastq shuffled.fastq"
  -d "dir4_shuffled" -o "-cov_cutoff 90 -min_contig_lgth 1000"
```

In this command -s represents the minimum kmer length to be tested, -e represents the maximum kmer length to be tested, -x represents the interval of kmer increase, -f is the velvet command (this shows read type, file type and actual file to be used) and -d is the output location. The commands after -o are options for velvet such as the coverage required to accept the contig and the minimum contig length. When trying to construct the S-fragment of FMDV the latter option was reduced. Contigs produce were entered in BLAST searches to confirm their origin was FMDV and to determine their orientation. Contigs that represented host were discarded. The orientated, FMDV contigs were built into a reference genome using BioEdit. This reference was then used in an alignment of all sickle trimmed reads. Alignments were completed using Bowtie2:

```
$ bowtie2-align -x ref_index -Read1sk.fastq,Read2sk.fastq -S Alignment.sam
```

Samtools was used to convert this SAM file to a BAM file:

```
$ samtools view -bS Alignment.sam > Alignment.bam
```

and subsequently sort it:

```
$ samtools sort Alignment.bam Alignment.sorted
```

3.6.3 Entropy Calculations

The BAM file created from the alignment pipeline was used to complete entropy calculations. The Samtools mpileup command was used to make an mpileup file;

```
$ samtools mpileup -d2000000 -f ref.fas file.sorted.bam > file_pileup.txt
```

where -d is the max depth and -f represents the reference sequence. This mpilup file was then counted by a script written by Rocio Enriquez-Gasca at the genome analysis centre (TGAC). This script counts the number of each base at at each positions on the genome. A shell script (nt count.sh, Appendix D) runs an R script (eleano.r, AppendixD) that counts the bases at each positions and outputs them in a chart formatted text file.

```
$ sh nt_count.sh file_pileup.txt
```

A python script was written to complete Shannon's entropy calculations on this text file (EntropyCalc.py, Appendix D).

```
$ python EntropyCalc.py inputfile.txt
```

The output of this file is an entropy score for each position along the genome.

3.6.4 Haplotype Reconstruction

Haplotype reconstruction was completed using QuRE:

```
$ java . QuRe read_file reference_genome_file
```

Where read file is an input fasta file of all reads and reference genome file is a comparison genome reference. Haploptypes were aligned using BioEdit.

3.6.5 Phylogentic trees

Evolutionary analyses were conducted in MEGA6. The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value.

3.6.6 dN/dS

Shannon's entropy is a measure of how much variation present at each nucleotide position. However, not all changes in the nucleotide sequence result in amino acid changes in viral proteins. Some nucleotide changes in the coding region of the genome, particularly those in the third codon positions, can result in the same amino acid being produced and are thus referred to as silent or synonymous. Changes that results in a change in amino acid sequence are referred to as non-synonymous. Considering if nucleotide changes result in synonymous or non-synonymous changes can be a powerful tool in interpreting genetic evolution.

Information regarding synonymous and non-synonymous mutations is considered in the form of the $d\mathrm{N}/d\mathrm{S}$ ratio.

$$d\mathrm{N}/d\mathrm{N} = \frac{d_N}{d_S} \tag{3.1}$$

Where d_N is the estimated number of non-synonymous substitutions per genome position and d_S is the estimated number of synonymous substitutions per genome position. This approximation is calculated with equation 3.2 and equation 3.3 respectively.

$$d_\mathrm{N} = -\frac{3}{4} ln\left(1 - \frac{4p_N}{3}\right) \tag{3.2}$$

$$d_\mathrm{S} = -\frac{3}{4} ln\left(1 - \frac{4p_S}{3}\right) \tag{3.3}$$

Where p_N and p_S are the proportion of non-synonymous and synonymous differences respectively.

The proportion of differences are calculated with equation 3.4 and equation 3.5.

$$p_\mathrm{N} = \frac{\sum \left(\text{Number of non-synonymous substitutions}\right)}{\sum \left(\text{Number of non-synonymous sites}\right)} \tag{3.4}$$

$$p_\mathrm{S} = \frac{\sum \left(\text{Number of synonymous substitutions}\right)}{\sum \left(\text{Number of synonymous sites}\right)} \tag{3.5}$$

The number of synonymous and non-synonymous sites is calculated by considering if changing each codon of the amino acid would change the amino acid.

Codon	Synonymous	Non-synonymous
ATG	0	3
GTG	1	2
AAA	$\frac{1}{3}$	$2\frac{2}{3}$
CGT	$1\frac{1}{3}$	$1\frac{2}{3}$
TTA	$1\frac{1}{3}$	$1\frac{2}{3}$
Sum	4	11

TABLE 3.2: **An example of the number of synonymous and non-synonymous sites for amino acids.** Synonymous and non-synonymous changes were calculated for an five exemplar amino acid codons.

For example, consider the start codon 'ATG'. Changing the nucleotide in any of the three codon positions would result in a non-synonymous change thus for this amino acid there are three non-synonymous sites and no synonymous sites. In many condons, a change in the third positions results in a non-synonymous change. Consider for example Valine. Valine can be coded by GTT, GTC, GTA or GTG. As such, a change in the first two positions would be non synonymous but a change in the third positions would be synonymous. For some positions, such as lysine (AAA), 1 of the three alternate options in the third coding potion would be synonymous (G) and two non-synonymous (T, C). As such the number of synonymous sites is $\frac{1}{3}$ and non-synonymous sites $2\frac{2}{3}$. The most complicated example is potentially arginine which can be coded for by 6 different amino acids (CGT, CGC, CGA, CGG, AGA, AGG). As both the first and last position can be more than one nucleotide (the first positions C or A and the last positions any of the four bases) the number of synonymous sites is $1\frac{1}{3}$ and the number of non synonymous sites is $1\frac{2}{3}$ (Table 3.2).

The number of non-synonymous substitutions is calculated by comparing the sequence of interest with the reference, and considering the number of synonymous or non-synonymous changes that have been made. This is relatively simplistic with one change: GCG to GGG is one non-synonymous change or AAA to AAG is one synonymous change. If there are two changes more consideration needs to be given in the possible order in which they happened. If CGT changes to CAC this could have been CGT-CGC-CAC or CGT-CAT-CAC. In the first example the first change is synonymous and the second is non synonymous. In the second example both changes are non-synonymous (as in the product is a non-synonymous change from the original.) If both of these options are equally probable then we consider the synonymous mutation to be 1 and the non-synonymous mutations to be 1.5 (Table 3.3).

Codon(ref)	Codon	Synonymous	Non-synonymous
ATG	ATG	0	0
GCG	GGG	0	1
AAA	AAG	1	0
CGT	CAC	1	$1\frac{1}{2}$
TTA	TTA	0	0
	Sum	2	$2\frac{1}{2}$

TABLE 3.3: **An example of the number of synonymous and non-synonymous substitutions.** Changes from a reference sequence were used to demonstrate synonymous and non-synonymous changes

Considering the discussed examples (Tables 3.2, 3.3) with equations 3.4 and 3.5 $p_S=\frac{2}{4}$ and $p_N=\frac{2\frac{1}{2}}{11}$. If these values are then used in equation 3.2 and equation 3.3 d_N=0.0.27076 and d_S=0.82396. The dN/dS ratio is therefore 0.3286.

The dN/dS ratio is used to determine if the protein is undergoing selection. For example, if the ratio of non-synonymous mutations to synonymous mutations is significantly <1 we can suggest that deleterious non-synonymous mutations may have been selected against. This population would be refereed to as undergoing purifying or negative selection. Conversely, if the ratio of non-synonymous mutations to synonymous mutations is significantly >1 the population is considered to be undergoing positive selection or Darwinian selection; this is to say that selection has allowed some amino acids changes to happen. Positive and negative selection are not mutually exclusive. During purifying selection some non-synonymous changes may occur and during positive selection purifying selection will also be occurring but at a lesser rate. However, the dN/dS ratio does act as a good indicator of the selective pressure the population is under.

dN/dS ratio's for viral swarms To adapt the use of the dN/dS ratio to NGS data sets each read that covers a position must be considered rather than just comparison of individual consensus sequences. So, for each read that covers a sequence the number of synonymous and non-synonymous changes from the consensus sequence are calculated. As before, this is then divided by the expected number. The average is then calculated and can be used in equations 3.2 and 3.3 as previously described. This is an approach first used by Morelli *et al* [116] to account for read coverage at each codon.

Software has been designed to consider dN/dS ratios in NGS data. One example of this is 'diversiTools' by Richard Orton and Joseph Hughes [218].

In order to test this tool on viral swarm NGS data FMDV was recovered from the infectious copy plasmid PT7S3. This sample was sequenced and reads aligned to the published FMDV O1K genome as previously described [170].

Protein	dN/dS(3 d.p.)
Lpro	1.278
VP4	1.067
VP2	0.968
VP3	1.209
VP1	1.116
2A	1.320
2B	1.230
2C	1.110
3A	1.130
3B1	0.870
3B2	1.048
3B3	1.173
3C	0.945
3D	1.167
Entire ORF	1.116

TABLE 3.4: ***d*N/*d*S ratio for viral proteins in FMDV O1K** The dN/dS ratio for each protein in the FMDV genome was calculated using deversi tools

diversiTools was then used to consider the dN/dS ratio of the swarm (Table 3.4). This analysis showed that three of the fourteen viral proteins had a dN/dS ratio less than one and all others had a dN/dS greater than one. This suggests that the majority of the genome is undergoing Darwinian selection. When the entire open reading frame is used in this analysis a dN/dS ratio of 1.116 is calculated. For these values to be considered relevant they must be significantly greater or less than one. Further repeats would need to be completed to achieve this.

It is important to note however that this software is still underdevelopment. There is no universally agreed way to consider dN/dS ratios from NGS data sets. This is due to the inherent biases. For example, if a non-synonymous mutation was present at a particular position 100 times out of 100 reads the dN/dS ratio would calculate that that mutation had arisen 100x individually. However it is not that simple. That mutation could have appeared once and then been selected via fitness or bottles necks. As such dN/dS ratios from NGS data can be skewed. Modern dN/dS accounts for phylogeny and so avoid this

problem but due to the short reads produced by short read high throughput sequencing technologies this isn't yet possible for these data.

Therefore $d\mathrm{N}/d\mathrm{S}$ ratios are a useful tool in comparing consensus sequences and will be used at this level but they will not be used in considering the swarm.

Chapter 4

Measuring population diversity in FMDV: a pilot study

4.1 Abstract

A key feature of RNA viruses is the sequence diversity within a viral population. Several of the objectives of this thesis involved measurement and comparison of this diversity. Therefore a selection of currently published diversity indices were compared to assess their suitability for considering the data sets generated in this thesis.

From this comparison, one diversity index, Shannon's entropy, was then optimised considering its inherent benefits and limitations. A pilot study was completed to compare the swarms of seven different viruses.

Differences were found between the seven entropy profiles. These differences were statistically significant. Entropy data for structural and non-structural coding regions correlated well with current published data on conservation in some regions. However, some proteins appeared to have different levels of selection acting on them at a within the swarm level in comparison to a consensus level. This work showed Shannon's entropy was a good tool for comparing variation within the swarm and identified the manner in which these profiles can be compared.

4.2 Introduction

4.2.1 Diversity indices

Diversity indices are mathematical estimates of the diversity present within a population. They provide a quantitative measure of biological variability.

There are three main groups of diversity indices; species richness, heterogeneity indices and taxonomic indices. Species richness measures depend upon counting the number of different options in a population or community. From a genetic stand point, consider two positions: positions 1 has 997 A, 1 C, 1 T and 1 G, position 2 has 250 A, 250 C, 250 T and 250 G. In consideration of the number of options present at each position both are comparable with 4. However, if we consider the abundance of each of these options it can be seen that an A at position 1 is relatively conserved whereas position two is highly variable. Including a measure of the rarity of each item adds more information to purely the number of items available. This is what heterogeneity indices do. These indices measure the individuals present within a community and the evenness with which they are distributed. If both species number and evenness of these species distribution are comparable taxonomic indices can be used. These take into account the taxonomic relation between organisms in a population. This latter class of indices are more commonly used in measuring diversity in an ecosystem.

4.2.1.1 Heterogeneity indices

Heterogeneity indices combine the number of species present within a population and their relative abundance. There are two categories of these indices: parametric and non parametric. Parametric heterogeneity indices are based upon the parameter of a species abundance model whereas equivalent non-parametric indices make no assumption about the species abundance/distribution. In order to avoid the requirement for a relevant species abundance model only non-parametric indices were considered.

Shannon-wiener diversity index The Shannon-wiener diversity index or Shannon's entropy is one of the most widely used heterogeneity indices in the literature. It is based upon information theory. This index makes three assumptions:

1. Individuals are randomly sampled.
2. Samples are taken from an indefinitely large community.
3. All individuals are represented in the sample.

Shannon-wiener diversity (H') is calculated using equation 4.1.

$$H' = -\sum_{b \in \{A,C,G,T\}} p_b log_2(p_b) \tag{4.1}$$

Where p_b represents the proportion of each base $\{A, C, G, T\}$ present.

Brillouin index For samples where assumption one cannot be guaranteed other indices are available. This can be useful if one species or individual is likely to be preferentially sampled. In this case heterogeneity measures such as the Brillouin index are appropriate (Equation 4.2).

$$HB = \frac{ln(N!) - \sum ln(n_i!)}{N} \tag{4.2}$$

Where N is the total number of individuals and n_i is the number of individuals in the ith species.

Simpsons index The Simpsons index (Equation 4.3) is another heterogeneity index that expresses the probability that two of the same species will be drawn from random from a community. This is used for large sampled communities as it assumes an indefinitely large community.

$$\gamma = \sum_i p_i^2 \tag{4.3}$$

p_i represents the proportion of individuals found in the ith species.

The Hill numbers Hill defined a set of diversity numbers of different orders. These are used to show the relation between species-richness indices and heterogeneity indices. For example, the diversity number of an order would be defined as equation 4.4.

$$H_a = (\sum_i p_i^a)^{(1/(1-a))} \tag{4.4}$$

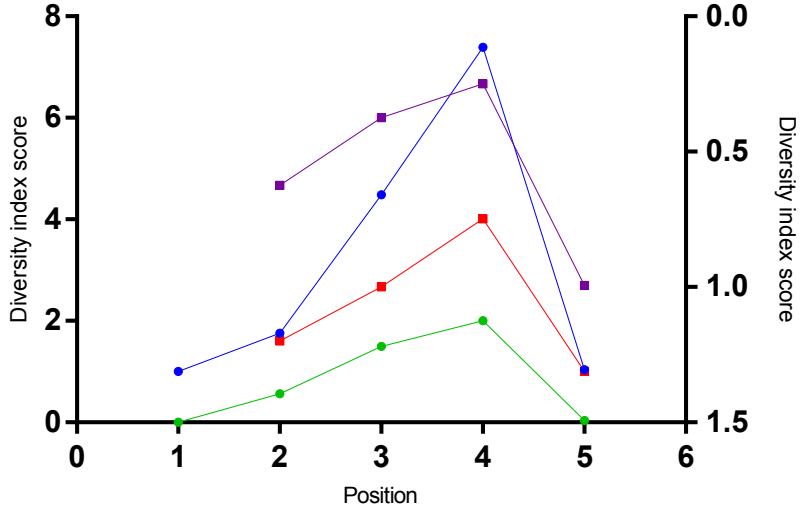
p_i represents the proportion of species i within the population. The most common Hill numbers are H_0 (Equation 4.5), H_1 (the exponential of Shannon-wiener diversity index)(Equation 4.6) and H_2(the reciprocal of Simpsons index)(Equation 4.7).

$$H_0 = S \tag{4.5}$$

$$H_1 = expH' \tag{4.6}$$

$$H_2 = \frac{1}{\gamma} \tag{4.7}$$

As one base is no more likely to be selected than another the Brillouin index was deemed unnecessary. Shannon-wiener diversity index and Simpsons index were compared.



	A	C	T	G	H'	γ	H_1	H_2
Position 1	1000	0	0	0	0	0	1	n/a
Position 2	750	250	0	0	0.5623	0.6246	1.754	1.601
Position 3	500	250	250	0	1.5	0.3744	4.481	2.670
Position 4	250	250	250	250	2	0.2492	7.389	4.0128
Position 5	997	1	1	1	0.03422	0.994	1.035	1.006

FIGURE 4.1: **Heterogenity measures present the data similarly** Diversity, as calculated with different indices, was compared for 5 genome positions with different proportions of bases. Diversity is shown measured by Shannons' entropy (H')(green), the Simpsons index (γ)(purple) and the hill number of these two measures (H_1(blue)and H_2(red) respectively). Shannon-wiener diversity index is show on the left y-axis and Simpsons index shown on an inverse right y-axis.

To asses how each of the diversity indices represents different distributions of nucleotides five potential examples were created (Fig. 4.1). These represented a range of possible degrees of variation. Both heterogeneity indices shows the same pattern revealing position 4 as the most variable and position 5 as the least variable (Fig. 4.1).

As all options were comparable, Shannon's entropy was used in future analysis due to the simplicity of comprehension (0 is least diverse, 2 is most diverse) and ease of calculation.

4.2.2 Shannon's entropy

4.2.2.1 Limitations

Coverage One factor affecting the robustness of the Shannons's entropy analysis is how the output is affected by areas of the genome with low coverage. Work was completed to explore the sampling properties of this diversity index and establish the level of coverage required to be reasonably confident the Shannon's entropy score produced is representative of the actual makeup of the virus population. This was first considered in a mathematical example. Five files containing 1000000nt were created. Each file contained a different proportion of each nucleotide designed to create an entropy score of 0.2, 0.5, 1, 1.5 and 2 (Table 4.2).

A	C	T	G	H'(1dp)
250000	250000	250000	250000	2
500000	250000	250000	0	1.5
500000	500000	0	0	1
89000	11000	0	0	0.5
97300	2000	700	0	0.2
1000000	0	0	0	0

TABLE 4.2: **A breakdown of the proportion of bases used to create files of each entropy score.**The value for 0.5 is accurate to 1dp. It was not possible to achieve an entropy score of exactly 0.2 or 0.5 with 4 positions split at whole number values.

It is apparent without mathematical analysis that any position with a Shannon's entropy score of 0 would be correctly represented regardless of the coverage as there would only be one base present at that position in the whole population thus there is only one possible outcome of sampling. For the other five examples the total files were randomly sampled to represent different levels of coverage and the Shannon's entropy calculated from each of these. This process was iterated 10 times and the average graphed.

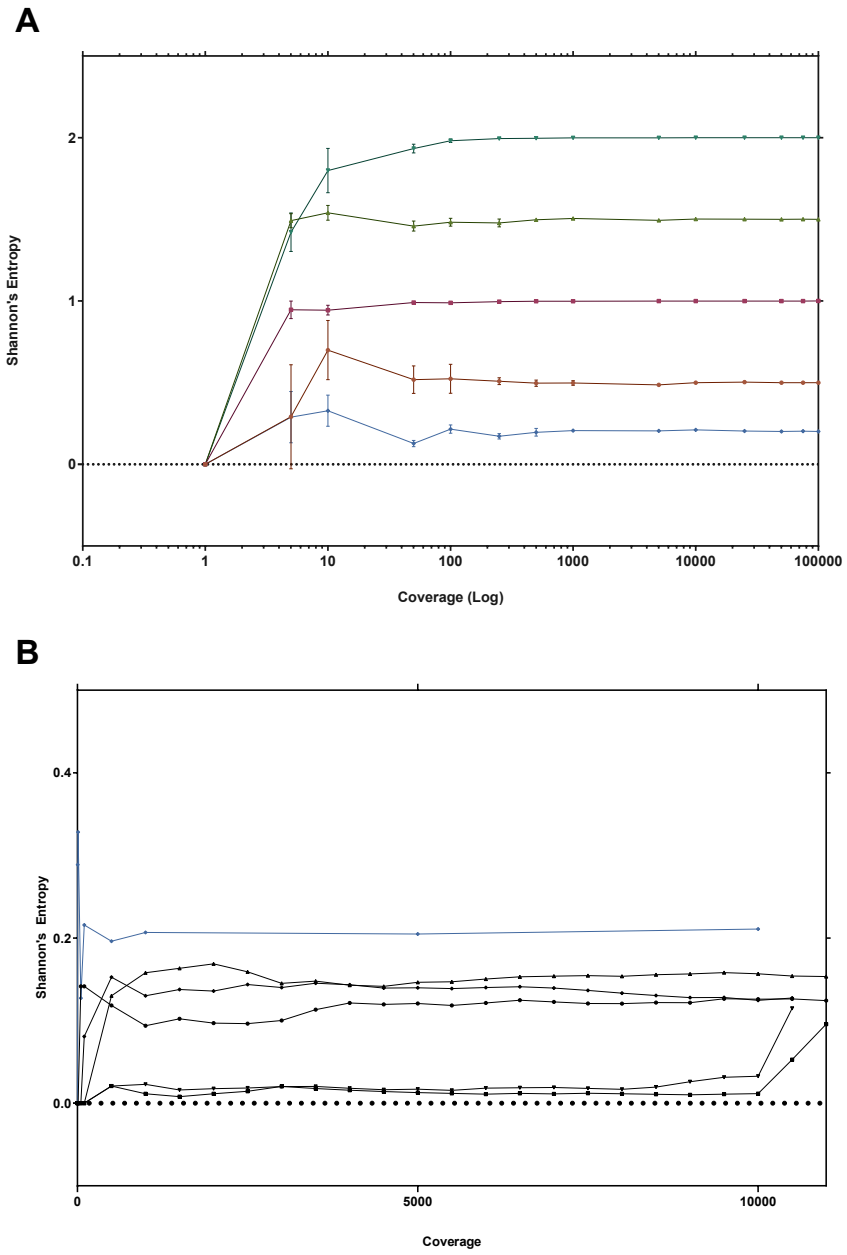


FIGURE 4.2: **Establishing a suitable coverage cut off for Shannon's entropy analysis.** A) Files of 100000 nt split into different proportions of bases were created to represent Shannon's entropy Scores of 0.2 (blue), 0.5 (orange), 1 (pink), 1.5 (green) and 2 (turquoise). These files were randomly sampled to represent different levels of coverage (1x, 5x, 10x, 50x, 100x, 250x, 500x, 1000x, 5000x, 10000x, 25000x, 50000x, 75000x and 100000x) and the Shannon's entropy calculated from these subsamples. This process was iterated ten times and the average plotted with a confidence interval of 95%. B) FMDV O1K was sequenced using the previously described method. An alignment to the known reference sequence was completed. The subsequent mpileup file was filtered to produce different levels of coverage (1x, 5x, 10x, 50x, 100x, 500x, 1000x, 1500x, 2000x, 2500x, 3000x, 3500x, 4000x, 4500x, 5000x, 5500x, 6000x, 6500x, 7000x, 7500x, 8000x, 8500x, 9000x, 9500x, 10000x, 10500x and (where sufficient reads were available) 11000x). Entropy analysis was completed at each of these levels of coverage. Five high coverage and high entropy (in comparison to other positions in this example) were plotted (black) alongside the mathematical example of 0.2 entropy (blue) (Fig. 4.2A)

For a file with a Shannon's entropy score of 0 (only one base represented) the score remains the same regardless of coverage. For all other entropy scores a plateau is reached from 250x coverage. This was reached more slowly and with more variation previous to the plateau in the files representing an entropy score of 0.5 and 0.2. This can be explained by the smaller subsets present; 11000nt and 700nt respectively. These smaller subsets have a lesser chance of being selected and thus require a larger sample to be accurately represented. The opposite of this phenomenon can be seen at an entropy score of 1.0. This file contains two equally large subsets of 50000nt. This is more quickly correctly represented being almost accurately viewed from 5x coverage and reaching a plateau by 50x coverage (Fig. 4.2A).

To assess whether the same cut-off was required in a real world example a similar analysis was completed using a sequenced virus. O1K was sequenced and analysed as previously described. The resulting sorted bam file was filtered to contain varying levels of coverage (Appendix D) and Shannon's entropy analysis completed on these files. Five high coverage nucleotides were compared (Fig. 4.2B). As with the lowest entropy score from the mathematical example these nucleotides exhibit some fluctuation before reaching a plateau. This is even more evident in these real world example due to the lower entropy score and the lack of iteration and averaging. In these examples a plateau is reached from 1000x coverage. This ignores the final data point for nt 7504 and nt 3352. In these positions there was a sharp increase in entropy at maximum coverage. This is a reflection of the way the script filters the sorted bam file. Reads in these files are sorted by read name, SAM flag, position, MAPQ and a series subsequent categories. Because of this the reads most diverged from the consensus would be listed last and thus present only in the maximum coverage files in the analysis. This issue would be overcome by using a bam file that had not been sorted but this proved challenging to design a script to process due to its format. In reality, these differences (if evenly represented in each entropy file) would results in only a slightly higher entropy score. This is an artefact that will only be apparent when using below the maximum coverage, such as in this example, but will not be reflected in future analysis of total viral populations.

Calculations completed on areas of the genome with lower coverage than 1000x may still be accurate but regarded with greater caution particularly for areas of the genome with a low entropy score.

Variation in Shannon's entropy Based on the analysis described above it was decided that in this work, Shannon's entropy score calculated at each position along the genome would be used as a measure of variation within the swarm. It is important therefore to show what this measure does and does not show.

First it is important to consider what Shannon's entropy analysis would define as most variable. The highest score possible is 2. This represents an equal proportion of all four bases being present in the swarm. The lowest score is zero. This is found at a position where only one base is represented. This fits in well with the concept of conserved and variable areas of the genome. Those positions that are variable can withstand having any of the four bases present (Shannon's entropy = 2) and those that are conserved always have the same base at that position (Shannon's entropy = 0). Although the score gives a good indicator of the variability of the genome position it does not reveal the underlying proportion of bases present at the position. Positions with a very similar Shannon's entropy score can reveal slightly different proportions of bases as shown in Table 4.3.

Position	A	C	T	G	H'(3 d.p.)
1	450	450	50	50	1.469
2	500	275	225	0	1.496
3	500	250	250	0	1.500
4	500	350	140	10	1.493

TABLE 4.3: **Differing nucleotide distributions can result in similar entropy scores** Hypothetical distributions of the nucleotide bases A, C, T and G are shown and the resulting Shannon's Entropy score (H') as calculated using equation 4.1.

For each positions shown in Table 4.3 H'≈1.5. This would suggest they were relatively variable positions. This is reflected in the nucleotide distributions with each of the positions being represented by three or four bases. However, the two positions with all four bases represented (Position 1 and Position 3, Table 4.3) show the lowest entropy scores (1.469, 1.493 respectively). This measure is considering not only the number of positions present but also the proportion of those represented bases so the positions with a higher proportion of bases not in the majority group have higher entropy scores. This highlights the need to be cautious about what the Shannon's entropy score shows as although it indicated a positions is variable, it may not show which position is the most diverse (with all four bases represented).

Overall Shannon's entropy is a good tool indicating well positions that are variable in comparison to those that are conserved. However, this method is limited on it's resolution and when dissecting minor changes at positions with similar entropy scores relative frequency of nucleotides may need to be considered.

4.3 Experimental Design

An experiment was designed to compare and contrast exemplars of FMDV to consider how the swarm structure varies.

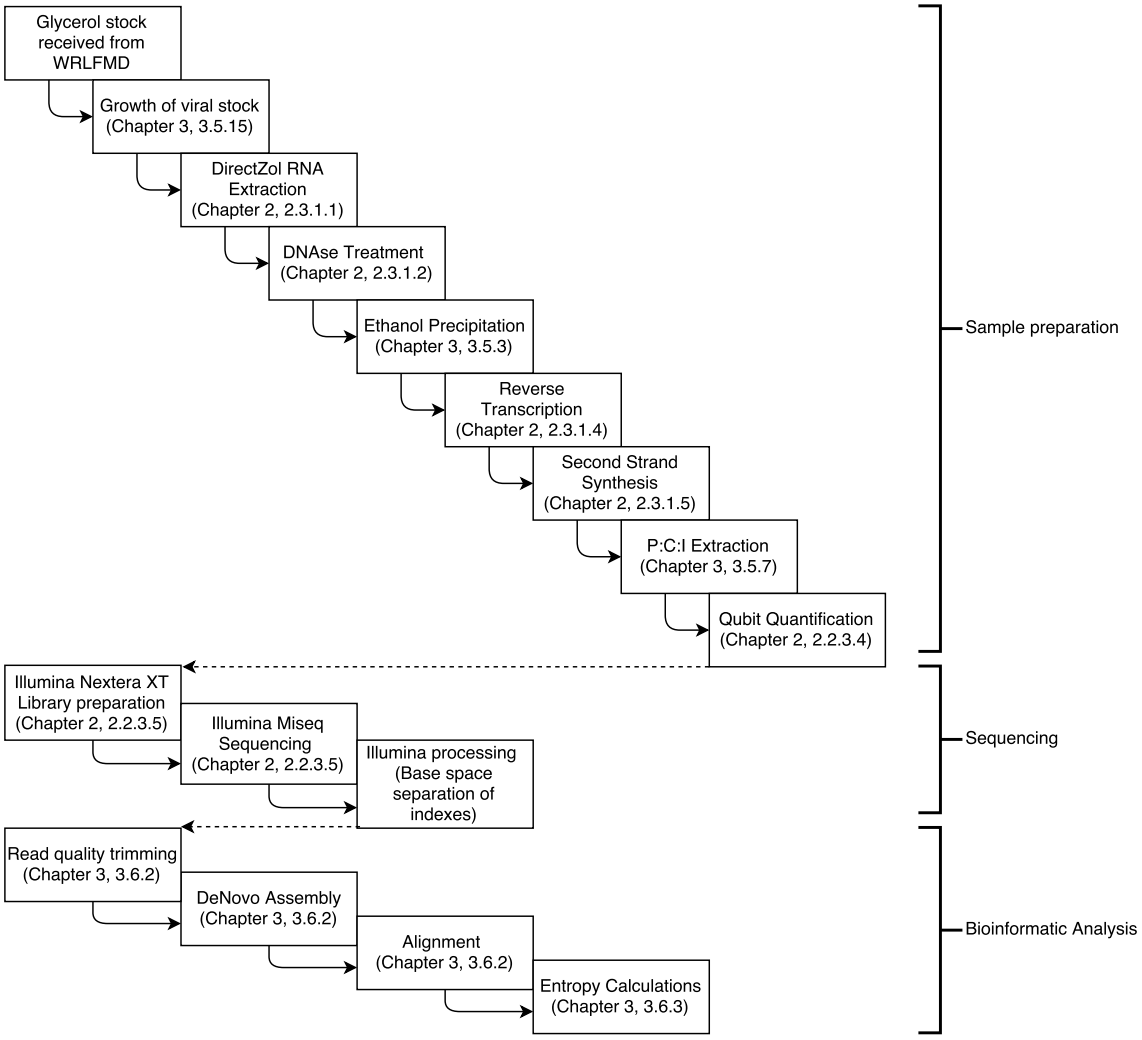


FIGURE 4.3: **Measuring population diversity flowchart of methods used** Viral samples were received from the WRLFMD and a viral stock created via one passage in BTy cells. Samples were prepared for sequencing as outlined in the published protocol [170] with some thesis specific amendments (Chapter 2). Library preparation was completed using the Illumina Nextera XT kit and subsequent sequencing completed on the Illumina MiSeq. Sequence analysis was completed as detailed in Chapter 3.

A panel of viruses were selected as an example. There were chosen for two reasons:

- Samples had a known passage history.
  - Field isolates of FMDV that are heavily passaged in cell culture show cell culture adaptation and are not necessarily representative of a wild type virus.

As such viruses were chosen from the World Reference Laboratory that had been isolated from an infected animal and subsequently passaged only once in primary cells to establish a large enough sample to sequence.

- Samples had a broad genetic distance.
  - To ensure a broad cross section of FMD viruses one exemplar was chosen from each serotype. This was hoped to ensure a broad genetic range.

The samples in use were received from the WRLFMDV at the Pirbright Institute and are detailed in Table 4.4.

Serotype	Isolate	Sample type	Material supplied by WRL	Passage history
Type O	TUR/11/2013	Epithelium	BTY1 ET/Glyc 21/05/2013	1x BTy
Type A	TUR/12/2013	Epithelium	BTY1 ET/Glyc 21/05/2013	1x BTy
Type C	KEN/1/2004	Epi Suspension	BTY1 ET/Glyc 05/02/2005	1x BTy
Asia 1	TUR/13/2013	Epithelium	BTY1 ET/Glyc 21/05/2013	1x BTy
SAT1	TAN/22/2012	Epithelium		1x BTy
SAT2	TAN/5/2012	Epithelium	BTY1 ET/Glyc 30/08/2012	1x BTy
SAT3	ZIM/6/91	Epithelium (Tongue)	BTY1 ET/Glyc 26/07/1991	1x BTy

TABLE 4.4: **Samples used for FMDV swarm comparison** Details of each isolate used and the serotype it has been classified as.

Samples were grown for one passage in bovine thyroid (BTy) cells to increase yield. This is a primary cell line and as such was used to avoid cell culture adaptation. Isolates then underwent sample preparation as detailed in the published method [170]. The NexteraXT library preparation kit was used and samples were sequenced on the Illumina MiSeq. Bioinformatics anlaysis was completed. Methods are outlined in Figure 4.3, method details are included in Chapters 2 and 3.

4.4 Results and Discussion

4.4.0.1 FMDV swarm variation

The Shannon's entropy score for each position along the genome was calculated for positions with a coverage of >1000x as detailed in the previous chapter. The distribution of entropy scores was then compared as a measure of variation within the

swarm. It was found that for each exemplar the majority of entropy scores were below 0.5 (Fig. 4.4). For two exemplars, SAT2/TAN/5/2012 and SAT3/ZIM/6/91, only one or two positions were above this score. Of the others, four revealed some entropy scores above one (TypeA/TUR/12/2013, Asia1/TUR/13/2013, TypeC/KEN/1/2004, TypeO/TUR/11/2013,) and one (SAT2/TAN/22/2012) had a large number of positions between 0.5 and 1 (33 positions) but none above one.

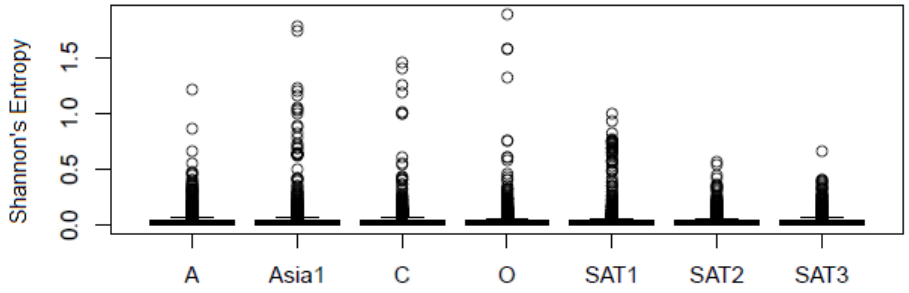


FIGURE 4.4: **The seven exemplars have varying numbers of high entropy positions** Shannon's entropy score (left y-axis) is shown for each exemplar TypeA/-TUR/12/2013, Asia1/TUR/13/2013, TypeC/TUR/11/2013, TypeO/TUR/11/2013, SAT1/TAN/22/2012, SAT2/TAN/5/2012 and SAT3/ZIM/6/91

4.4.1 Assessing if the sample viruses vary significantly in their entropy profiles.

The entropy scores were statistically compared. An F-test showed the variances for the samples are significantly different and thus a parametric ANOVA test was not appropriate. A non-parametric alternative (The Kruskal-Wallis test) was used. This showed there were statistically significant differences between the seven exemplars (chi-squared = 987.5852, df = 6, p=2.2e-16).

	Type A	Asia1	Type O	Type C	SAT1	SAT2
Asia1	2.795e-13					
Type O	2.2e-16	1.062e-09				
Type C	2.2e-16	2.2e-16	1.861e-05			
SAT 1	2.2e-16	2.2e-16	2.2e-16	2.2e-16		
SAT 2	2.2e-16	0.141	3.796e-06	2.2e-16	2.2e-16	
SAT 3	9.562e-06	0.09028	1.416e-12	2.2e-16	2.2e-16	0.0007189

TABLE 4.5: **Pairwise Wilcoxon tests to determine differences between exemplar entropy scores** A pairwise comparison of each exemplar with the other was completed using a Wilcoxon test and the p values are displayed.

A pairwise Wilcoxon tests was completed to understand which samples were different from one another. This showed the median entropies are significantly different for all serotypes except Asia1/TUR/13/2013 vs SAT2/TAN/5/2012 and Asia1/TUR/13/2013 vs SAT3/ZIM/6/91 (Table 4.5). This suggests that the amount of variation within a swarm does vary isolate to isolate. Although this shows variation isolate to isolate the differences can not be said it be indicative of the serotype as there is only one exemplar of each.

4.4.1.1 Can selection be detected within the swarm?

Work was completed to consider if selection could be detected in the swarm. This was completed by comparing structural and non-structural coding regions and comparing it to published data on conservation and selection. The majority of genome sites of each exemplar had >1000x coverage with all bar SAT3/ZIM/6/91 missing only 1.4 to 3.01% of their genome (Fig. 4.5). These missing regions were at the beginning and end of the genome which are notoriously under-represent as discussed in Appendix B (Fig. B.3). The area surrounding the poly(c)region also lack coverage as the sequencing chemistry struggles to successfully call long cystine tracts.

The sample with the least of it's genome represented was SAT3/ZIM/6/91. This isolate was missing coverage >1000x in 10.1% of it's genome. This is likely due to a lower proportion of the over all genome reads being attributed to this sample. There are numerous causes of this. Firstly, incorrect quantification of sample for use in the Nextera XT kit can affect the tagmentation reaction. Samples with more than one nanogram of DNA could result in undertagmentation. This results in an excessively large library that does not cluster efficiently on the flow cell. Conversely, samples with less than one nanogram of DNA could results in overtagmentation which produces a reduced library yield and therefore decreased coverage. Although all samples were quantified in the same manor, using the Qubit (flurometric quantification) there could still be discrepancies in the quality of the DNA. For example, the tagementation reaction requires DNA longer than 300bp to work. On the Qubit 5 x 200bp fragments of DNA would quantify the same as one 1000bp fragment however the 200bp fragments would not work in the tagmentation reaction. In a repeat experiment this could be checked using the bioanalyser to consider not only the quantity of DNA being put into the Nextera XT kit but also the quality.

Alternatively there could be elements of the sample preparation that are inhibiting the enzymatic reactions in the library preparation procedure. Three main issues have been identified by Illumina in this area.

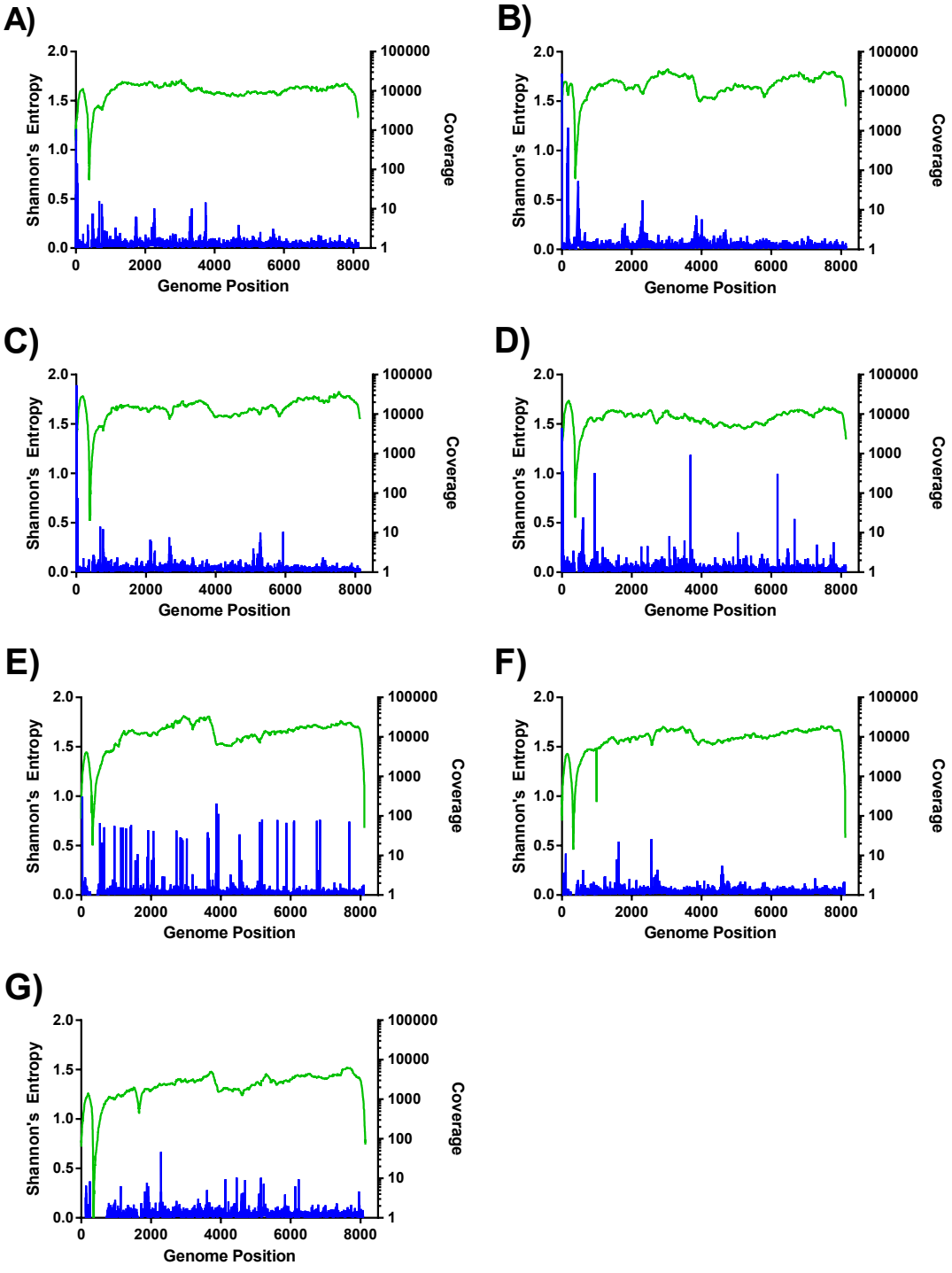


FIGURE 4.5: **Six of seven isolates have a comparable entropy profile** Shannon's entropy score (blue, left y-axis) is shown for each positions across the genome (x-axis) when coverage is greater than 1000x. Coverage (green, right y-axis) is shown for each positions along the genome on a log scale. This is detailed for each exemplar A) TypeA/TUR/12/2013 B) Asia1/TUR/13/2013 C) TypeO/TUR/11/2013 D) TypeC/KEN/1/2004 E) SAT1/TAN/22/2012 F) SAT2/TAN/5/2012 G) SAT3/ZIM/6/91

- Residual proteins have been found to bind DNA preventing subesequent enzymes binding.
- Enzymes can be degraded by proteinases or detergent/phenol contaminants left from sample preparation
- EDTA has been found to sequester enzyme co-factors

As the regions of the genome that were not represented remained at the ends of the genome and the area surrounding the poly(c)tract, although a slightly larger region, analysis was continued comparing regions with sufficient coverage.

Of the seven samples, 6 exemplars had a Shannon's entropy profile of low level fluctuations across the genome with some small clusters of high entropy positions. SAT1/TAN/22/2012 does not appear to fit in with this pattern (Fig. 4.5E). This isolate has a series of high entropy points across the genome each exhibiting a similar score. This difference is discussed in more detail in Chapter 5. TypeC/KEN/1/2004 also shows evidence of a number of high entropy positions across the genome although in a less consistent manner than SAT1/TAN/22/2012. The three points with high entropy scores in TypeC/KEN/1/2004 are at nt932 (5'UTR), nt3676 (VP1) and nt3171(3C). They have an entropy score of 1.0048, 1.18735 and 0.995289 respectively. The highest score is the nucleotide in VP1. This equates to the first coding positions in amino acid 138 of the protein. This is situated at the base of the GH loop which has well characterised relevance to immunogenicity and could be evidence that this swarm was under immune pressure. The positions identified in 3C is in amino acid 48. This is on positions 3' of *alpha*-helix 1. This is the first coding position of a lysine residue and therefore any other nucleotide will result in change. Asp^{45}, three positions 5' of this lysine, is involved in a Cys-His-Asp triad configuration important for protein structure. Therefore, a change at this point could be functionally relevant although without further structural analysis it is impossible to know [219, 220].

Although the overall entropy profile from sample to sample are similar (Fig. 4.5) the number of high entropy positions varies (Fig. 4.4, Table 4.5).

4.4.1.2 Conservation of the genome

Previous studies have been completed of FMDV genome conservation [221]. This work has indicated that some areas of the genome are more conserved than others. For example the region of the genome encoding for the structural proteins is generally considered to be more variable than the region coding for the non-structural proteins.

To consider if this previously described genome conservation was reflected by the swarm the location of high entropy positions was considered. To do this the frequency distribution for each exemplar was calculated. The 75^{th} percentile sat between 0.030 and 0.039 for all seven samples. Therefore an entropy score of 0.04 was classified as 'high'. A comparison was first made between the structural (VP4, VP2, VP3 and VP1) and non-structural (2A-3D) coding regions. When a virus is under immune pressure it is expected that the structural coding region will undergo Darwinian selection to facilitate immune escape. Consequently, it would be expected that there is more variation in the structural than non-structural coding region which might be reflected in the entropy scores.

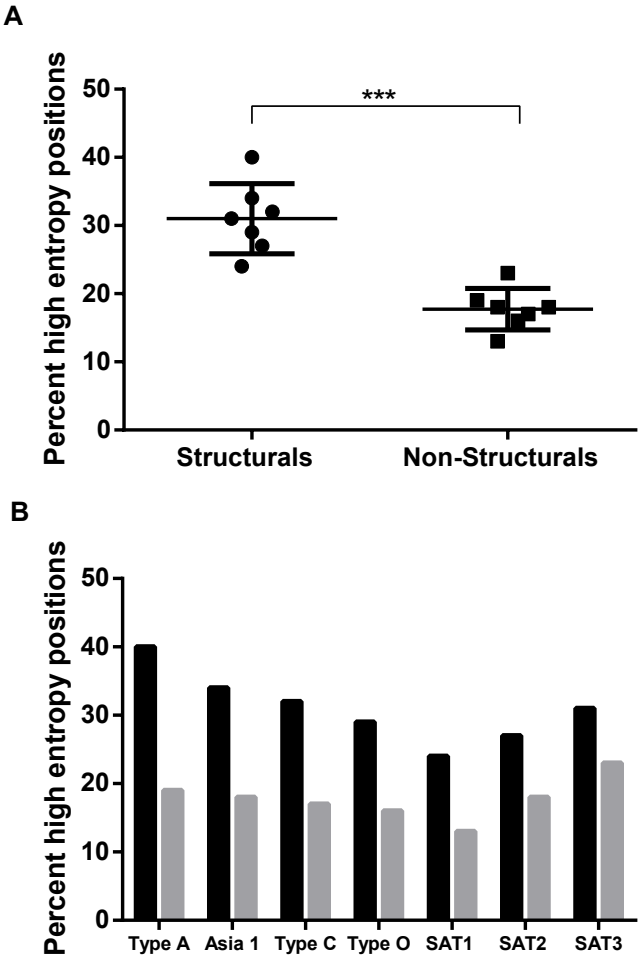


FIGURE 4.6: **More variation is evident in the structural coding region in comparison to the non-structural** The percentage of the coding region for the structural and non-structural proteins were compared. Positions above the 75^{th} percentile (>0.04) were considered high. The percentage of positions in each coding region above this cut off is shown in A. Due to the non-normal distribution statistical significance was analysed using the Mann-Whitney test, the values were found to be significantly different (p=0.0006). B) The percentage of high entropy positions in the structural coding region (black bars) and non-structural coding region (grey bars) is shown for each isolate.

The percentage of the structural region with high entropy positions was significantly higher than the percentage of the non-structurals with high entropy positions (Mann-Whitney, p=0.0006)(Fig. 4.6A). The percentage of high entropy positions in the non-structual coding region is more consistent between isolates (range=10%) than in the structural coding region (range=16%). When considered in closer detail, five of the seven isolates considered differ by only 3% in the proportions of their non-structural coding region with high entropy scores. This fits in well with the theory that the structural region is under greater selective pressure than non-structural coding region. However there are also other explanations. High entropy can arise as a result of relaxed purifying selection in areas of the protein that are not as important to viral survival or because of positive selection. Only a $dN/dS > 1$ is unambiguously indicative of positive selection.

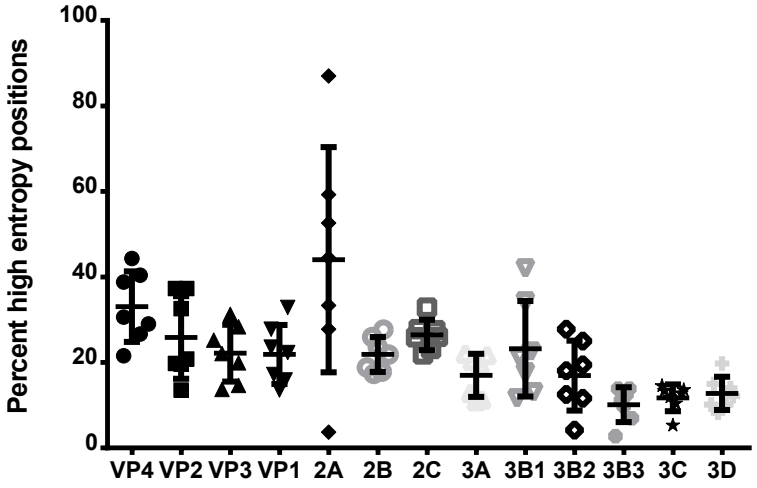


FIGURE 4.7: **Proteins have a varied proportion of high entropy positions and variable consistency of this proportion between samples** The percentage of positions in each proteins' coding region with an entropy score of >0.04 was considered. The mean and standard deviation is shown for each protein.

When considered on a protein to protein basis the entropy pattern is not so clear cut. The mean percentage of high entropy (>0.04) varies from 9.9% (3B3) to 43.9% (2A). However, The range for the data sets varies dramatically. The values for 2B, 2C, 3A, 3B3, 3C,and 3D all have a small range (9.2-11.6%), VP4, VP3, VP2, VP1, 3B1 and 3B2 have a slightly larger range (17.6%-30.4%) and 2A has the largest range (83.3%) (Fig. 4.7).

Published data on conservation Work by Carrillo *et al* [221] considered the conservation of FMDV proteins using consensus sequences of FMDV. Some of the data on

conservation from this study correlates with the results of these experiments. For example 3C and 3D have been found to be highly conserved with 76 and 74 % of their amino acids shown to be invariable respectively. This was echoed in the entropy scores with just under twenty percent of nucleotide positions in this coding region having high entropy scores. Data from Carillo *et al*'s paper also correlates with the findings for 3B. They found overall amino acid conservation to be 50%. Interestingly they found that 3B3 was highly conserved in all variants and reflected by the low number of high entropy positions found in our analysis.

In some genome positions these two sets of data show different things. For example, in the non structurals genes, higher percentages of nucleotides had high entropy positions, the highest of these was found to be VP4. In Carillo's work however VP4 was found to be the most conserved protein in the genome with 81% of it's amino acids being invariable.

This suggests that although low level variation can be tolerated in the swarm at these positions the variants rarely become fixed. This could be due to the lack of immune pressure on VP4. VP4 is the only entirely internal capsid protein. It has a similar amount of high entropy positions as the other capsid proteins. A Mann-Whitney test shows VP4 is not significant different from VP2 (p=0.12) and is weakly significantly different from VP3 and VP1 (p=0.0379, p=0.379 respectively). However VP4 is highly conserved (81% amino acids invaraibale) compared to the other capsid proteins that are more variable (VP2=47%, vP3=39% and VP1=24%). This suggest each capsid protein is equally variable but VP4 lacks the same selection pressure due to it's internal position.

The other notable outlier in this data is 2A. 2A is an 18 amino acid long protein whose function is not well understood. It facilitates its own separation from 2B by modifying cellular translational machinery resulting in ribosomal skipping preventing the original binding of 2A and 2B. Carillo found amino acids in this protein to be quite highly conserved with 65% of positions being invariant. In this work from 3% to 87% of genome positions were found to have high entropy positions hinting at some subconsensus variation although this level of variation appears not to be translated in the consensus level changes. Carillo *et al* suggests this could be due to the small size of the proteins, only 18AA in length, resulting in changes affecting both the structure and functionality. This could mean that nearly all mutations that arise are detrimental.

This work has detected a difference between the structural and non-structural coding regions at the within swarm level. This suggests that there is selection occurring or some molecular mechanism that allows less variation to be introduced in the non-structural coding region. However, the differences with Carrilo *et al's* consensus data conflicts with this. Where the data matches the consensus data from Carillo's paper it suggests

that selection is occurring at the virus replication level. Were it differs it suggests selection is potentially happening at a different level (for example the immune response and VP4). This raises an interesting question about the functionality of 2A and what would explain the disparity between selection at the consensus level (and the resulting genome conservation) and the selection within the swarm.

Cumulative Shannon's entropy Work was then completed to consider if the samples vary significantly in the amount of variation across the genome as a whole. Summing the entropy score of each position provides a cumulative entropy score indicative of the amount of variation in the swarm across the whole genome. This analysis was completed on the polyprotein of each sample to ensure the work was comparable. The 7th sample (SAT3/ZIM/6/91) was not used in this analysis as it was lacking sufficient coverage for the last 20nt of the polyprotein.

The cumulative entropy for each sample is similar varying from 199.9831 to 232.4176 (range = 32.4345). This suggests that the swarm for each sample is able to contain a similar amount of variation.

The entropy scores contributing to this value were then considered. Genome positions were binned into low entropy scores (<0.04), high entropy scores (>0.04<0.25) and very high entropy scores (>0.25). The number of low entropy positions shows an inverse patten to that of cumulative entropy (Fig. 4.8 purple triangles, blue circles respectively). That is to say that samples with high cumulative entropy scores have a lower number of positions with low entropy scores and samples with low cumulative entropy scores have a higher number of positions with low entropy scores. When considering the high entropy positions (red squares) it can be see that the pattern mimics that of cumulative entropy. The higher the cumulative entropy score the higher the number of high entropy scores in this range. However, when considering the very high entropy scores (green diamonds) we see that the pattern is not so simple. In some samples it is inverse of the cumulative entropy (sample 5) and in some it reflects it (sample 1). In sample 5 it appears that with an increased number of extremely high entropy positions the number of high entropy positions decreases and the number of low entropy positions increases. However, in a sample one with a lower number of very high entropy positions the number of high entropy positions is higher and the number of low entropy positions is lower.

This pattern hints at the concept that a viral swarm can only withstand a certain amount of variation. As such the cumulative entropy score of all six examples is relatively similar. This results in fluctuation among the distribution of positions with different entropy scores. As the number of extremely high entropy scores increase the number of high entropy scores decrease and low entropy scores increase. However, if the number of

extremely high entropy scores is low, the number of high entropy scores increases and the number of low entropy scores decrease. A more powerful test of this theory would be to observe cumulative entropy over time.

This could correlate with the concept of an error threshold. It may be that a swarm with above this level of variation is unable to survive.

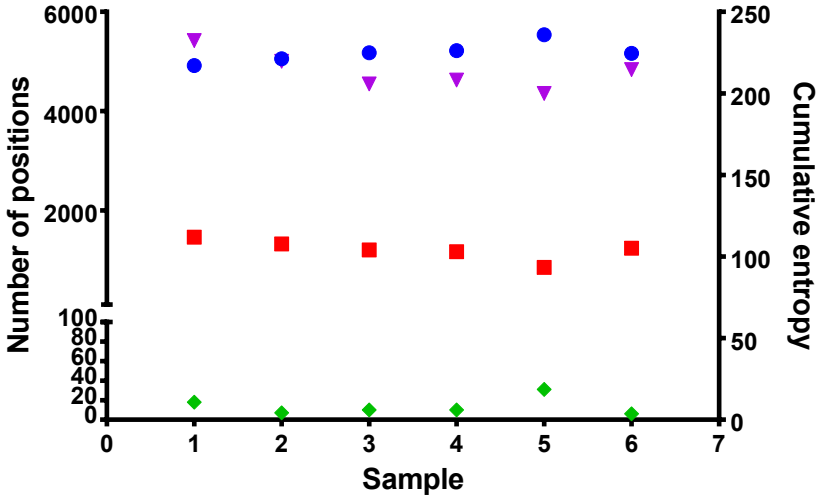


FIGURE 4.8: **Distribution of entropy scores in comparison with cumulative entropy** The cumulative entropy, as determined by summing the entropy scores for each nucleotide positions in the poly protein is shown on the right y-axis (purple triangles) for six samples 1) TypeA/TUR/12/2013 2)Asia1/TUR/13/2013 3)TypeO/KEN/1/2004 4)TypeC/TUR/11/2013 5)SAT1/TAN/22/2012 6)SAT2/TAN/5/2012. The distributions of nucleotides in each entropy score bin is shown on the segmented left hand y axis. This axis is divided into 0-100 for the lower 25% and 100-6000 for the upper 75% to allow visualisation of the lowest group. The number of nucleotide positions in the three bins is shown; <0.04 (blue circle), >0.04<0.25 (red square) and >0.25 (green diamond)

Although the cumulative entropy cannot be used as an indicator of if the virus is undergoing any form of selection the distribution of entropy scores could be reflective of this. As a virus adapts to a new environment, adaptive changes will appear and develop to the consensus level. These positions will move from low to high to very high entropy positions and then back down to low as they become fixed. As such, it could be that genomes with high numbers of extremely high entropy positions and associated high number of low entropy positions are under adaptive pressure. This is a concept that is discussed further in Chapter 6.

4.5 Summary

Shannon's Entropy is a good measure of diversity within the swarm. Shannon's entropy was found to be a good indicator of variation within the swarm at each position of the genome. Limitations relating to coverage provided by the NGS data were considered and the entropy calculation was found to be robust if 1000x coverage was achieved.

Entropy profiles between exemplars of FMDV are sometimes statistically significantly different. Visual comparisons of the entropy profiles produced showed some difference and these were confirmed statistically. This highlights the level of variation within the swarm structure of several strains of the same virus.

Selection can be identified at a within swarm level. There is evidence of selection occurring below the consensus, and selection differs in places from that previously identified at the consensus level. This has the potential for further defining the functional roles of FMDV proteins as important for replication, or higher level processes.

The amount of variation within the swarm (as represented by cumulative entropy) remains comparable. Although there is variation between the swarms from virus to virus, as detailed above, the cumulative entropy score for each virus remained comparable. This suggests that the swarm can only withstand a certain level of variation within it.

Chapter 5

FMDV genetic variability is influenced by the host species

5.1 Abstract

Southern African territories (SAT) serotype viruses have been shown to be more genetically variable at the consensus level than their European counterparts. The root of this increased level of variation was investigated.

SAT viruses were received from the WRLFMD and sequenced on the Illumina MiSeq. *De novo* assemblies were completed to produce consensus sequences which reads were subsequently aligned to. This alignment was used to calculate Shannon's entropy as a measure of variation at each genome position within the swarm.

This analysis indicated that the increased variability seen in SAT virus samples could be host derived. When viruses were divided according to the host they had been isolated from, samples isolated from cattle had less high entropy positions than samples from buffalo. Both the structural and non-structural coding regions of viruses isolated from buffalo show plasticity. Therefore, this increased variability does not appear to be solely due to immune pressure. Bioinformatic dissection of the swarm showed that the buffalo derived samples sometimes contained a second subconsensus level population. This could be due to either a secondary infection, co-infection or viral evolution within the host. Some subconsensus level variant populations were antigenically important and should therefore be considered in disease control. This work was completed on a small number of exemplars (n=24) and thus further work needs to be completed to confirm these preliminary findings.

5.2 Introduction

5.2.1 Genetic variability of SAT serotypes

The southern African territories (SAT) serotypes 1, 2 and 3 are endemic to sub-Saharan Africa. In this region they have been found to circulate along with type A, O and C. Type C has not been isolated since 2004 however and may be extinct in the region and potentially globally. These seroptypes have all been causative agents of outbreaks to varying extents. Of the 350 epizootics known in southern Africa since 1931 41 % have been attributed to SAT2, 25% SAT 1, 7 % SAT 3, 7 % type A, 6% type 0 and 1% type C . A proportion of outbreaks (13%) were caused by viruses that were untyped [222]. In more recent years, 2000-2010, SAT2 has been responsible for 41% of outbreaks, SAT1 19% and SAT3 only two outbreaks during this time period [223].

SAT serotypes are genetically diverse. They can differ by up to 40% of their nucleotides in VP1 [224]. They have approximately 10% more heterogeneity than the European types when comparing P1 [225]. This high level of variation produces a challenging disease control environment. High antigentic variability poses problems for single serotype vaccines as one vaccine will not protect against the large variety of viruses present in the area.

5.2.2 Wildlife Reservoirs

Wildlife reservoirs are thought to have an affect on FMDV transmission and maintenance in southern Africa. It has been found that in this area cattle (*Bos taurus*), kudu (*Tragelaphus strepsiceros*), warthog (*Phacochoerus aethiopicus*), bush pig (*Potamochoerus porcus*), wildebeest (*Connochaetes taurinus*), eland (*Taurotragus oryx*), waterbuck (*Kobus ellipsiprymnus*), sable (*Hippotragus niger*), buffalo (*Syncerus caffer*) and impala (*Aepyceros melampus*) are susceptible to FMDV infection [226–228]. Of these, two main species have been suggested as potentially relevant to FMDV spread and maintenance: impala and African buffalo. Impala and African buffalo have lesser severity or no clinical signs respectively but high susceptibility to infection. Impala are thought to be relevant in wildlife to cattle transmission and African buffalo have long been implicated as a maintenance host.

5.2.2.1 Buffalo

The relevance of African buffalo in FMDV spread in southern Africa remains controversial. Although it has been shown that buffalo can transmit disease to cattle, transmission

is erratic [229]. Although inconsistent, buffalo to cattle transmission has been shown to be the source for outbreaks [230]. African buffalo are a natural host of FMDV but they do not suffer clinical signs [231, 232]. Calves are infected after the lessening of maternal immunity that originally protects them [232]. Infection of buffalo is widespread. A study in Uganda found that of the 207 buffalo sampled 85% had antibodies against FMDV non-structural proteins, of these 96% were positive for antibodies against more than one serotype [233]. Buffalo have in fact been found with evidence of all three SAT serotypes in probang samples. This ability to maintain several serotypes, topoypes or strains could result in an environment that allows for intra and inter serotype recombination.

The sequence variation found between buffalo in the same herd can be great. In work by Vosloo *et al.* it was found that FMD viruses from two buffalo in the same herd sampled on the same day could be very similar (less than 2% nucleotide variation) or genetically diverse (up to 20% nucleotide variation) when comparing VP1 [224]. This is potentially due to a significant amount of genetic diversity of FMD viruses can be produced within a single African buffalo. Furthermore, the rate of change observed over a viral infection can increase via transmission to another animal [229].

It is generally accepted that buffalo can be persistently infected for 5 years. This stems from work completed by Condy et al [234]. However, this may not be accurate. Mechanisms of persistence remain inconclusive. Whether the virus persisted within the animal or within the group was not entirely clear.

Regardless of how infection is preserved in the individual or herd it is clear that FMDV can be maintained in African buffalo alone for extended periods of time and within this time the level of sequence variation has the capacity to increases.

5.2.2.2 Impala

Impala are widely distributed across southern Africa. The body of data relating to their involvement in FMDV epidemiology comes from work in the Kruger National Park (KNP) due to the high density of impala and the numerous scientific studies. All three SAT serotypes have been isolated from impala in the last 40 years although the majority of outbreaks have been SAT2 [235]. Impala show little to no clinical signs when infected with FMDV [227]. It is likely that impala are infected with FMDV via contact with African buffalo [236]. As buffalo and impala have little natural cause to interact it has been suggested that shared grazing and waterholes due to high density populations and seasonal droughts respectively could be relevant in between species transmission [235, 236]. Although buffalo continue to be considered the most relevant animal reservoir there has been some evidence of impala transmitting disease across

control fencing designed to prevent spread from infected buffalo [237]. Although clearly relevant in transmission from buffalo to cattle they are not considered as a maintenance host and therefore not considered in this body of work.

5.3 Experimental Design

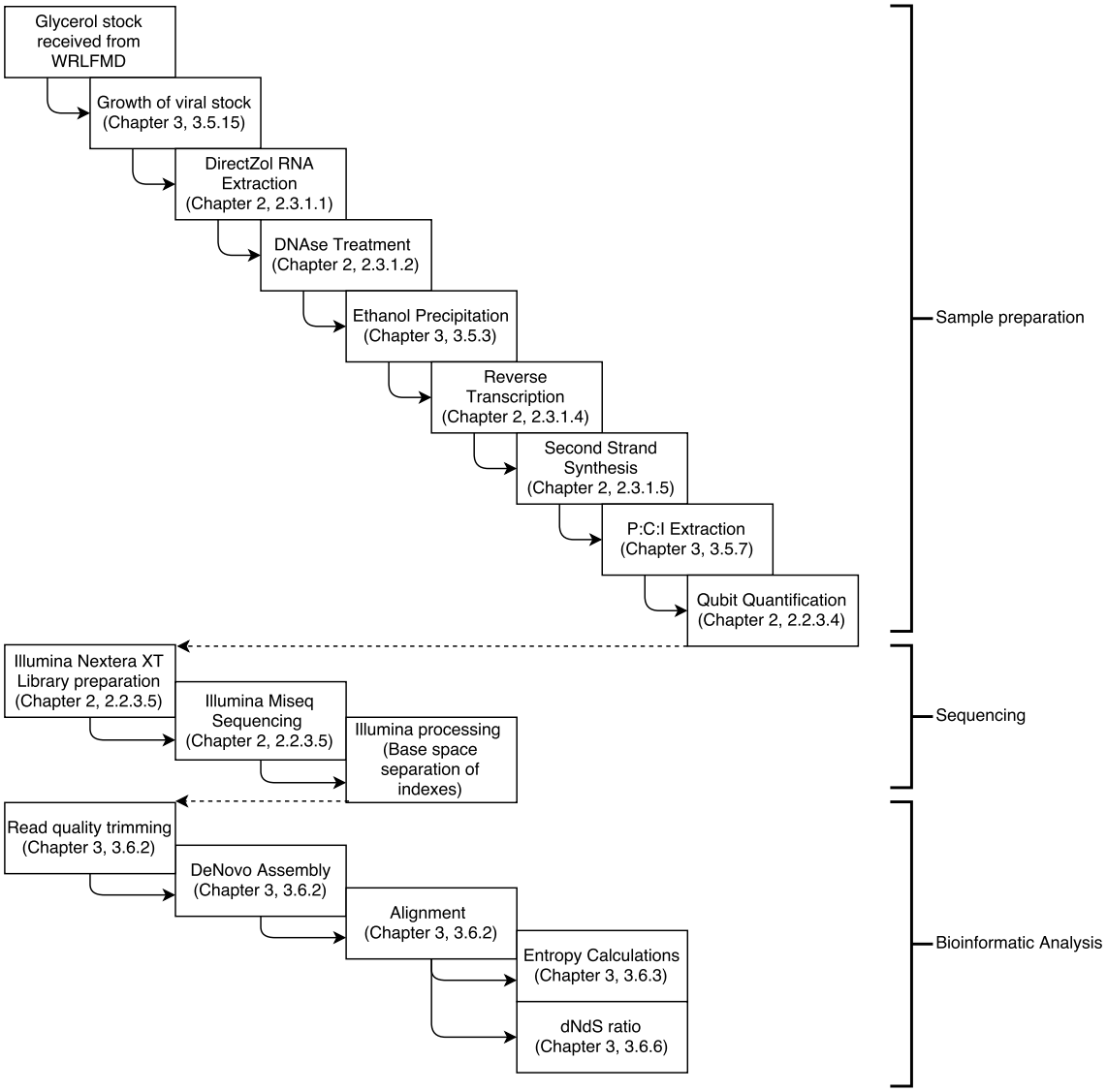


FIGURE 5.1: **FMDV genetic variability is influenced by host species flowchart of methods used** Viral samples were received from the WRLFMD and a viral stock created via one passage in BTy cells. Samples were prepared for sequencing as outlined in the published protocol [170] with some thesis specific amendments (Chapter 2). Library preparation was completed using the Illumina Nextera XT kit and subsequent sequencing completed on the Illumina MiSeq. Sample preparation and sequencing was completed by Lidia Lasecka and Caroline Wright. Sequence analysis was completed as detailed in Chapter 3.

Isolate	Serotype	Host	Date isolated	Country of origin	Sample type	Known passage history	Outbreak information
UGA/Buff/10/1970	SAT1	Buffalo	1970	UGA			
UGA/Buff/21/1970	SAT1	Buffalo	1970	UGA			
KNP/196/1991	SAT1	Buffalo	1991	SA			
BOT/Buff/62/1974	SAT1	Buffalo	1974	BOT			
SWA/2/1989	SAT1	Buffalo	1989	SWA			
TAN/22/2012	SAT1	Cattle	30/04/2012	TAN	Epithelium	1x BTy	
MOZ/5/1981	SAT1	Cattle	26/08/198	MOZ	Epithelium		
BOT/Buff/2/1969	SAT2	Buffalo	1969	BOT			
BOT/Buff/2/1968	SAT2	Buffalo	1968	BOT			
BOT/Buff/170/1974	SAT2	Buffalo	1974	BOT			
BOT/Buff/107/1972	SAT2	Buffalo	1972	BOT			
BOT/Buff/17/1969	SAT2	Buffalo	1969	BOT			
NR/1/1964	SAT2	Cattle	1964	NR			
SRHO/1/1965	SAT2	Cattle	1965	SRHO	Tongue epithelium		
ETH/1/1990	SAT2	Cattle	15/08/1989	ETH	Epithelium		
UGA/Buff27/1970	SAT3	Buffalo	1970	UGA			
ZIM/P6/83/Buff/19	SAT3	Buffalo	1983	ZIM			
BOT/Buff/13/1970	SAT3	Buffalo	1970	BOT			
ZAM/Nan/11	SAT3	Buffalo		ZAM			
BOT/109/1966	SAT3	Cattle	1966	BOT			
RHO/7/1974	SAT3	Cattle	31/07/1974	RHO	Vesicular Fluid		
RV/7/1974	SAT3	Cattle	1974	RV			
ZIM/2/1984	SAT3	Cattle	28/08/1984	ZIM	Lingual epithelium		Outbreak approx 10 days old

TABLE 5.1: **Known information regarding virus isolates used.**

During initial experiments to characterise the viral swarm (Chapter 4, Fig. 4.5), it became apparent that the SAT1 sample showed a number of high entropy positions throughout the genome not mirrored by any other sample. Previously published research in this area has found SATs to be highly variable. To investigate the cause of this variation and the unusual entropy pattern seen in this isolate it was considered in the context of a larger sample of SAT viruses sequenced and assembled by Dr. L. Lasecka and Dr. C.F. Wright. Methods used in this chapter are outlined in Figure 5.1. Details of these methods can be found in Chapter 2 and 3.

5.4 Results and Discussion

5.4.1 The viral swarms in buffalo show more genetic variation

Exemplars of SAT1 (n=8), SAT2 (n=8) and SAT3 (n=8) were considered in comparison to non SAT samples described in the published method (type O(n=1), type A(n=1), type C(n=1) and Asia 1(n=1))(Chapter 2).

As previously described, Shannon's entropy was calculated at each position along each genome. This score provides a measure of swarm diversity at each position. As this analysis can only be considered representative at positions with a coverage of >1000x (Fig. 4.2), only these scores were considered. Not all of the genomes considered had >1000x coverage at the same number of positions. Therefore these results are shown as a percentage of the genome analysed. Positions with a score of >0.25 were considered to have a high level of variability. In the non-SAT exemplars a relatively consistent percentage of positions are above this cut off (0.06-0.4%). Some SAT exemplars from each serotype sit in this low range, but some have a much higher proportion of their genome with high entropy position (range of 0.19-3.01%, 0.25-1.26% and 0.07-1.88% for SAT1, 2 and 3 respectively). On average, all three SATs had a higher proportion of entropy scores >0.25 than the non-SATs although this difference was only significant for SAT1 and 2 (Mann-Whitney p=0.0081, p=0.0283)(Fig. 5.2A). A proportion of SAT viruses appear to be similar to non-SATs, the others do not. This suggests the difference between the SATs and non SATS may be due to something other than the SAT viruses themselves. This explains why in the original seven serotype panel only the SAT1 showed a different pattern in comparison to SAT2 and SAT3.

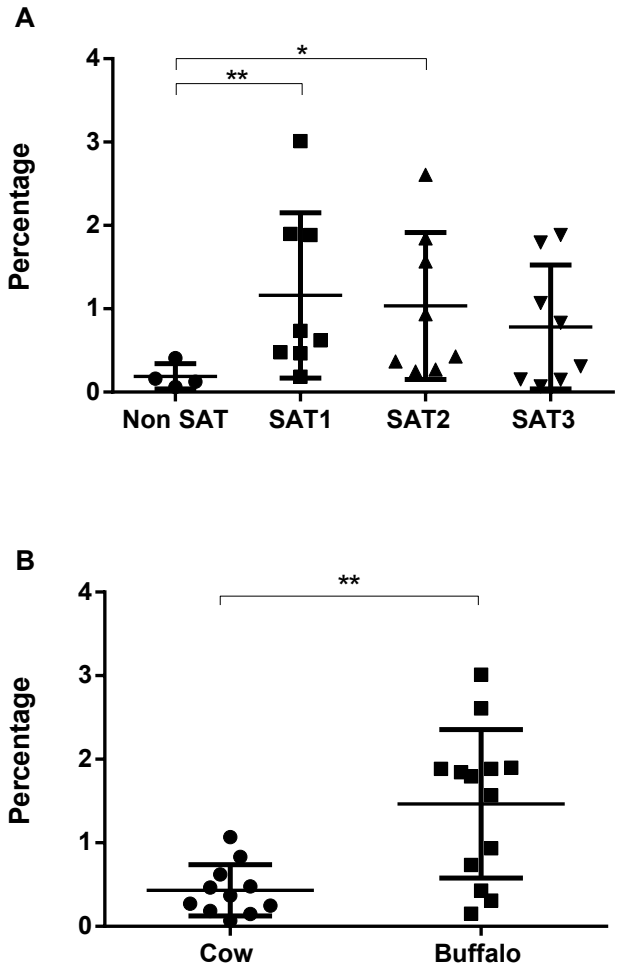


FIGURE 5.2: **Increased high entropy positions are a product of the buffalo host.** The Shannon's entropy score for each position along the genome was calculated for 'Non SAT' (This data set contained one exemplar each of type O, type A, type C and Asia 1). This analysis was also completed for 8 exemplars of SAT1, SAT2 and SAT3. In each graph the values are plotted as percentages (y-axis) of the positions with sufficient coverage for entropy analysis. Statistical comparisons were completed using a non-parametric(Mann-Whitney) test due to non-normal distributions. A) Percentage of positions for Non SATs, SAT1, SAT2 and SAT3 with an entropy score of greater than 0.25. Mean and standard deviation of each group is shown. There was a significant difference between Non SATs and SAT1 (p=0.0081) and Non SATs and SAT2 (p=0.0283). There was no significant difference between the SATs and each other or SAT3 and the non SAT exemplars. B) Percentage of positions for SAT samples isolated from cattle and SAT samples isolated from buffalo with an entropy score of greater than 0.25. Mean and standard deviation of each group is shown. There was a significant difference comparing bovine derived and buffalo derived sampes (Man-Whitney, p=0.0038).

In attempt to dissect the within serotype difference in entropy positions >0.25, the host from which the SAT samples was taken was considered. SAT samples were split into those isolated from bovine hosts (n=10) and those isolated from African buffalo (n=14). Samples isolated from cattle show a low proportion of positions above the

>0.25 cut off (0.07-1.07%). This is higher on average than the non-SAT samples (0.432 compared with 0.190). The buffalo derived samples show a much larger range (0.15-3.01%). Some values are similar to those found in the cattle derived samples (5/14) whilst the others have a higher proportion of positions with an entropy score >0.25. This was a statistically significant difference between samples isolated from buffalo and those isolated from cattle (Mann-Whitney, p=0.0038) (Fig. 5.2B). This supports the hypothesis that the differences seen in the number of high entropy positions is due not to the SAT viruses themselves and suggests it could be a product of the host from which they are derived.

Having shown there are more high entropy positions within swarm isolated from buffalo the genetic variability in the swarm was further analysed by calculating the relative frequency of each nucleotide at each position. This represents a more detailed view of the data from which the entropy scores described above are calculated. The majority of the buffalo samples show several nucleotides present at many positions (Fig. 5.3). The samples isolated from cattle generally show less positions with several nucleotides represented in the swarm (Fig. 5.4). This indicated that in general the viral swarms isolated from buffalo are more genetically diverse than those isolated from cattle. It shows that using the entropy scores as an indicator of within swarm variation works well.

The cause of this host specific difference was considered. The overall trend shows more variation within the swarm isolated from buffalo in comparison with those isolated from cattle. However there are some exceptions. In the samples isolated from buffalo SAT2/BOT/Buff/2/1968, SAT3/ZIM-p6-83-Buff-19 and SAT3/BOTBuff/13/1970 (Fig. 5.3 B.2, C.2, C.3) the level of variation is more comparable to the majority of cattle derived samples (Fig. 5.4). Amongst samples derived from cattle, SAT1/TAN/22/2012 and SAT3/BOT/109/1966 show more variation than the other cattle samples (Fig. 5.4 A.1, C.1). SAT3 sample BOT/109/1966 (Fig. 5.4 C.1) was isolated from a 'carrier cattle' (personal communication from Nick Knowles). It could be that the difference observed between cattle and buffalo is due to the longevity with which they are able to maintain an infection. If this is true, it would not be surprising to observe the same pattern in carrier cattle as buffalo and conversely the same pattern between recently infected buffalo and cattle. The variation observed could also be animal specific. Parallel challenges on their immune systems from other infections or well being of the animal more generally could be a contributing factor. Without a controlled direct comparison it is difficult to determine.

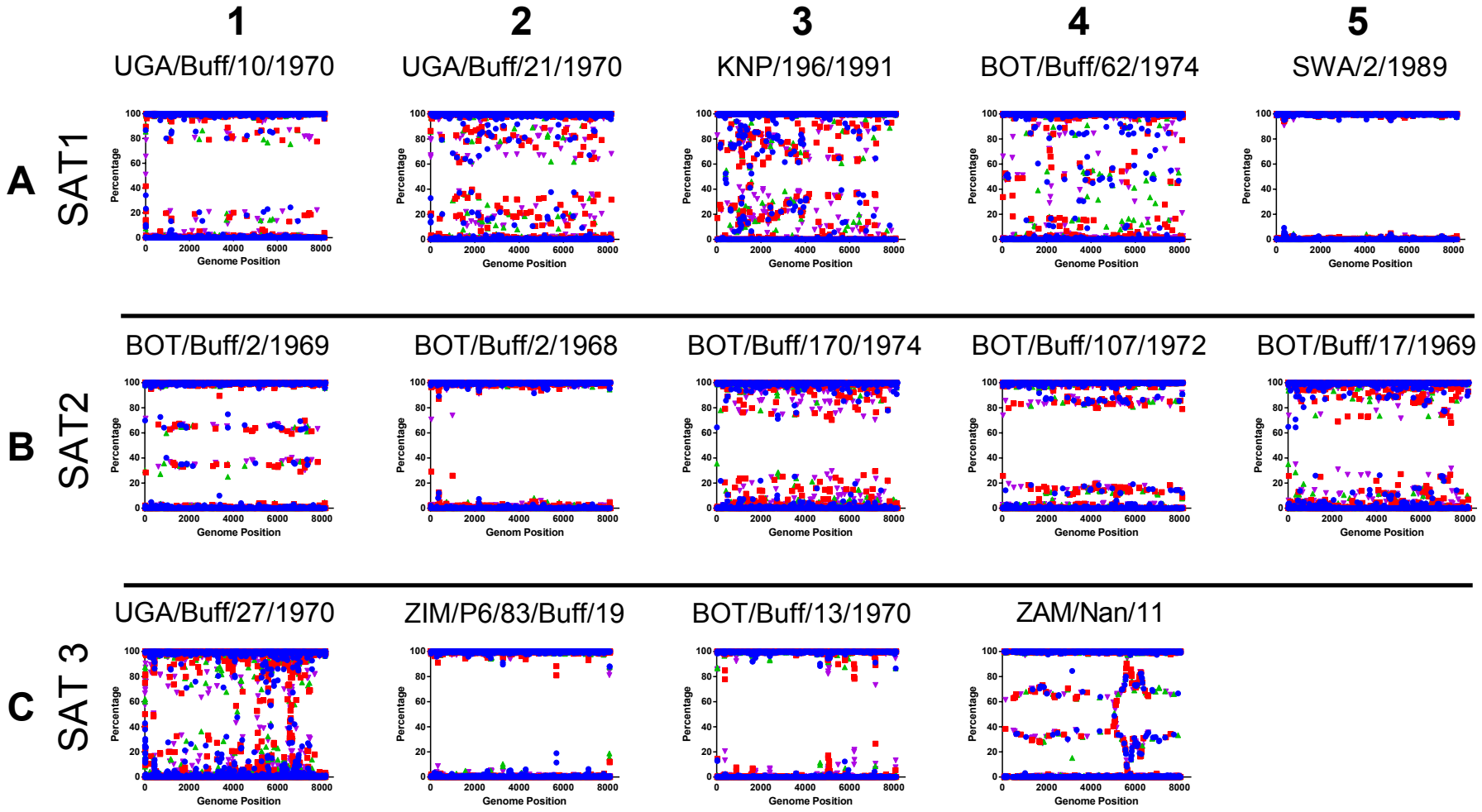


FIGURE 5.3: **Samples isolated from buffalo: Relative frequency of nucleotides at each genome position suggests two viral populations may be present** The relative frequency of each base (A-blue circle, C-red square, G-green triangle, T-purple triangle) at each nucleotide position was calculated and graphed against genome position.

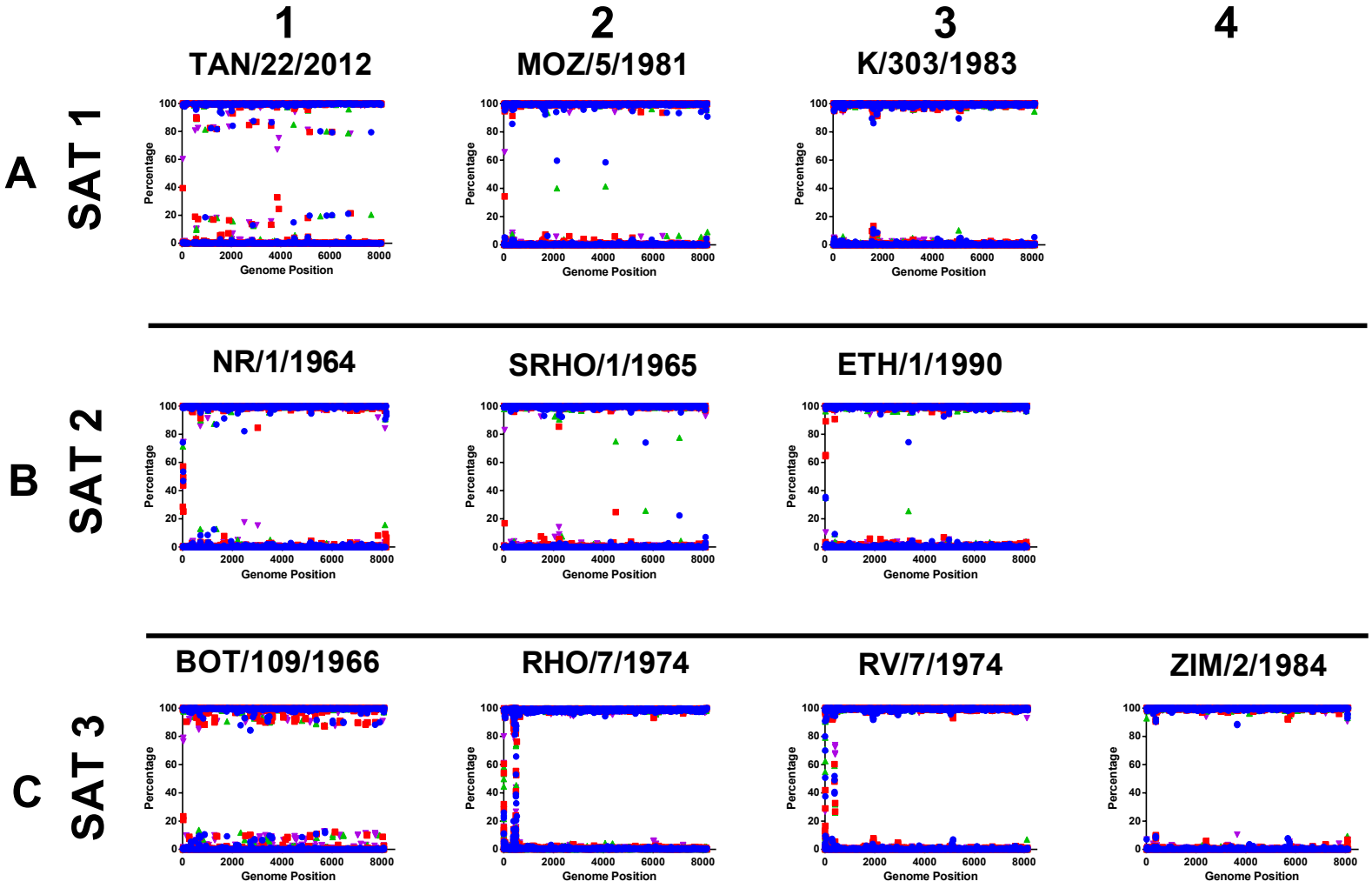


FIGURE 5.4: **Samples isolated from cattle: genome wide variability in the swarm appears low.** The relative frequency of each base (A-blue circle, C-red square, G-green triangle, T-purple triangle.) at each nucleotide position was calculated and graphed against genome position.

5.4.2 The effect of immune pressure on high swarm variability

To consider the evolutionary pressure on the buffalo derived viral swarm the location of the high entropy positions were considered.

5.4.2.1 Immune pressure acting on the virus particle

The immune systems recognises viral capsid epitopes. Therefore many immune changes can be found in the capsid coding region. There are numerous known antigenic sites allowing for immune escape.

5.4.2.2 Cytotoxic T-cells

The capsid monomers are not the only viral elements exposed to the immune system. Antigens (such as viral proteins or peptides) can be presented on the outside of an infected cell on MHC molecules. These antigens can then be recognised by the T-cell receptors (TCRs) on the surface of cytotoxic T-cell. TCRs are specific to particular antigens, if a specific interaction is made between the TCR and MHC presented antigen the cell is destroyed. In this way, traditionally internal epitopes could be under immune selection [238, 239].

The percentage of positions with high entropy scores (>0.25) in the structural coding region (VP4-VP1) and non-structural coding region (2A-3D) were compared for bovine derived and buffalo derived samples. For the cattle derived samples the average percentage of positions with high entropy scores in the structurals proteins was slightly higher than the non-structurals proteins (mean=0.34% and 0.19% respectively) although this difference was not significant (Mann-Whitney, p=0.1666)(Fig. 5.5).

In buffalo derived samples the average percentage of positions with high entropy scores is actually slightly higher in the non-structurals proteins (mean=1.3%) than the structurals (mean=1.15%). Again there is no significant difference between these data sets (Mann-Whitney, p=0.5712).

When comparing cattle and buffalo derived samples there is a statistically significant difference between the percentage of high entropy positions for both the structural and non structural coding regions (Mann-Whitney, p=0.004 and p=0.0009 respectively). This statistically significant difference highlights that there is a difference between cattle derived and buffalo derived viral swarms.

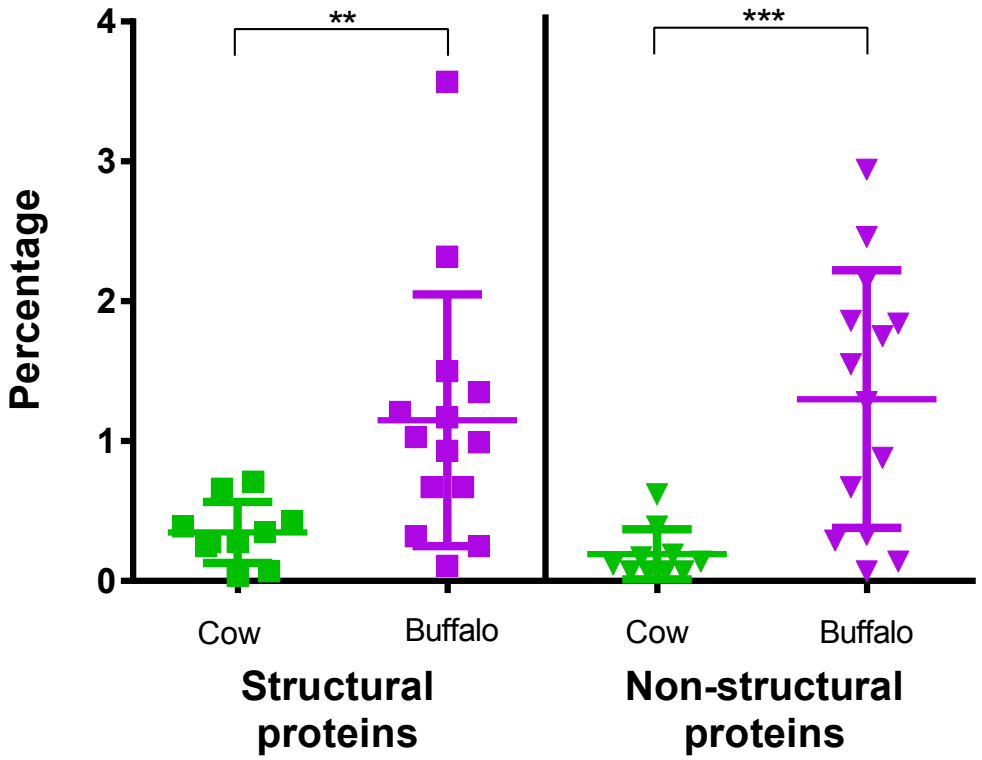


FIGURE 5.5: **High entropy positions in buffalo derived swarms are not concentrated in the structural region** Percentage of positions with high entropy scores (>0.25) is shown for the coding region of structural proteins (square points) and non structural proteins (triangular points). This is shown for cattle derived samples (green) and buffalo derived samples (purple). The mean and standard deviation of each data set is displayed. There is a statistically significant difference between cattle and buffalo in the percentage of both structural coding potions with high entropy scores (Mann-Whitney, p=0.004) and non-structural coding positions with high entropy scores (Mann-Whitney, p=0.0009)

These high entropy positions could represent both synonymous and non-synonymous changes. To understand which areas of the genomes may be undergoing genetic change, the high entropy positions where the second most prevalent nucleotide at that positions would result in a non-synonymous change were considered.

The number of non-synonymous changes in the structural protein coding region for samples isolated from cattle and samples isolated from buffalo was not significantly different (Mann-Whitney, p=0.3830). The mean of the samples isolated from buffalo is only marginally higher (0.35% in comparison with 0.25%) (Fig. 5.6). The number of non-synonymous changes in the non-structural protein coding region for samples isolated from cattle and samples isolated buffalo is statistically significantly different (Mann-Whitney, p=0.0018). Interestingly, 72 % of the high entropy positions in the structural coding region of viruses isolated from cows result in non-synonmous changes.

In comparison, only 31 % of the high entropy positions in the structural coding region of viruses isolated from buffalo results in non-synonymous changes. In the nonstructural coding region high entropy positions resulting in a non-synonymous change for viruses isolated from cattle was also higher; 45% in comparison to 17% for buffalo. This suggests there is a greater amount of noise or synonymous changes in the buffalo derived swarm.

This analysis suggests viral swarms isolated from cattle and bovine hosts are both under immunogenic pressure in the capsid region and this pressure results in a similar proportion of sites under going non-synonymous changes.

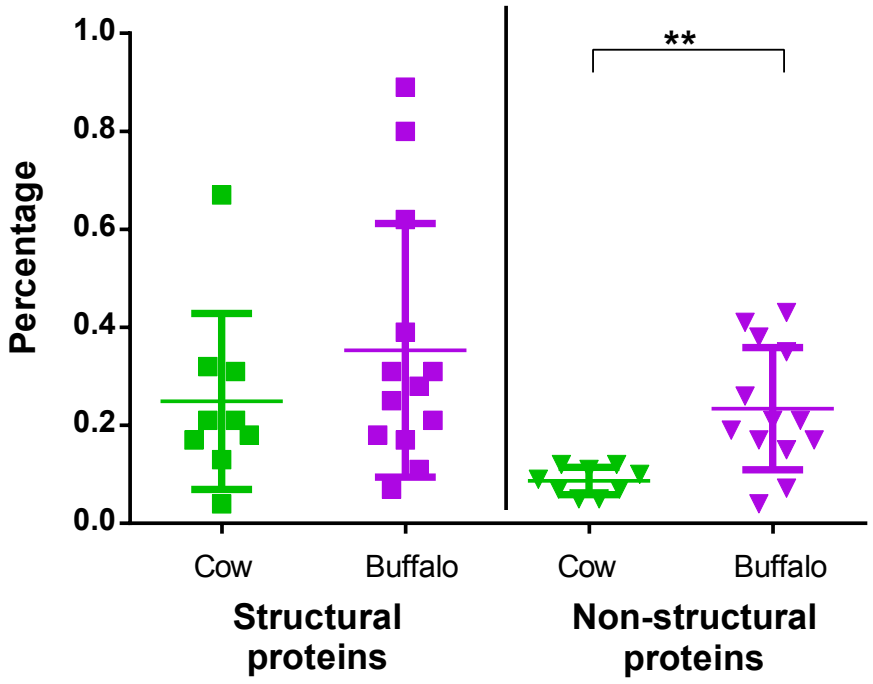


FIGURE 5.6: **Buffalo and cattle derived samples have comparable numbers of non-synonymous changes in their structural protein coding regions but differ in the non-structural coding region** Percentage of positions with high entropy scores (>0.25) that result in a non-synymous change when considering the next majority nucleotide is shown for the coding region of structural proteins (square points) and non structural proteins (triangular points). This is shown for cattle derived samples (green) and buffalo derived samples (purple). The mean and standard deviation of each data set is displayed. There is a statistically significant difference between cattle and buffalo in the percentage of non-structural coding positions with high entropy scores (Mann-Whitney, p=0.0018)

The comparable percent of the structural coding region undergoing non-synonymous change may be reflective of this region's capacity to undergo change. These sites are likely associated with immune escape. The level of noise seen in the buffalo derived

viral swarm coupled with the number of non-synonymous changes in the non-structural proteins is a marked difference from bovine derived samples.

5.4.3 Increased diversity in buffalo derived viral swarm may be due to co-infection

The differences described above could hint at a sub-consensus majority variant. If two virus swarms were present in the same sample the capacity for both non-synonymous and synonymous differences would increase. The pattern originally identified in SAT1/TAN /22/2012 was one of a consistent entropy score at several positions across the genome (Fig. 4.5). From the relative frequency graphs it can be seen that this was due to an equal split of two nucleotides at several positions across the genome (Fig. 5.4A.1). This may be due to two majority viruses being present within the swarm. If there were two majority viruses present, one present at 60% and one at 40% we would expect to see a consistent split of nucleotides at 60/40% at positions that vary between the two populations' consensus sequences. This pattern is repeated to an extent in one other bovine derived sample SAT2/BOT/109/1966 (Fig. 5.4C.1) and at least six buffalo derived samples (Fig. 5.3 A.1, A.4, B.1, B.4, B.5, C.4). It is difficult to clearly say from visual inspection if this pattern is repeated in the most variable buffalo samples as it would be obscured by the high level of variation present (Fig. 5.3 A.2, A.3, B.3, C.1).

The pattern observed in the samples isolated from buffalo is less evident in the samples isolated from cattle. Two of the samples have low level variation at a small number of positions along the genome and SAT1/TAN/22/2012 clearly shows the split in nucleotide detailed in the buffalo. This split has a lesser percentage of the minority variant (80/20%) and less positions are showing this pattern (33 nucleotide positions, approximately 0.4 %).

5.4.3.1 Bioinformatic dissection of swarm haplotypes

Analysis was completed to dissect the swarm's majority variants to confirm if more than one viral populations was present. If a sufficient number of differences are present in the genome then software can be used to piece together a prediction of the genomes present in the population. This requires the genomes to be sufficiently different to have several differences in one read that can be matched to differences in the next read along. As a large number of positions were identified, and they were spread across the genome (Fig. 5.4, Fig. 5.3) there seemed to be a high chance of establishing the original haplotypes.

Several programs have been designed to consider this question to date four of which are reviewed in Prosperi *et al's* paper: ShoRAH [240], QuRE [241], PredictHaplo [242] and Geneious *de novo* [243]. Prosperi *et al's* work used each of the programs to consider hepatitis c virus (HCV) and human immunodeficiency virus (HIV). They found that using contigs of 500bp and 900bp respectively the percentage of variants produced that were correct, were highest for PredictHaplo (100-75%) and QuRE (89-56%). Both were better with the shorter contigs (100% and 89%). Although ShoRAH output 200-1247 possible variants only 0.4-0.005% of these were found to be correct. However it did provide the best recall. The program that had the best recall/accuracy of the four compared in this paper was QuRE and thus was the one used for this work.

QuRE analysis was completed on SAT1 KNP/196/91. This was a sample that showed high levels of variation hopefully providing sufficient mutations for the QuRE algorithm to be successful. It was also a sample where the presence of a subconsensus variant was not obvious. This analysis reconstructed three variants at a prevalence of 80.2%, 18.3% and 1.4%. The QuRE program was unable to produce a full genome length sequence; each sequence ran from 503nt to 7026nt in comparison to the consensus. This resulted in a genome lacking 1147nt of the non-structural coding region. The 80.2% variant throughout the region represented was identical to the consensus sequence created *de novo*.

Comparison of the minority variants (18.3% and 1.4%) to the consensus revealed 48 and 55 nucleotides different in the structural coding region respectively. 30 of these differences were present in both. To evaluate if the QuRE reconstruction could be considered accurate it was compared to the Shannon's entropy analysis . An approximate 80/18/2 split, as seen in the QuRE variants, would be represented by an entropy score of approximately 0.8. Therefore the positions different between the QuRE variants should have an entropy score of between 0.8-1. Therefore, the number of positions with a score of between 0.8 and 1 should be comparable with the number of positions that differ between the QuRE variants if the reconstructions are accurate. There are 51 positions with entropy scores of between 0.8 and 1 which correlates well to the 48-55 differences between the 18.3% or 1.4% variant and majority. From this we can suggest the QuRE reconstruction of variants in the structural coding region is accurate. The non-structural coding region was also considered. Both the 18.3% and 1.4% variants differ from the majority by 4 nucleotides none of which were represented in both. This equates to only 1 amino acid in the 18.3% variant and none in the 1.4% variant. When compared to the number of positions identified in the entropy analysis the numbers are not comparable. In the non-structural coding region represented by the QuRE analysis 26 positions have a score between 0.8 and 1 compared to the 4 differences identified in the QuRE variants. This suggests QuREs reconstruction of the variants on assessment of the non-structural

region may not have been as successful. This could be due to the variants in these regions being more spaced out. 1.7% of nucleotide positions in the structural coding region were identified in Shannons entropy analysis compared to just 0.9% of the non-structural coding region. As this software relies upon mutations being frequent enough to map reads together, this lower number of variants may have provided insufficient framework on which the program could build.

5.4.3.2 Maximum likelihood tree of swarm haplotypes

To consider the relatedness of the QuRE variants reconstructed from the KNP/196/91 population a maximum likelihood tree was constructed. The variants were considered in comparison to all currently published full genome SAT sequences on genbank and unpublished sequences generated by Lidia Lasecka and Caroline Wright (Chapter A, Table E.1). The QuRE variants sit in a clade with other SAT1 viruses. They are most closely related to the virus from which they are derived used in Maree *et al*'s original experiment [231].

This analysis suggested they are quite closely related in the larger scale of all SAT viruses. However, there are not a large number of full genome SAT sequences available and therefore the power of this analysis is limited. Due to the lack of time data it is challenging to identify how each variant evolved. There is a 1.6% difference in the structural coding region in comparison to the reference. This correlates to the predicted rate of change in VP1 in a buffalo over a year. Therefore, it is conceivable that the 18.3% variant evolved from the same starting population as the majority population if the animal was infected for a sufficient length of time. However, if this infection was less than a year in duration, or if the virus does not replicate at a consistent rate in a carrier animal the 18.3% population could be a secondary population. This bioinformatic analysis could be corroborated through molecular techniques such as plaque purification.

5.4.4 Subconsensus level swarm haplotypes can be immunogenically relevant

Having established that there are several populations present in KNP/196/91, work was completed to dissect what these differences were and the evolutionary significance behind them. As the QuRE algorithm was only able to reconstruct the structural coding region only those changes were considered.

Samples were grown in cell culture and thus these changes were compared to known cell culture adaptations. Some previous work has been completed on SAT adaptation

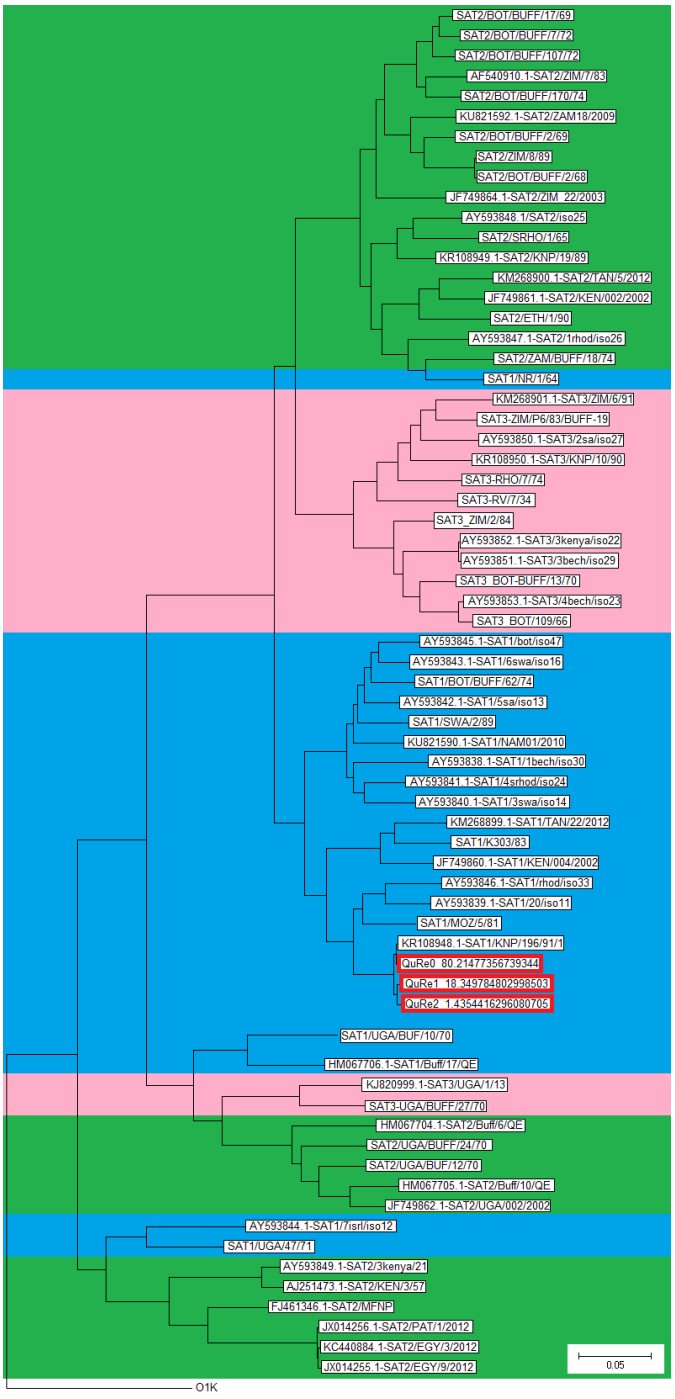


FIGURE 5.7: **Maximum likelihood tree of KNP/196/91 variants** The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model [244]. The tree with the highest log likelihood (-103411.8809) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 69 nucleotide sequences. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 5660 positions in the final dataset. Evolutionary analyses were conducted in MEGA6 [245]. The tree is routed with an FMDV Type O (Kaufbeuren) exemplar. QuRE variants are highlighted in red. SAT1 samples are showing in blue, SAT2 in green and SAT3 in pink. All viruses used in the Maximum liklihood tree (MLT) are listed Chapter A, Table E.1.

to cell culture. In a study by Maree *et al* a virus isolated from an impala was passaged in a number of cell lines. Cell culture adaptation mutations were observed in VP3/192, VP3/217, VP1/69, VP1/84 and VP1/110-112 [223]. None of these difference have been seen between the QuRE variants.

The difference were also compared with known antigenic sites. A paper by Maree *et al* predicted antigentically variant sites using structure information, cross protection titres in virus neutralisation assays and amino acid variants [223]. This study identified surface exposed structural loops in VP2, VP3 and VP1. A potentially antigenic region in one of these loops (βH-βI) spanned VP1/177-VP1/179, this includes VP1/178 which has been identified in this study (Table 5.2). The C-terminus of VP1 was also implicated, a feature that has been identified in the variants at VP1/205 and VP1/209 (Table 5.2). In VP1 the βH-βI loop and C-terminus correlate with epitopes previous identified in other FMDV serotypes: type A, type O and type C. None of the loops identified correlated with the variants found in VP2 or VP3.

The identification of mutations in positions previously associated with immune escape is important from a diagnostic and epidemiological stand point. These mutations would not be identified by current diagnostic procedures but could be relevant in the epidemiology of FMDV and indeed vaccine matching.

Amino acid	18.3%	1.4%	Genome Region
10	K → R	K → R	Lpro
19	T → K	T → K	
81	D → G		
96		I → V	
103		T → V	
147		D → E	
151		A → V	
166		E → D	
168		T → A	
332	Q → R	Q → R	VP2(57)
454	S → N	S → N	VP3(10)
716	H → Y		VP1(17)
877	E → V		VP1(178)
904	R → K		VP1(205)
908	R → K		VP1(209)

TABLE 5.2: **Amino acid changes between QuRE variants** 18.3% and 1.4% variants were compared to the 80% majority. Amino acid changes are listed annotated with the genome region in which they fall. Four changes appear in both variants and are highlighted in red.

QuRE analysis could not be completed on all samples due to insufficient number of variants on each read. Sequencing with a longer read length would likely overcome this difficulty. The sample completed (SAT1/KNP/196/91) suggests that within the higher level of variation seen in buffalo a secondary populations exists and that subconsensus populations can have antigenically relevant changes.

5.4.5 SAT3 swarm's distinct genetic features are not host derived

In much of the analysis described above SAT3 viral swarms appear to differ from SAT1 and SAT2. In comparing the bovine and buffalo derived samples SAT3 appears to display some interesting genetic features in comparison to SAT1 and SAT2. In two of the cattle derived samples and two of the buffalo derived samples regions of extreme variability appear to be present. This can be seen at the 5'end of the genome in SAT3/RHO/7/1974 and SAT3/RV/7/1934 (Fig. 5.4C.2, C.3) and around the 6000nt mark in SAT3/UGA/Buff/27/1970 and SAT3/ZAM/Nan/11 (Fig. 5.3C.1, C.4). The latter hot spot of variation is particularly unusual given the usually conserved nature of the non-structural region. In SAT3/ZAM/Nan/11 it is clear that this area of variation appears to be a fluctuation between two majority variants at a 60/40 split. One possible explanation for this could be that it is representative of a recombination hotspot. SAT3 has previously been found to have the most evidence for recombination within it's genome in comparison with SAT2 and SAT1 [139] and recombination sites around the 6000nt point have previously been suggested [138, 139]. These nucleotide positions equate to 3B and 3C. Numerous recombination sites have also been suggested in the 5'UTR of the genome which could equate to the hot spot seen in the two cattle derived samples (Fig. 5.4C.2, C.3).

These difference between SAT3 and SAT1 and SAT2 are of particular interest due to their differences from an epidemiological stand point. SAT3 causes the lowest number of outbreaks of the three south African serotypes and these differences in variation within the swarm could hint at the genetic cause behind this. If SAT3 is the most able to recombine, this could result in fitness loss. Recombination reduces the ability to undergo purifying selection due to the exchange of genetic material between genomes. If a buffalo host produces increase variability, but the SAT3 predisposition to recombine

results in a decreased ability to undergo purifying selection, it may result in SAT3 being less fit than the other two SAT seroptypes and thus is out competed by them from an epidemiological stand point.

5.4.6 Bias introduced by sample collection

The sample used in this study were supplied by the World Reference Laboratory for FMD (WRLFMD) at the Pirbright Institute. These samples are sent for diagnostic testing from all over the world and their sample collection method is not always known. To ensure the pattern viewed is not a sampling bias, samples with a known collection history were compared.

5.4.6.1 Probang vs Epithelium

African buffalo infected with FMDV show no clinical signs. Sampling from infected livestock often focuses on sampling epithelial lesions in the mouth or foot.

However, due to the lack of these clinical signs in buffalo, epithelial sampling is unlikely to be viable. Therefore it is probable an oesophageal sample is taken from these animals using a probang. As the probang samples the length of the oesophagus it could be sampling several different replication sites, meaning this sampling methodology could show more variation than a sample extracted from an epithelial lesion. This potential difference in sample collection from buffalo and cow could be behind the observed difference detailed in the section above.

Direct comparisons of probang and epithelial samples from African buffalo infected with SAT viruses were not available for analysis. Therefore, sequence data from a previous controlled animal experiment performed at the Pirbright Institute was used to ensure known sample history. This comparison established if probang sampling in cattle produces the pattern observed in buffalo. Data was provided by Caroline Wright, from samples previously extracted and sequenced as detailed in Morelli *et al*'s paper [116].

The samples used represented a probang sample (PB) taken at four days post contact (4DPC) and an epithelial sample taken from the front right foot (FRF) taken six days post contact (6DPC). The coverage, entropy score and relative frequency of nucleotides at each positions was considered for each of these samples, and the two samples as one.

The samples analysed from the FRF and PB show very similar Shannon's entropy and relative nucleotide frequency (Fig. 5.8A,B). The PB sample does not show a different pattern than the sample retrieved from the FRF lesion. To further investigate

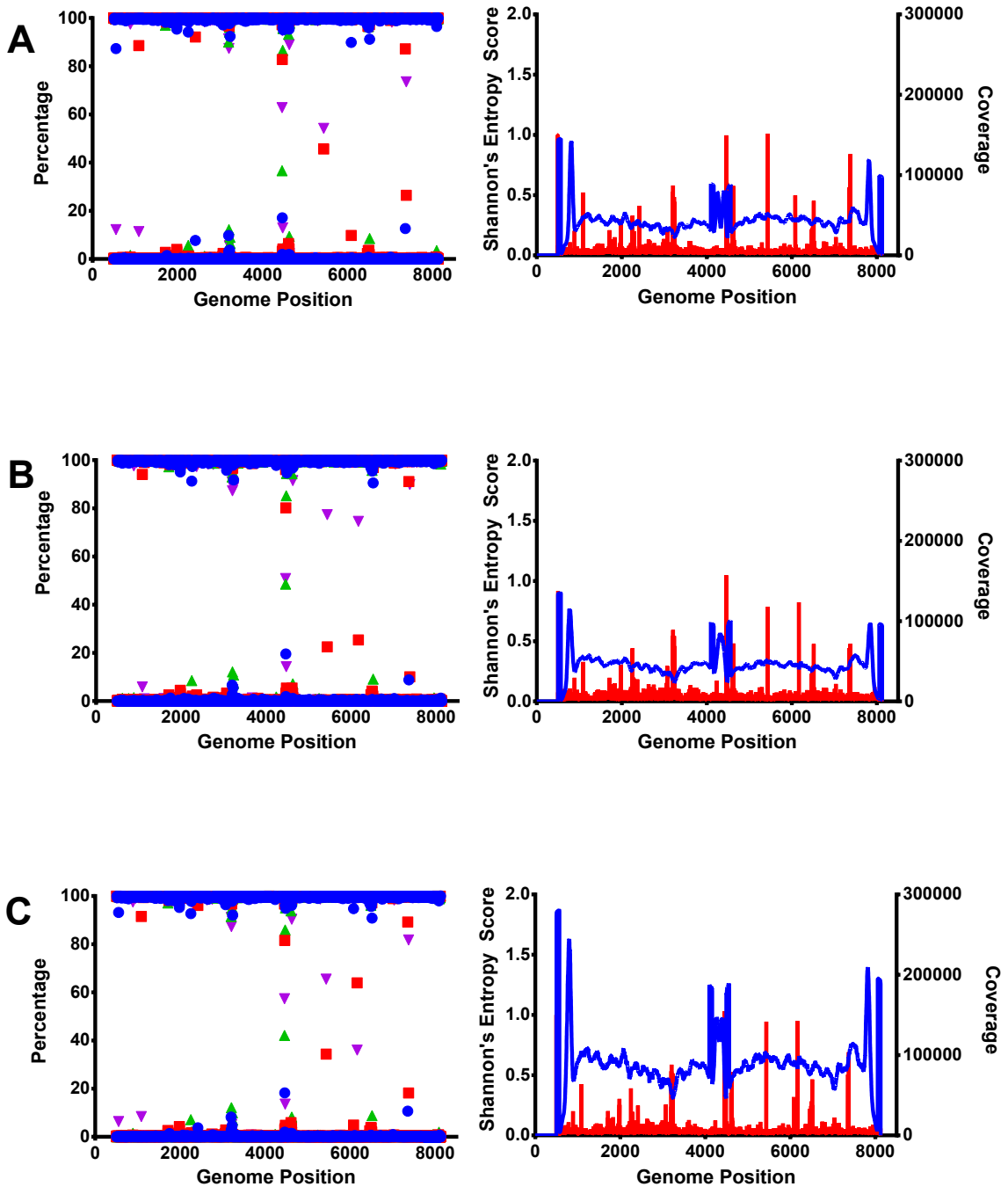


FIGURE 5.8: **Probang sampling, or sampling two replication sites in one, does not produce the pattern observed in buffalo.** The relative frequency of each base (A-blue circle, C-red square, G-green triangle, T-purple triangle.) at each nucleotide position was calculated and graphed against genome position (Left most graphs). Concurrently (right most graph) Shannon's entropy at each position (left axis, red) was compared with genome coverage (right axis, blue). This was completed for three samples A) Probang sample (PB 4DPC) B) Foot lesion (FRF 6DPC) C) Probang and foot lesions combined (PB 4DPC + FRF 6DPC)

the possibility that probang sampling identifies several sites of replication, the data from the FRF lesion and the oesophagus were combined and the analysis completed again. This did not result in the pattern observed in buffalo (Fig. 5.8C).

This comparison is limited due the the different host (cattle) and serotype (FMDV type O). It could be that the amount of variation generated and maintained in type O infection is less than in comparable SAT infections. The original data also suggest that bovine host may be less able to sustain or create such high levels of variation so this may also limit the appearance of a sub consensus variation. This experiment does highlight that sampling more than one site of replication in one animal (foot lesion and probang) does not results in the pattern observed in buffalo. These two lesions are days apart during the infection and are not in close proximity to each other so should represent two distinct replication sites.

5.5 Summary

SAT's increased variability is host derived Samples isolated from buffalo have more high entropy positions than those isolated from cattle (Fig. 5.2). The differences seen between SATs and non-SAT serotypes (type O, type A, type C and Asia 1) can be attributed to differences between bovine derived and buffalo derived samples. The southern African serotypes (SAT1, SAT2 and SAT3) have previously been shown to be more variable than their European counterparts when comparing P1 [230]. This work shows that this may not be a virus specific feature as previously assumed but actually a product of host range.

Variability is not solely due to immune pressure Having established that buffalo derived swarms have more high entropy positions than their cattle derived counterparts, the regions of the genome with this plasticity was considered. Viruses under immune pressure are expected to produce mutations in their capsid coding region to allow for immune escape. This is not the case for the buffalo viral swarms. A higher percentage of the non-structural coding region had high entropy positions than the structural coding region (Fig. 5.5). This suggests that the production of this increased level of variability is not solely immune pressure as it is genome wide.

Buffalo swarms can obscure a subconsensus level populations Considering the relative nucleotide frequencies at high entropy positions revealed a consistent nucleotide split at several regions across the genome (Fig. 5.3). The consistency of this

split suggested a subconsensus majority variant. Bioinformatic haplotype reconstruction confirmed that one buffalo (SAT1/KNP/196/91) swarm contained a 18.3% variant that would not have been identified with current diagnostic techniques. This suggests increased variability in buffalo derived swarm is due to the presence of several viruses.

Subconsensus level variants can be of antigenic importance Having established two subconsensus level variants, the antigenic difference between the variants and consensus level genome were considered. Changes were found at known antigenic sites (Table 5.2) suggesting the swarm could contain subconsensus immune escape variants. These variants would not be identified by VP1 consensus level sequencing as currently used in diagnostic test.

Epidemiological differences seen in SAT3 could be a result of genetic differences This analysis revealed serotype specific differences between SAT3 and the other SAT serotypes. When considering relative nucleotide frequencies SAT3 showed regions of the genome with unique variability patterns (Fig. 5.3, 5.4). One explanation of this has been proposed as SAT3 having an increased ability for recombination. This could be the cause of the unique patterns observed. This would also result in decrease efficiency of purifying selection potentially explaining the different distribution of entropy scores. If SAT3's increased level of recombination decreases it's ability to undergo successful purifying selection it could be less fit and therefore result in less outbreaks.

In short, this work suggests that the reservoir host African buffalo may be subject to co-infection, thus resulting in more variation within the swarm across the entire genome. This reservoir of variation can hide antigenically important variants. Work should be completed on how buffalo maintain co-infection and if viruses transmitted from these animals produced a different disease outcome in subsequently infected cattle. Current diagnostic techniques may also need to change to include subconsensus level sequencing. This would allow for more accurate epidemiology and disease control strategies. Furthermore, SAT3 viruses show interesting genetic differences relative to SAT1 and SAT2 viruses that could be relevant to the reduced spread seen with SAT3. Work could be completed to establish if these difference are indeed related to the viruses ability to recombine and why this is a feature specific only to SAT3.

Chapter 6

Swarm dynamics during adaptive evolution

6.1 Abstract

Picornaviruses, like many positive strand RNA viruses, have error prone polymerases that lack proof reading capability [246]. This, coupled with short generation times and recombination, results in a diverse population of viruses that exist as a swarm. The reservoir of diversity that exists in this swarm allows for a viral infection to evolve quickly to adapt to new environmental challenges. FMDV's broad host range shows evidence of this ability to adapt to different cellular environments with nearly all *Artiodactyla* and *Camelids* susceptible to infection [18].

How variation within a viral swarm affects or allows for this broad host range has not been considered. To understand how the swarm allows FMDV to evolve in this scenario an experiment was designed to observe the dynamics of variants present in the swarm during viral adaptation to a new cell line. The historical model for *in vitro* study of FMDV involved using lab strains of virus adapted to baby hamster kidney (BHK) cells. A virus previously adapted to growth in a BHK cell line was serially passaged through a bovine foetal aorta (BFA) cell line. Time taken to achieve complete cytopathic effect (CPE) was observed as an indicator of viral growth rate and the virus was sequenced after each passage to identify genetic changes and consider swarm dynamics.

Passage through BFA cells produced decreased time to observe CPE and increased rate of replication. This was associated with four consensus level changes; one change in the internal ribosome entry site (IRES), one change in VP2, one change in VP3 and one change in 2C. All of these changes appeared after a drop in titre produced a potential

genetic bottle neck. A consistent amount of cumulative entropy was apparent through out the passages but there was variation in the proportion of genomes in high, mid and low entropy ranges. Two different distribution were apparent in viruses under adaptive pressure and those not.

6.2 Introduction

In vitro studies to consider FMDV adaptation to new hosts have found two main elements involved in viral adaptation; virus entry into cells and virus intracellular function. For cell entry FMDV has been shown to amend its capsid proteins to facilitate cell entry through different receptors. Within the cell, non-structural protein 3A and the internal ribosome entry site (IRES) have been shown to have an affect on tropism.

6.2.1 Entry adaptation

The presence or absence of receptors on the cell surface can determine if a particular FMDV can enter the cell. Therefore, many viral adaptations are found in the capsid proteins. The receptors used by field isolates of FMDV are RGD-dependent integrins [52–54, 247, 248]. FMD viruses interact with these integrins using an RGD motif located in the G-H loop of capsid protein VP1 [55]. Adaptations in cell culture can be a product of minimal changes in the capsid proteins. Generally a small number of residues will change increasing the overall positive charge of the virus. This allows for the use of negatively charged receptor molecules such as heparan sulphate. FMDV adaptation to use heparan has been studied with the analogue heprin sulphate. A structure of FMDV and heprin sulphate produced by Fry *et al* showed that instead of binding areas of the GH loop the heparin sulphate binding site is actually an indent in the capsid surface formed by VP1, 2 and 3 [249]. The central point of this depression (VP3 56) and surrounding positions (VP2 135 and VP1 195) have been shown to be key in adaptation to negatively charged receptors [50]. Cell culture adaptation has not been found solely at these residues. Viruses have also been identified with mutations at the fivefold axis of the capsid. Again these sites introduce positive charges [57]. Changes in this regions can also allow integrin dependant entry for viruses lacking the RGD motif although the mechanism for this is not fully elucidated [57]. Not all receptors used by FMDV are known. Viruses have been identified that use neither heparan sulphate or integrin receptors and work needs to be completed to identified these novel receptors [56].

6.2.2 Intracellular adaptation

6.2.2.1 IRES

Work by Sun *et al* has suggested that the internal ribosomal entry site could affect cell tropism. An FMDV mutant containing the IRES element from bovine rhinitis B virus (BRBV) replicated comparably to wildype virus in baby hamster kidney cells but showed decreased replication efficiency in three porcine cell lines (IBRS2, PK15 and SK6) [250]. Further dissection revealed that this effect was due to IRES domains 3 or 4. The IRES element has also been found to be important in poliovirus tropism [251–254].

6.2.2.2 Non-structural protein 3A

Non-structural protein 3A has previously been implicated in host range in rhinoviruses, enteroviruses and hepatoviruses. FMDV 3A is a partially conserved protein 153aa long. This is approximately 60aa longer than other 3A proteins in the picornavirus family such as poliovirus. It contains a highly conserved N-terminus containing hydrophobic and hydrophilic domains associated with membrane binding and a much less conserved C-terminus [75]. Deletions and mutations in this C-terminus have been associated with adaptation to host. For example, egg-adapted vaccine strains have been found to develop deletions of 19-20aa which produce viruses attenuated in cattle. An outbreak of FMDV in pigs in Taiwan showed a 10aa deletion in the C-terminus of 3A resulted in attenuation in cattle but disease in pigs [76, 77]. The effect of mutations on host range suggests this protein may interact with host factors. Gladue *et al* completed a yeast two-hybrid study which identified DCTN3 as a likely interaction factor. This is a subunit of the dynactin complex which has a role in dynein (a microtuble base motor) function. This suggests a potential role for intracellular organelle transport in viral replication.

6.3 Experimental Design

An experiment was designed to observe viral adaptation to a new cell line. A sample of FMDV O1Kaufbeuren (O1K) was recovered from infectious copy plasmid (ICP) pT7S3. This virus is cell culture adapted to baby hampster kidney (BHK) cells. This virus was expected to use heparan sulphate to enter the cells. As this sample was recovered from an ICP an original 'population expansion' step was included to try and ensure the swarm was not unrealistically clonal.

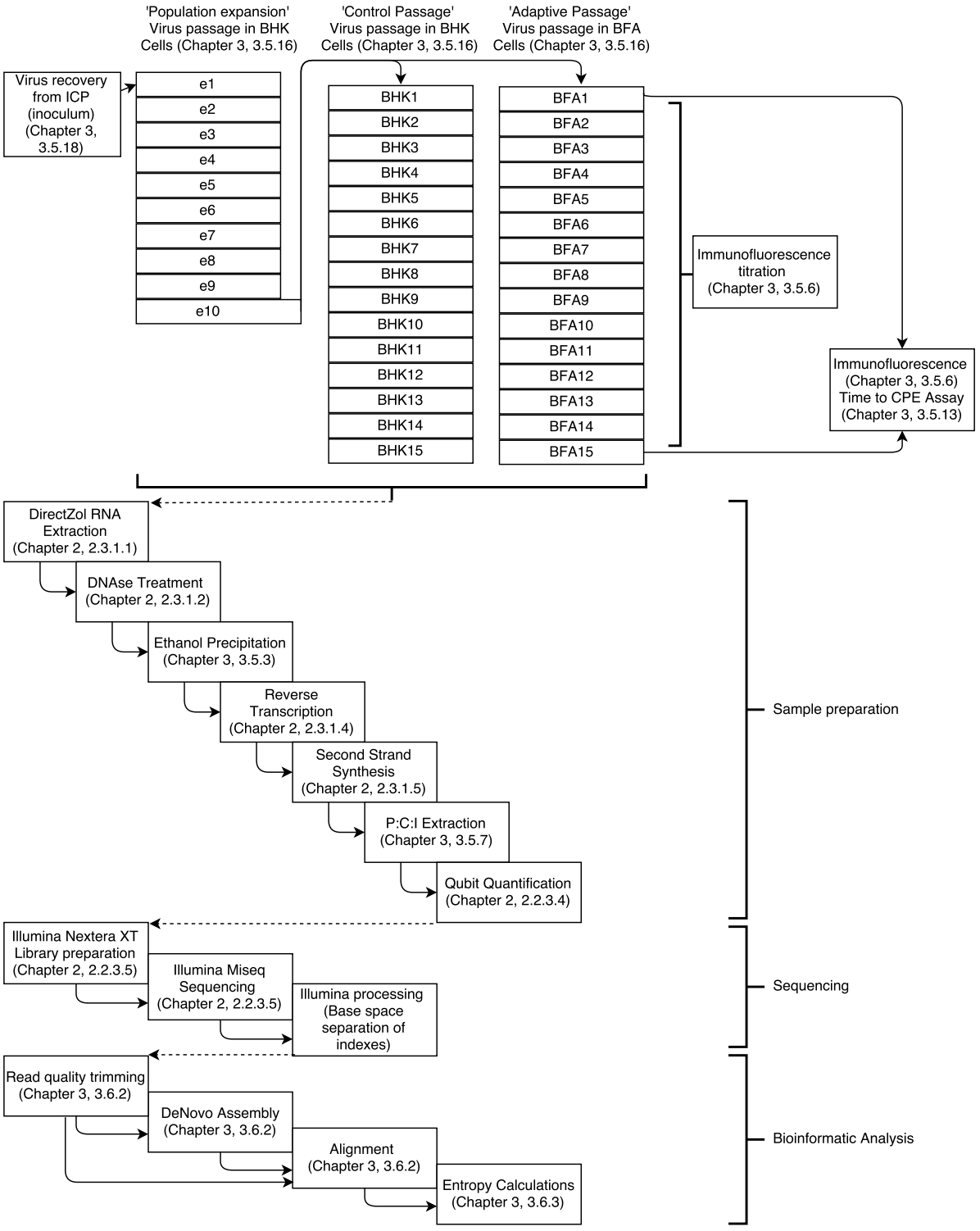


FIGURE 6.1: **Adaptive passage flowchart of methods used** FMDV O1K was recovered from an ICP and passaged 10x through BHK cells as a population expansion step (e1-10). e10 was then passaged through BFA cells in an attempt to promote adaptation and BHK cells as a control (BFA1-15 and BHK 1-15 respectively). All samples were prepared for sequencing as outlined in the published protocol [170] with some thesis specific amendments (Chapter 2). Library preparation was completed using the Illumina Nextera XT kit and subsequent sequencing completed on the Illumina MiSeq. Sequence analysis was completed as detailed in Chapter 3. DeNovo assembly was completed on e1 to provide a reference sequence. All subsequent sample's sequencing data was aligned to this reference. BFA samples were titrated using IF. BFA1 and 15 were compared directly on the MiniMax Cytometer.

To achieve this the recovered sample, referred to as 'inoculum', was blind passaged through BHK cells 10 times. These ten passages were called expansions 1 to expansion 10 (e1-e10). e10 was then passaged through bovine foetal aorta cells (BFA) fifteen times to observed if the virus would adapt. Concurrently it was passaged through BHK cells fifteen times as a control. These samples were referred to as BFA1-BFA15 and BHK1-BHK15 respectively. The viruses were titred and an equal MOI infection of BFA1 and BFA15 in BFAs was completed on the SpectraMax MiniMax 300 Imaging Cytometer to consider time to achieve cytopathic effect. Each sample was also sequenced on the Illumina Miseq to identify any genetic changes that could be responsible for phenotypic changes observed and to consider the dynamics of a viral swarm undergoing adaptation. Methods used in this chapter are outlined in Figure 6.1. Further detail on methods is included in Chapters 2 and 3.

6.4 Results and Discussion

6.4.1 Adaptive passage produces a phenotype change

6.4.1.1 Optimisation of Immunofluorescence titration

To allow for equivalent amounts of virus to be compared the cell lysate from each passage needed to be titred in BFAs. To consider the time point at which a titration should be completed, BFA adaptive passage one (BFA1) and fifteen (BFA15) were compared at three different time points (4hpi, 5hpi and 6hpi) in BFA and BHK cells. The number of cells were counted by ToPro3 staining.

In BFA cells unpaired t-tests showed a significant difference between the control and wells infected with BFA15 at 4hpi infection ($p=0.347$) but no significant difference between the control and BFA1 ($p=0.2167$). There was no significant difference between the control and BFA1 or BFA15 at five hours post infection in BFAs ($p=0.9517$ and $p=0.1811$ respectively). There was a significant difference between the control and BFA1 and the control and BFA15 at six hours post infection in BFAs ($p=0.0003$ and $p=0.0002$ respectively.). In BHK cells there was a significant difference between the control and monolayers infected with BFA15 at 4hpi ($p=0.0165$) but not between the control and BFA1 ($p=0.2167$). At 5 and 6hpi infection in BHK there was a statistically significant difference in the number of cells between the control and BFA1 ($p=0.002$, $p=<0.0001$) and the control and BFA15 ($p=<0.0001$, $p=<0.0001$) From this it can be see that the number of cells in infected wells is significantly different than the uninfected control by 6hpi infection for both BFA and BHK cells. This suggests that by this point the virus

has started to kill the cells. This time point was therefore discounted for titration. This is also the case for five hours post infection in BHK cells. However, at 5hpi in BFA cells there is no statistically significant difference in the number of cells in infected wells when compared to the control (Fig.6.2). This indicates that both viruses (BFA1 and BFA15) cause delayed cell death in BFA cells compared to BHK cells.

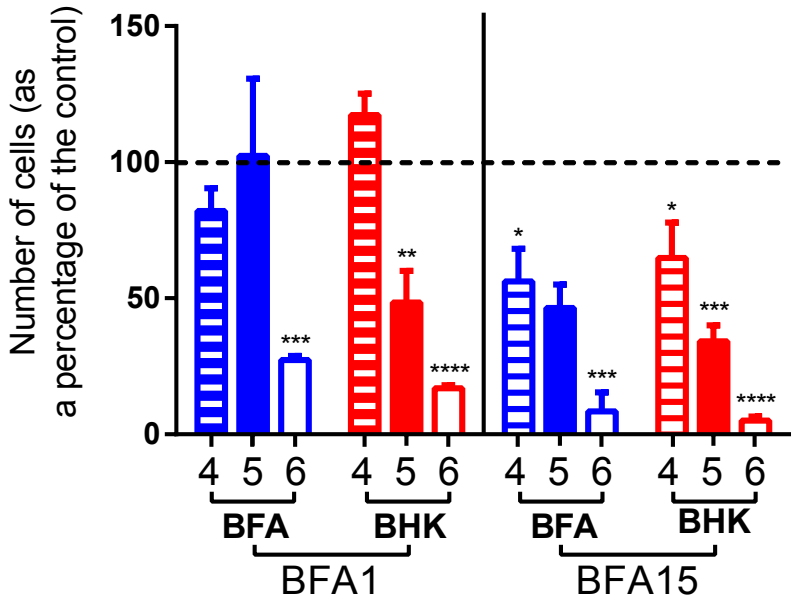


FIGURE 6.2: **CPE in virus infected cell monolayers** Cell monolayers of BFA (blue) and BHK (red) cells infected with cell lysate from BFA1 and BFA15. Monolayers were fixed at 4 (horizontal stripe), 5 (solid fill) and 6 (no fill) hours post infection and stained with ToPro3. Cells were the counted using the SpectraMax MiniMax 300 Imaging Cytometer. Number of cells was normalised in comparison to an uninfected cell monolayer. The un-infected cell monolayer was considered 100% and the cell number for each infected monolayer is shown as a percentage of that.

This variation from timepoint to timepoint creates a challenge in determining the optimum time to calculate titre as calculations at different time points would produce a different titre. The time point of 6hpi was deemed too late due to extensive cell death and at 4hpi it was decided it was likely that the full extent of infection would not be identified therefore 5hpi was chosen as the time at which to titre the viruses.

6.4.1.2 BFA15's replication cycle appears faster than BFA1

Although the experiment described above was designed as an optimisation step it is also interesting to note that for BFA15 in BFA and BHK cells there are less cells in comparison to the control suggesting some cell death before 4hpi. Furthermore, in

BFA1 in BFA cells there is no decrease in cell number until 6hpi. In comparison BFA15 in BFAs show cell death by 4hpi that continues to decrease at 5 and 6hpi. This would suggest that BFA15 is causing CPE in BFA cells faster than BFA1 and has therefore successfully adapted. To consider the number of infected cells causing this monolayer degradation (to insure CPE was not solely due to higher virus titre) immunofluorescence against FMDV non-structural protein 3A was performed.

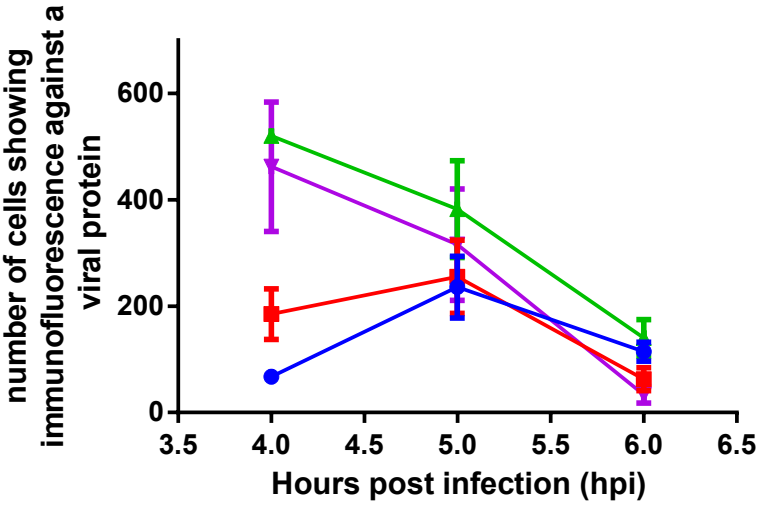


FIGURE 6.3: **Number of infected cells** Cell monolayers were infected with adaptive passage one (BFA1) and adaptive passage fifteen (BFA15) and fixed at 4, 5 and 6hpi. The number of infected cells was calculated using IF against non-structural protein 3A. The infected cell counts for passage 1 in BFAs (blue circle) and BHKs (green triangle) and adaptive passage fifteen in BFAs (red square) and BHKs (purple triangle) are shown with standard deviation.

This showed that the peak number of cells infected in BHKs was at 4 hours post infection. There was only one replicate for BFA1 at this time point in BHKs (due to a pipetting error). For all other samples three replicates were available. For the BFA1 sample in BHKs 505 cells were infected. For BFA15 in BHKs at 4hpi an average of 462 cells were infected. This number decreases by five hours post infection (382 and 315 cells on average respectively) and again by six hours post infection (140 and 33 respectively). For BFA cells more cells were identified as infected at five hour post infection than 4 or 6hpi. At this time point an average of 382 cells were infected for BFA1 and 315 cells for BFA15. BFA1 and BFA15 in BHK cells produced a similar number of cells identified as infected at four and five hours post infection. This is not the case for viruses in BFAs. Although comparable at five hours post infection, at four hours post infection less cells are identified as infected in passage one than in passage fifteen (67.3 on average compared to 185 respectively)(Fig. 6.3). It is possible that less cells were identified in BFA1 at four hours post infection not because less cells were infected but because the

cells had not completed a sufficient amount of replication and translation to be identified by IF at this time point. More cells were identified in BFA15 at 4hpi, a similar number at 5hpi (both at their maximum replicative rate) and less at 6hpi (as cells start to die). This could suggest that BFA15's replication cycle is ahead of BFA1.

This work was completed on three separate plates as samples needed to be fixed at different timepoints. The plates were seeded with the same cells at the same density and infected with the same virus stock in order to make the results as comparable as possible. The difference observed were therefore assumed not to be due to difference in set up.

6.4.1.3 Viruses in BFA15 have higher replication efficiency than those in BFA1

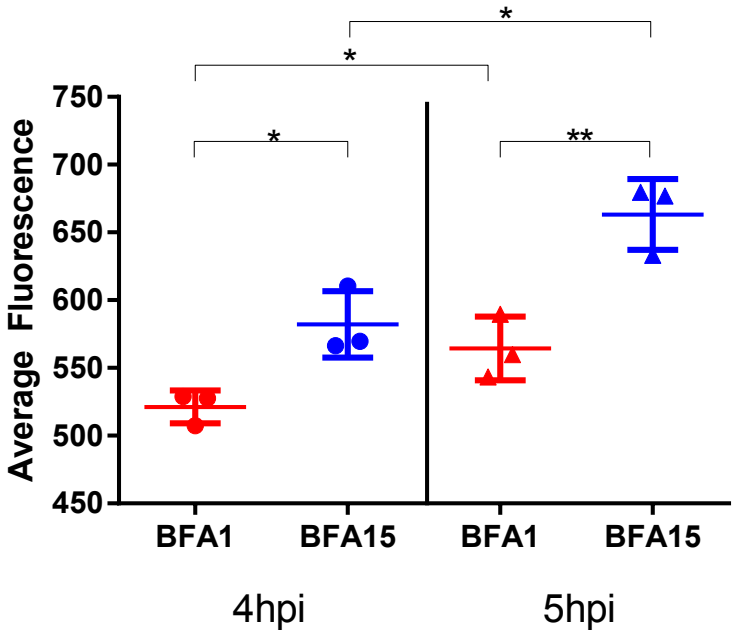


FIGURE 6.4: **Average fluorescence of infected cells as an indicator of replication** The average fluorescence of infected cells (as determined by IF) was calculated for adaptive passage one (red) and adaptive passage fifteen (blue) at four hours (circle) and five hours (triangle) post infection. There was a statistically significant difference between average fluorescence of adaptive passage one at 4 and five hours post infection (unpaired t-test, p=0.0475). There was also a statistically significant difference between adaptive passage fifteen at 4 and 5 hours post infection (unpaired t-test, p=0.0171). There was a significant difference between adaptive passage one and fifteen at both 4 and 5 hours (unpaired t-test, p=0.0181 and p=0.0081 respectively)

To observe if the replication/translation efficiency differed the average fluorescence of each infected item was compared at four and five hours post infection. The average fluorescence was lower in BFA1 than BFA15 at both four and five hours although it has increased by five hours post infection. The level of fluorescence for BFA1 at 5hpi is similar to the level of fluorescence BFA15 at 4hpi (Fig. 6.4). These results suggest BFA15 is replicating more efficiently than BFA1. However, IF is indicative of translation efficiency not replication efficiency. Further work could be done to complete qPCR comparisons of these samples as a more direct indicator of replication.

6.4.1.4 There is a drop in titre at BFA5

Having identified 5 hour post infection as an appropriate time to complete IF to determine titre, a cell monolayer was infected with 100 μL of diluted lysate (10^{-1}) in triplicate. This was left for 1 hour before the supernatant was removed and replaced with VGM. Five hours after the original infection the cell monolayers were fixed. Immunofluorescence against FMDV 3A was performed. The number of infected cells were calculated from this.

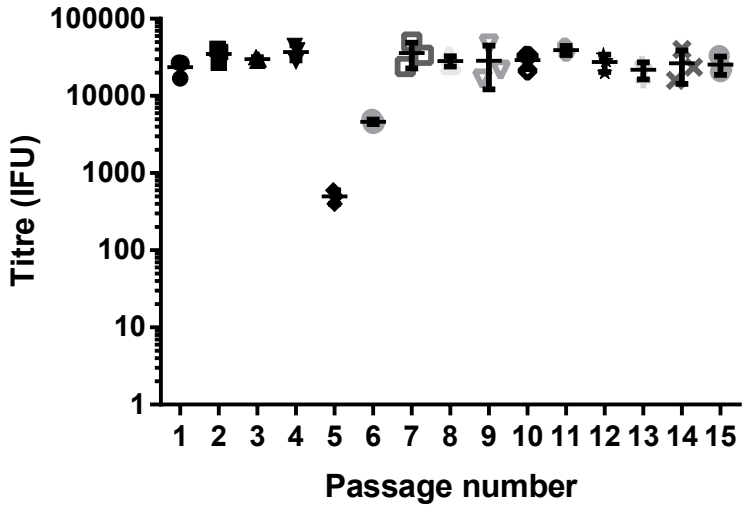


FIGURE 6.5: **Titre of lysate from adaptive passage series** The number of infectious units per mL was calculated by infecting a cell monolayer of BFA cells with 10^{-1} of each cell lysate. The cells were fixed at 5hpi and IF labelled for non-structural protein 3A. The number of infected cells was counted and adjusted to represent the number of infectious particles per mL.

There is a dramatic drop in titre at BFAF5 (drop of 1 log from the previous passage; 3.7 x 10^4 to 1.5 x 10^3). This is recovered partially in BFA6 (4.6x10^3) and fully by BFA7

(3.58×10^4) (Fig. 6.5). The cause of this is unknown. Due to the process of blind passage, titre transferred to the next passage is not controlled for (Fig. 6.5).

6.4.1.5 BFA15 is slightly more efficient at destroying the cell monolayer

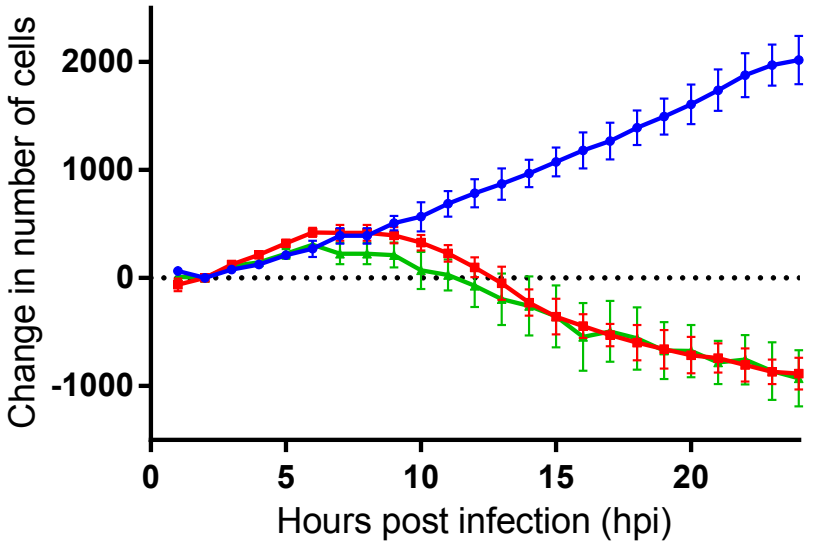


FIGURE 6.6: **BFA1 and BFA15 cell death over time** The change in the number of cells in a cell monolayer is shown for a control uninfected BFA wells (blue, circle), BFA cells infected with adaptive passage one (BFA1)(red, square) and BFA cells infected with adaptive passage fifteen (BFA15)(green triangle). Each value is plotted with standard deviation

To compare the time BFA1 and BFA15 take to cause CPE an experiment was designed on the SpectraMax MiniMax 300 Imaging Cytometer to measure time to observed CPE in an equal MOI infection. This titre data was used to produce an equal MOI infection of BFA1 and BFA15 in BFAs. The number of cells infected were compared using IF at five hours to ensure the samples were correctly calculated as comparable. There was no significant difference between the number of cells infected (unpaired t-test, p=0.2222). Just under 10% of the cells in each monolayer were infected. Both infections produced full CPE within 24 hours. When observing the decrease in the number of cells over time it can be seen that both infections mimic the control (with the number of cells increasing) until 6 hours post infection. At this point passage one appears to plateau and passage 15 decreases and plateaus. By ten hours post infection both experiments are showing a decrease in the number of cells infected although the difference between passage 15 and the control is greater than passage one and the control. By 15 hours post infection both experiments are directly comparable. This represents complete CPE

in each experiment although the number of cells continued to decrease slowly as cells detach and are no longer counted by the machine (Fig. 6.6).

BFA15 appears to be destroying the monolayer slightly more efficiently than BFA1 in an equal MOI infection although the difference is not extensive. This data does not correlate with what has previously been seen which showed a more notable change in phenotype. This may be because the MOI was too high resulting in complete CPE in under 24 hours. Due to the speed of cell death it may be that the difference between the viruses were not apparent. Previous work has found that to produce an infection in BFA cells a higher quantity of virus needs to be added (more than, for example, BHK cells)(data not shown). This may be due to the cells innate immunity having an effect on infection, a concept that needs to be further investigated.

6.4.2 Virus adapted to BFA cells shows consensus level changes.

During the 15 passages completed in BFA cells there were seven consensus level changes. Of these, three genome positions showed nucleotide fluctuation in the control passage also and four were uniquely identified in the BFA passage with no variation apparent at these positions in the control passage (Table 6.1).

Genome position	Coding region	Nucleotide change	Amino acid
1056	5'UTR	G to A	n/a
2651	VP3	G to C	Serine to Threonine
3916	VP1	A to C	Lysine to Glutamine
5180	2C	T to C	Isoleucine to Asparagine and Threonine
5203	2C	G to A	Alanine to Threonine
5681	3A	C to T	Alanine to Valine
6411	3D	C to T	Valine to Valine

TABLE 6.1: **Consensus level changes in BFA adaptation passage.** Positions in red showed nucleotide fluctuation of more than 5% during the control passage series. (That is to say, during the control passage the majority nucleotide at that position went below 95 percent)

6.4.2.1 Three consensus level changes in the adaptive passage also showed plasticity in the control

Of the three changes identified both in the control passage and the BFA adaptation, one was synonymous (nt 6411) and two were non-synonymous (nt5203 and nt5681)(Table

6.1). Synonymous changes are often ignored as they have no affect on the amino acid code. However, it should be noted that in some cases synonymous changes can have a relevant affect on RNA secondary structure. The synonymous change in this instance is consensus level by e3 (Fig. 6.7 C). This transition of C to T in the adaptive passage reaches almost 100% in comparison to the control where the two variants are maintained a 60/40 split for the majority of the passage slowly reaching an 80/20 split.

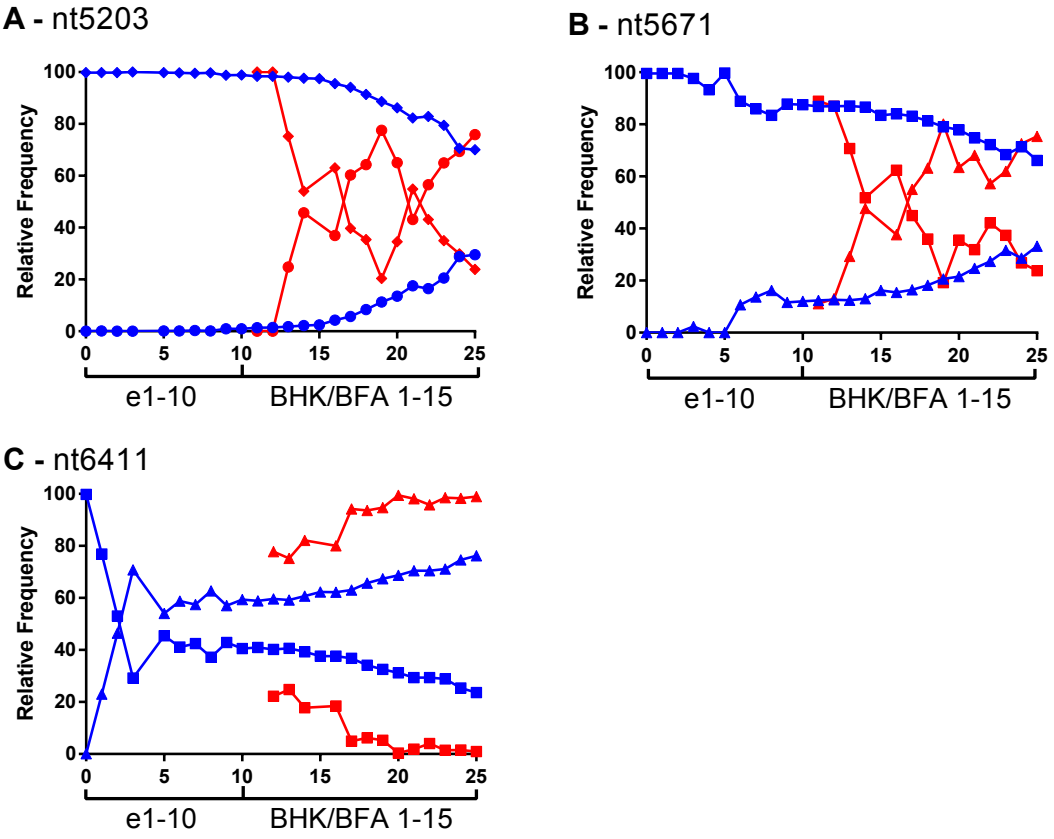


FIGURE 6.7: **Three sequence changes that became fixed in the BFA adaptive passage that showed flexibility in the control passage.** The relative frequency of the fluctuating nucleotides (y axis) is plotted for each passage (x axis) for the BFA adaptive passage (red) and the BHK control passage (blue) at positions with a coverage >10x. This includes the expansion of the original BHK passage (p0-10). Fluctuating nucleotides in the BFA passage and the corresponding genome position in the BHK is plotted for positions A) 5203 B) 5681 C) 6411. The changing bases are distinguished by the marker shape: Adenine (circle), Cytosine (square), Thymine (triangle) or Guanine (diamond).

This split population suggests both variants are viable but the original change suggests a thymine at this positions is preferable. This synonymous change at position 6411 is in the coding region for 3D. The non-synonymous changes at genome positions 5203 and 5671 are in 2C and 3A and equated to an alanine to threonine or valine substitution respectively. Both alanine and the substituted amino acids are neutral although it should

be noted that threonine has a polar side chain (in comparison to the original non-polar alanine and other change, valine).

With variability in both the control passage and the adaptive passage it is possible these positions in the genome might be less conserved. This correlates with the findings of Wang et al [255] in poliovirus. They identified a region tolerant to variation down stream of Motif C between amino acids 252 and 263. This region in poliovirus correlates with the position of the non-synonymous change found in 2C that showed flexibility in both the control and adaptive passage in this study (nt5203, Table 6.1). This could explain the slightly more erratic distribution of nucleotides in the BFA passage (Fig. 6.7), a pattern that is only seen in these three changes (Fig. 6.8).

6.4.2.2 Four consensus level changes showed no plasticity in the control

Of the four consensus level changes at genome positions with little to no variation in the control one position was non-coding (nt1056) and the other three were non-synonymous (nt2651, nt3916 and nt5180)(Table 6.1).

Each of the four changes observed, that were unique to the adaptive passage series, appeared at BFA5. This conincided with a large drop in titre. For nt1056, 2651 and 3916 the variant slowly increases from BFA5 and becomes fixed in the populations by BFA14. The proportion of the variant continues to rise from this point and it is possible with further passages it would reach 100% (Fig. 6.8A, B, C). The pattern for nt 5180 differs slightly. The variant appears in the population at BFA5. By BFA8 the majority variant is present at approximately 30% with a second variant population also present at 20%. This plateaus until BFAF12 where the minor variants begins to decrease as does the original majority and the majority variant increases and is fixed (Fig. 6.8D).

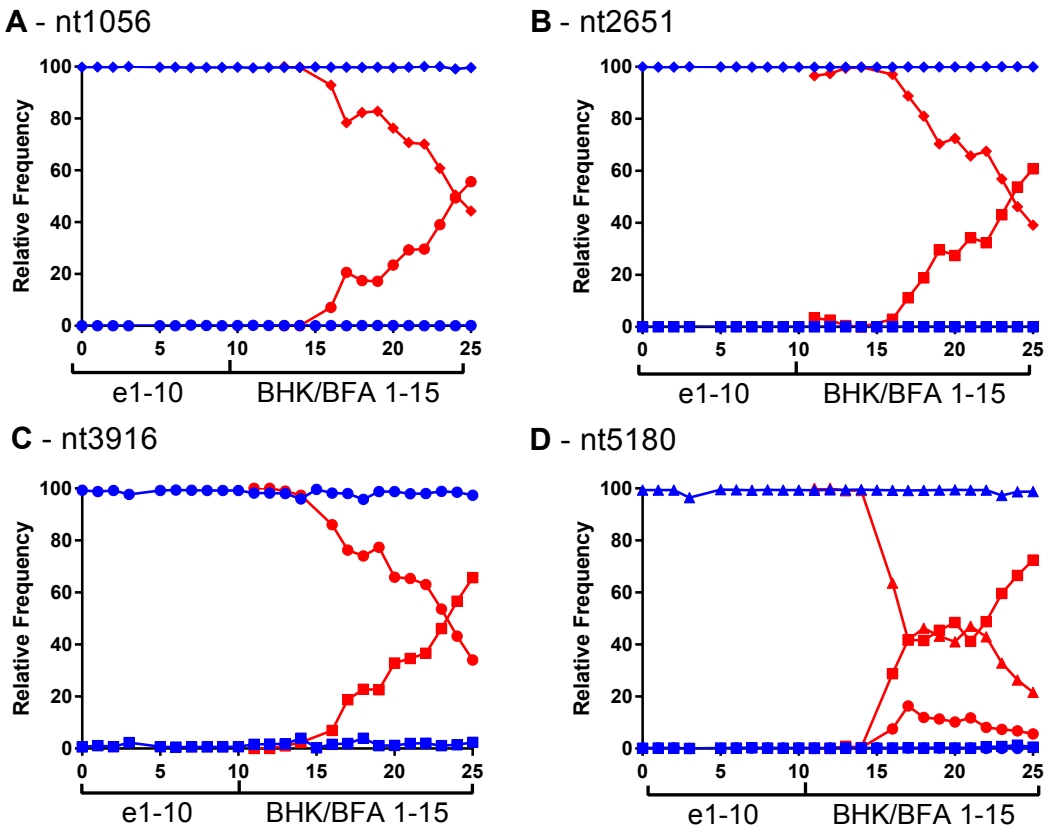


FIGURE 6.8: **Four sequence changes became fixed in the BFA adaptive passage that showed little to no variability in the control.** The relative frequency of the fluctuating nucleotides (y axis) is plotted for each passage (x axis) for the BFA adaptive passage (red) and the BHK control passage (blue) at positions with a coverage >10x. This includes the expansion of the original BHK passage (p0-10). Fluctuating nucleotides in the BFA passage and the corresponding genome position in the BHK is plotted for positions A) 1056 B) 2651 C) 3916 D) 5180. The changing bases are distinguished by the marker shape: Adenine (circle), Cytosine (square), Thymine (triangle) or Guanine (diamond).

6.4.2.3 Change in FMDV IRES could be associated with eIF4G binding

The non-coding change (G to A) was at genome positions 1056 which is situated within the 5'UTR. This nucleotide maps to the internal ribosome entry site (IRES) of FMDV. It lies within stem loop 4 at the 3' end of the adenine rich nucleotide bulge associated with eIF4G binding [88, 256](Fig. 6.9). This is a feature that is completely conserved and essential in another picornavirus, EMCV [257]. eIF4G plays a key role in internal initiation as well as CAP dependant cellular translation. The N-terminal region associated with the latter is cleaved by FMDV Lpro and 3C to result in a truncated protein

that no longer functions in its role for cellular translation. eIF4G contains a core sequence of approximately three hundred amino acids between aa745/772 and aa941/949 that specifically bind the IRES [258]. When comparing the amino acid sequence for this region in bovines and hamsters it can be seen there are some differences (Fig. 6.9B) which could explain the associated change in the IRES sequence when passaging a hamster cell line adapted virus into a bovine cell line. Equally the amino acid change results one more adenine at the end of a poly a chain of 5, potentially adding further stability to the interaction.

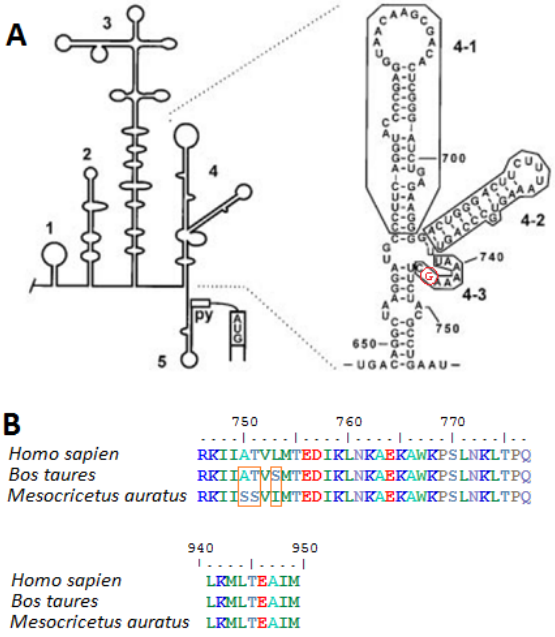


FIGURE 6.9: **FMDV IRES loop 4 and associated species specific binding sequence of eIF4G** A) The secondary structure of the FMDV IRES is shown (adapted from Saleh *et al* [88]) with the sequence of loop 4. Position 744 in loop 4-3 is the residue changed in this experiment. B) The binding site for eIF4G's interaction with the IRES is shown for human, cattle and hamster and differences between them highlighted.

6.4.2.4 Two changes are in the viral capsid

The three changes in the coding region of the genome were in the capsid proteins VP3 (nt2651) and VP1 (nt3916) as well as one change in the non-structural protein 2C (nt5180). The amino acid change in VP3 was from the neutral serine to the also neutral threonine whilst the change in VP1 was from the positively charged lysine to neutral glutamine. The change in VP1 is surface exposed whereas the change in VP3 is mostly obscured, sitting close to the fivefold axis (Fig. 6.10). Cell culture adaptation often results in the addition of positive charges. As this is adaptation from cell culture to bovine cells it could be the opposite (although this is unlikely as the ability to bind

heparan does not prevent binding of integrin). Previously changes on the five fold axis have been identified as involved in adaptation to new receptors [57] although not in these positions (VP3 74 and VP1 629).

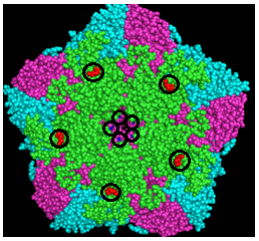


FIGURE 6.10: **Consensus level capsid changes mapped on a pentamer** Changes in the structural capsid proteins map to one surface exposed position in VP1 and one positon on the fivefold axis in VP3. Non-structural changes were mapped on to an FMDV O pentamer using pymol. The change in VP1 is indicated in red and the change in VP3 is indicated in blue.

6.4.2.5 Changes in the ORF could affect autophagosome formation

The mutation in the non-structural protein 2C is at nucleotide positions 5180. This non-synonymous substitution results in a change from isoleucine to a mix of threonine and asparagine although it is the threonine mutation that becomes fixed in the consensus sequence. Isoleucine is non-polar whilst both threonine and asparagine are polar. This change is in a multifunctional domain of 2C (Fig. 6.11).

This region has been found to be involved in ATPase activity (yellow). 2C has been defined as a hexameric AAA+ protein that is involved in ATP hydrolysis [72]. It requires all three of the labelled domains to be functional (walker A, walker B and 3) and is relatively specific for ATP in comparison to GTP, CTP, UTP which show a 10 fold lower level of hydrolysis. The mutations sit in an area associated with ATPase but not one of the essential binding domains. This suggests that potentially one amino acid change would not result in sufficient amendment to protein structure to have a notable effect. The region in which the variant mutation was identified has also been associated with Beclin 1 binding activity (Fig. 6.11(orange))[73]. Work by Gladue *et al* found that mutations in the Beclin binding regions identified resulted in non-viable virus and they suggested that this interaction is therefore essential for replication. They do acknowledge however that they cannot rule out the possibility that their alanine scanning mutagenesis affect the structure of the protein and thus disrupted one of its many other functions.

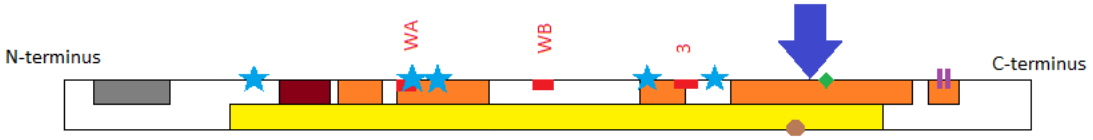


FIGURE 6.11: **A schematic of the known functional elements each region of 2C codes for** FMDV 2C includes a predicted amphipathic helix at the N-terminus (grey), ATPase binding activity (yellow) and beclin binding activity (orange). It is a hexameric AAA+ protein with a co-ordinated ATP hydrolysis mechanism requiring the Walker A, B and site three domains (red). It has also been found to interact with the cellular component vimentin (maroon). Various 2C mutants have been characterised in poliovirus including Hydantoin resistant mutants (light blue stars), uncoating defective mutants (purple) and 2C packaging mutants (brown circle). An interaction has also been found in CAV three between 2C and VP3 (green). The 2C mutation found only in the BFA adaptation and not in the control of this experiment is marked on the diagram with the blue arrow.

6.4.2.6 There is evidence of these mutations in WT viruses

To confirm the possibility of these mutations occurring in WT samples, an alignment of all currently published full genome sequences was completed and the presence of these mutations investigated.

Mutation:	5'UTR	VP3 2	VP1	2C
Bovine host	14	1		5
Ovine host	1			1
Buffalo		1	1	
BHK-21 passage				2
Other	1 (porcine)		1(murine)	
Unknown	4	11		12

TABLE 6.2: **Presence of mutations in published sequences.** An alignment of all currently published FMDV genomes was completed and the presence of each mutant in published sequences identified. The host from which the sample was taken was considered

From this it was found that the mutations in the IRES and the mutations in 2C were present in 20 sequences (4%) of the currently published sequences. For the IRES mutation 70% of the changes were derived from bovine hosts supporting the hypothesis that this is an adaptation to bovine cells (Table 6.2). These regions are relatively conserved and therefore it is easy to determine the presence of a variant. The same analysis is more complicated in the variable capsid proteins. The mutation in VP3 resulted in a C at the position nt2651, this is true of 24% of the published sequences. However, in our

WT virus the base after this positions (nt2652) is a C, so the mutation results in GC to CC. The CC motif is only present in 2.5% of the genomes; other genomes with C at position 2652 had a variety of bases subsequent to them although the majority had CG (17%). For the mutations in VP1 only two published sequences showed this mutation (0.3%) although there was a great amount of variation in the vicinity.

6.4.3 Swarms dynamics during adaptive passage

Having considered the consensus level changes, and associated phenotype, the dynamics of the swarm in achieving this adaptation were investigated. Although each passage was sequenced not all passages had sufficient coverage to be used in this analysis. This is expanded on in Appendix C.

6.4.3.1 Cumulative entropy was comparable between expansion, adaptation and control passages

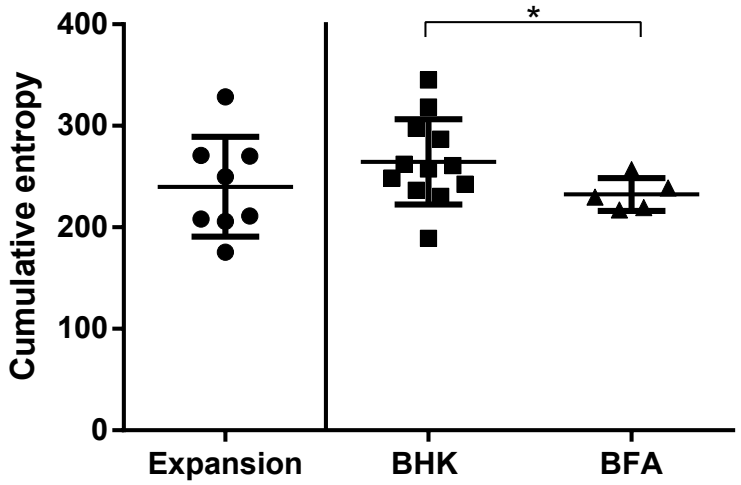


FIGURE 6.12: **Cummulative entropy of the population expansion, BFA adaptation and associated control** Shannon's entropy was calculated at each position of the genome and summed to produce a cumulative entropy score.

Shannon's entropy was calculated at each position along each genome as a measure of variation within the swarm. These values were added together to give a cumulative entropy score. Cumulative entropy was calculated using the region of the L-fragment represented at >1000x coverage in all compared genomes (nt646-8118). The L-fragment represents the entirety of the ORF and is usually well represented as it does not include the beginning of the genome or poly-c fragment, both of which are challenging

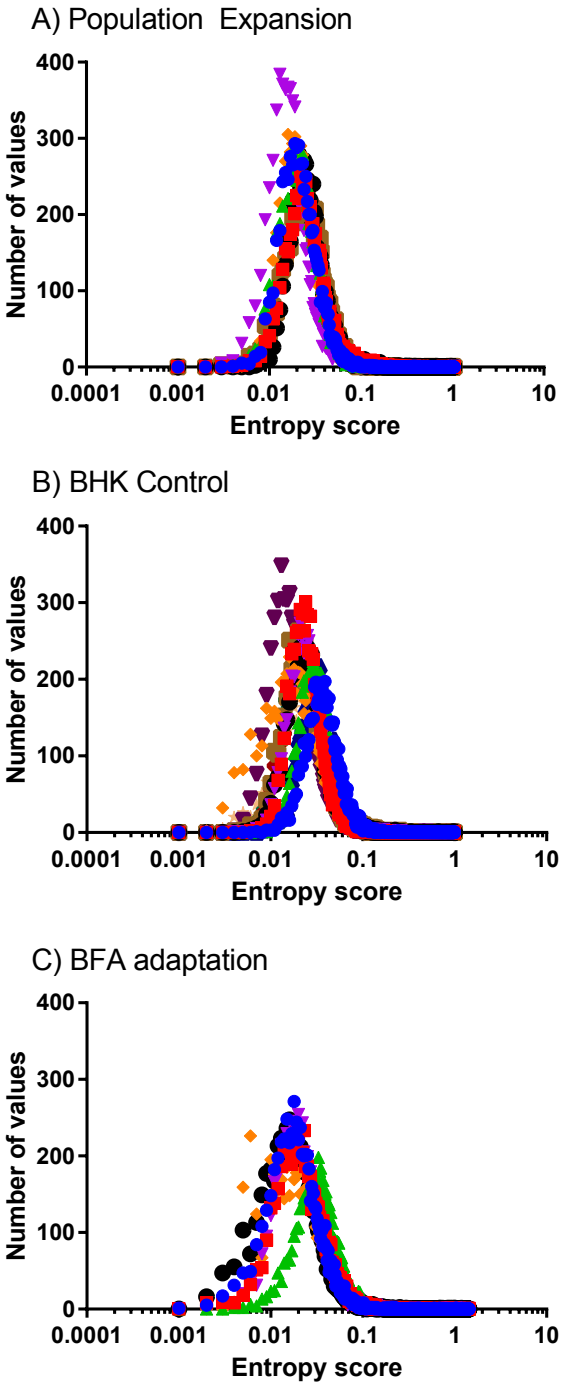


FIGURE 6.13: **The majority of genome positions have low entropy positions** Entropy scores were sorted into bins of 0.001. The number of positions (y-axis) is shown against the bin centre (x-axis)(log) for each passage of A) the population expansion (Inoculum - blue circle, E1 - red square, E2 - green triangle, E5 - purple triangle, E6 - orange diamond, E8 - black circle, E9 - brown square) B) the BHK control (H1 - blue circle, H2 - red square, H3 - green triangle, H4 - purple triangle, H5 - orange diamond, H6 - black circle,H7 - brown square, H8 - navy triangle, H9 - purple triangle, H10 - maroon square, H11 - dark green asterisk, H15 - beige star) and C) the adaptive passage (F4 - blue circle, F8 - red square, F10 - green triangle, F12 - purple triangle, F13 - orange diamond, F14 - black circle).

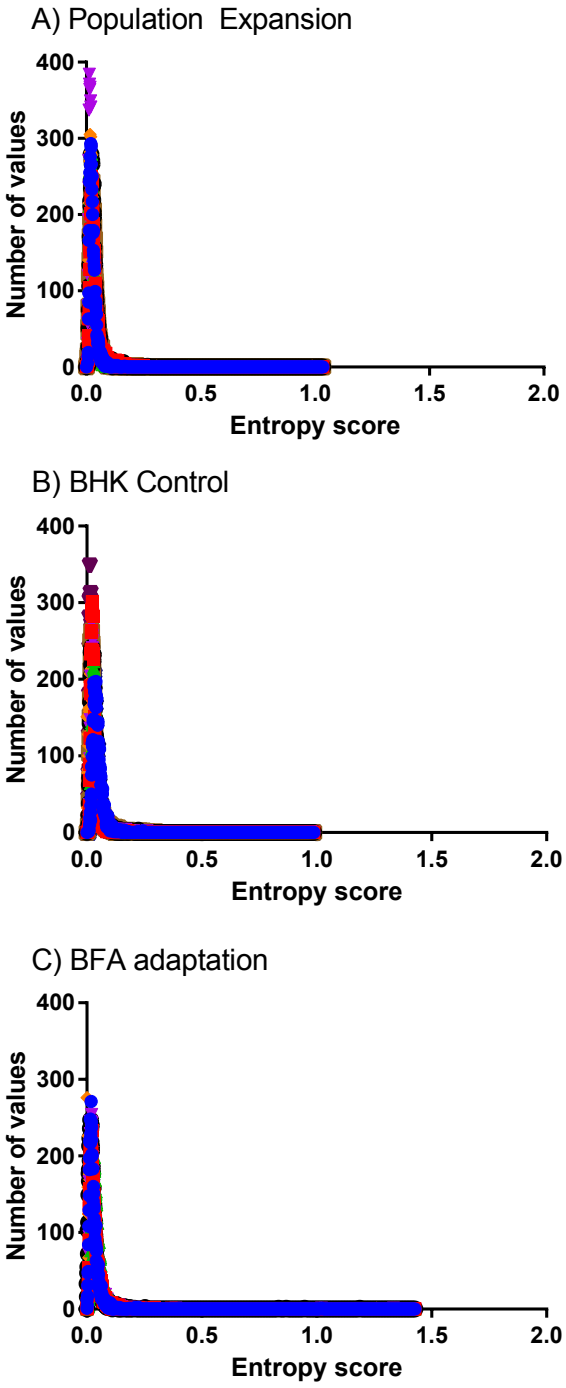


FIGURE 6.14: **The majority of genome positions have low entropy positions** Entropy scores were sorted into bins of 0.001. The number of positions (y-axis) is shown against the bin centre (x-axis) for each passage of A) the population expansion (Inoculum - blue circle, E1 - red square, E2 - green triangle, E5 - purple triangle, E6 - orange diamond, E8 - black circle, E9 - brown square) B) the BHK control (H1 - blue circle, H2 - red square, H3 - green triangle, H4 - purple triangle, H5 - orange diamond, H6 - black circle,H7 - brown square, H8 - navy triangle, H9 - purple triangle, H10 - maroon square, H11 - dark green asterisk, H15 - beige star) and C) the adaptive passage (F4 - blue circle, F8 - red square, F10 - green triangle, F12 - purple triangle, F13 - orange diamond, F14 - black circle). This graph is the same as Fig. 6.13 without the log scale.

to sequence. Cumulative entropy for the BHK population expansion appears to fluctuate with an average of 240.14 and a range of 153.23. The following control passage in BHK cells had a similar average of 264.61 and an almost identical range of 155.95. The adaptive passage in BFAs have a similar but slightly lower average (232.47) and a much smaller range of 39.58. There was no statistically significant difference between the expansion and the BHK control passage or the expansion and the BFA adaptation passage (Mann-Whitney p=0.3328, p=0.9293 respectively). There is a statistically significant difference between the BHK control passage and the BFA adaptation passage (Mann-Whitney, p=0.0469) (Fig. 6.12). It should be noted however that less of the BFA passage series had sufficient coverage for analysis (5/15) than the BHK control (12/15) and the population expansion (8/10). This could be the reason for the comparably small range in values.

6.4.3.2 The majority of genome positions in each passage have a low entropy score but this majority is smaller in the adaptive passage

Although the cumulative entropy of each passage was comparable, the distribution of the entropy scores resulting in this value may not be. For example, if four positions had an entropy score of 0.25 their cumulative entropy would be 1. If three positions had entropy scores of zero and one an entropy score of 1 this region would also have an cumulative entropy of 1.

Distribution of entropy scores was therefore considered in a histogram. It can been seen from this analysis that the distributions of entropy scores does differ slightly. Each set of passages had a peak bin of low entropy values. That is to say, that for each passage in the expansion, the control and the BFA adaptation the majority of genome positions had low entropy scores. On average, the number of positions in the majority bin for the expansion was 287.37 (range 139) for the control 250.58 (range 151) and for the adaptation 250.42 (range 43). Although there was some variation in the expansion and control in this figure the variation between the adaptive passages was low (range 43). When considering the high entropy scores, the tail of the histogram, the population expansion and BHK control have no positions with a score over 1.04 or 0.97 respectively. In comparison, every passage of the BFA adaptation has positions with a score >1 ((Fig. 6.14).

Passages with a higher number of positions within their majority bin have a lower entropy score as the majority. This can be seen most clearly in e5 (Fig. 6.13 A) purple triangles) and BHK9 (Fig. 6.13 B) purple triangles. These peaks are tall and thin. There is a statistically significant correlation between the majority bin and the values within this

bin (p=0.0018)(i.e. passages with a lower majority score have more genome positions in the bin). This gives rise the the idea of two different distributions of entropy scores: low majority with a smaller distribution and a higher majority with a more broader distribution.

This pattern appears to be cyclical in the BHK control. BHK1 shows a higher majority bin with less values in it, BHK2 shows a lower entropy bin with more values in it, BHK3 shows a higher majority bin score with less in it and BHK4 shows a lower majority bin with more positions in it (Fig. 6.15). It could not be observed if this pattern continued due to a drop in titre potentially resulting in a genetic bottle neck. This is evident in e5 having a much lower entropy majority bin with more positions in it (Fig. 6.13A, purple triangle). This is expected as the smaller the population the lower the capacity for variation.

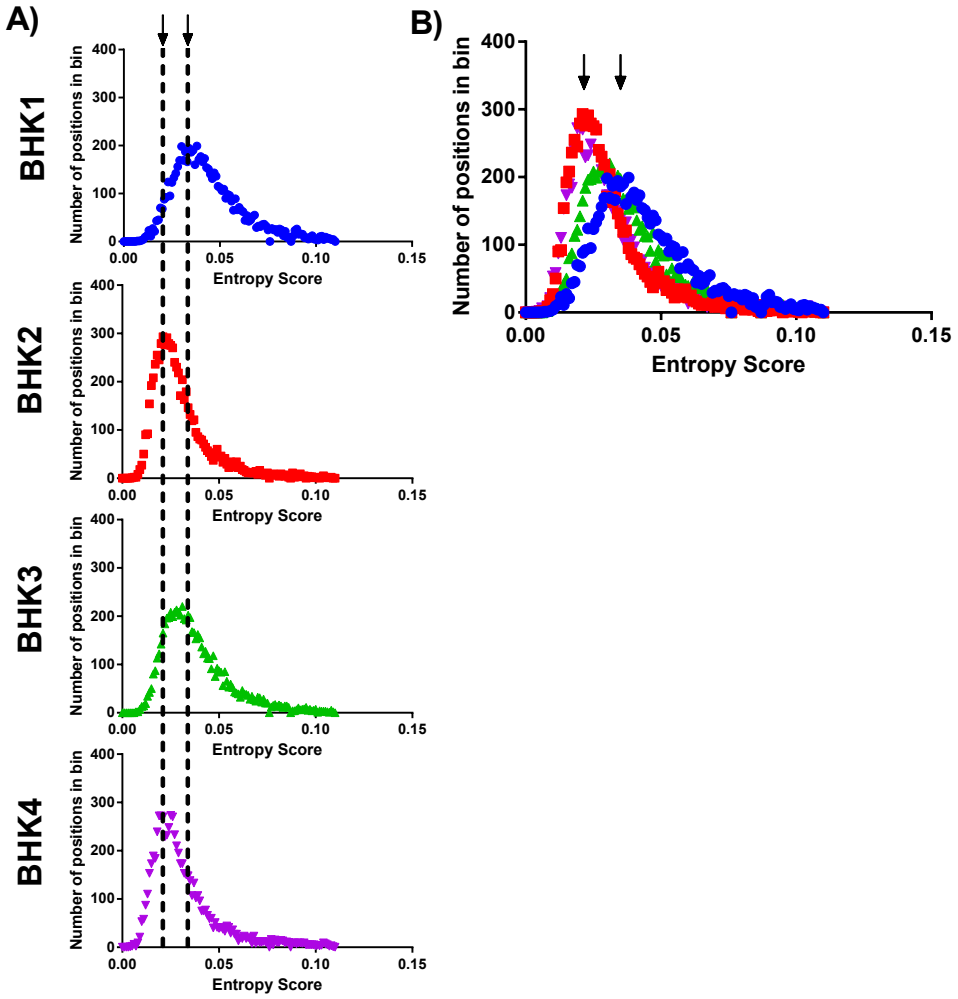


FIGURE 6.15: **Two distributions of entropy scores are apparent** Entropy scores were sorted into bins of 0.001. The number of positions (y-axis) is shown against the bin centre (x-axis) for BHK1, BHK2, BHK3 and BHK4 A) separately and B) overlayed

6.4.3.3 Decreases in the majority midrange scores are reflected by an increase in low and/or high entropy positions

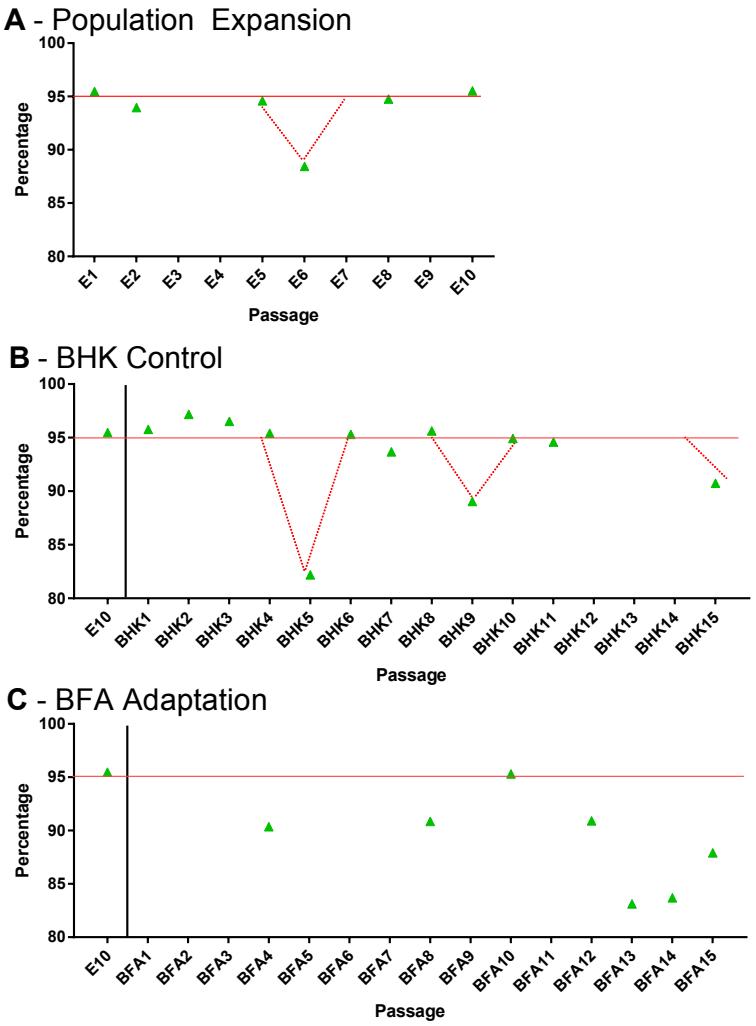


FIGURE 6.16: **Percentage of the genome with a Shannon's entropy score of 0.01-0.1** The percentage of the genome with an entropy score of between 0.01-0.1 is shown for (A) the BHK control (B) and the adaptive passage (C)

As shown above, the majority of genome positions for each of the passage series have low entropy scores. To consider the distribution of entropy scores outside this majority scores were sorted into three larger bins: 0-0.01 (low entropy scores), 0.01-0.1 (mid range entropy scores) and 0.1-0.15 (high entropy scores). These values are considered as a percentage to make it comparable regardless of the region of the genome with >1000x coverage. For both the population expansion and the BHK control passage series approximately 95% of genome positions fall between 0.01-0.1.

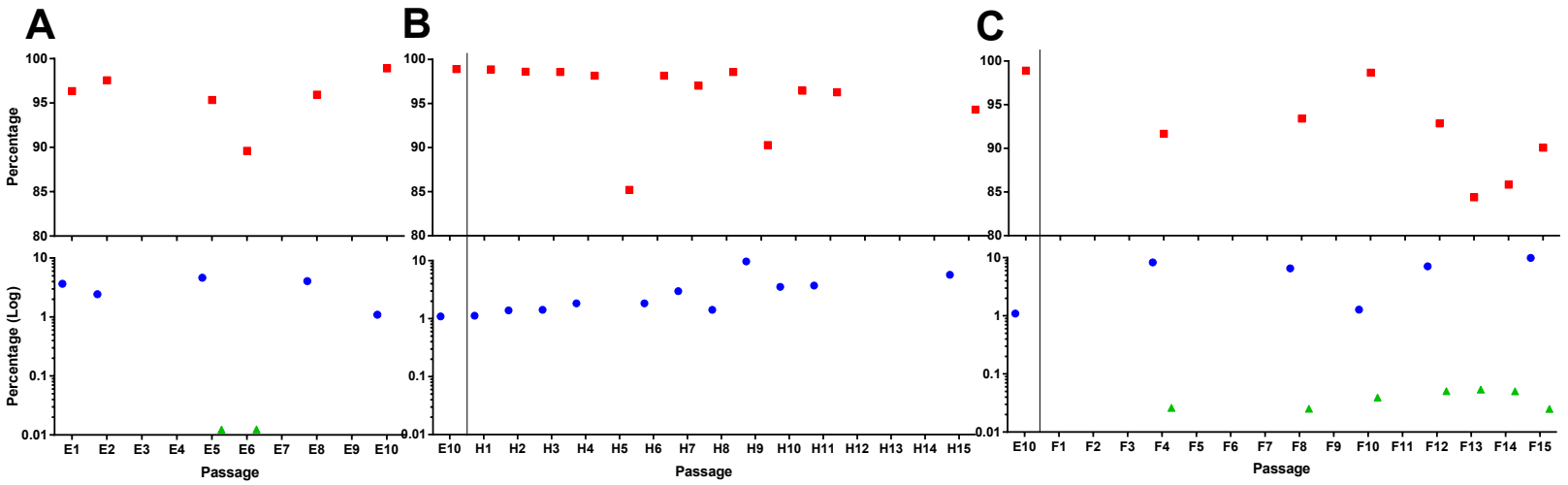


FIGURE 6.17: **Fluctuations between entropy score histrogram bins** Shannon's entropy scores were binned into three bins low entropy (0-0.01(blue circle)), mid-range entropy scores (0.01-1 (red squares)) and high entropy scores (1-1.5(green triangles)). This is shown for the populations expansions (A) the BHK control (B) and the adaptive passage (C).

This consistent proportion only dips for one passage before recovering. This 95% is annotated on Fig. 6.16 with a red line with the one passage dips shown with a dashed red line. In comparison, the adaptation series is generally below this proportions and the dips last longer than one passage (Fig. 6.16).

When the percentage of values in the majority bin (0.01-0.1) drops (as displayed in Fig. 6.16) a concurrent increase is seen in the low range entropy scores (0-0.01, blue circle) and vice versa (Fig. 6.17). In some situations this is coupled with the appearance of a proportion of high entropy scores (1-1.5, green triangles) for example expansion passage five and six and BFA adaptation passage 12-15 (Fig. 6.17).

Obviously, as these are percentage values, if one drops something else must increase. What is interesting is that if high entropy values appear there is also an increase in low entropy values. As high entropy changes appear in the population the number of mid range entropy scores drop and the number of low entropy scores increase leaving the cumulative entropy relatively consistent as previously shown.

As these high entropy positions become fixed in the population they would go from having a high entropy score to mid range and potentially to low. This can potentially be seen at the latter end of the BFA adaptation. Dips where no associated increase in high entropy position are shown (i.e. BHK5 or BHK9) could be evidence of a bottle neck with no adaptive pressure. If there is a bottle neck of some form, a percentage of the swarm is removed. This would cause the overall level of variation to decrease (less midrange values, more low level values). One of these outlying values equates to a drop in titre of inoculating virus (from the previous passage). Inconsistency of blind passage sampling could explain this. The other drop does not relate to lower titre from the input virus. However, cell number, quality or time of CPE could all affect the quality of virus produced by this experimental procedure. It does seem however that these changes are rapidly recovered with the next passage showing the expected 95/5% split between midrange and low entropy positions.

The observation of the cycling pattern between very low tight distribution and the higher broader distribution lead to the suggestion that when under no adaptive pressure the swarm cycles between a swarm with minimal variation and swarm with slightly more variation. This makes logical sense as as more variation is introduced some of it is likely to be non viable and be eradicated leading to less variation. However in periods of adaptive pressure, when some highly variable positions appear as adaptive changes come to fixation, the low level variation is somehow restricted with high entropy positions resulting in less midrange positions and more low level positions. These proposed population dynamics are outlined in Fig. 6.18.

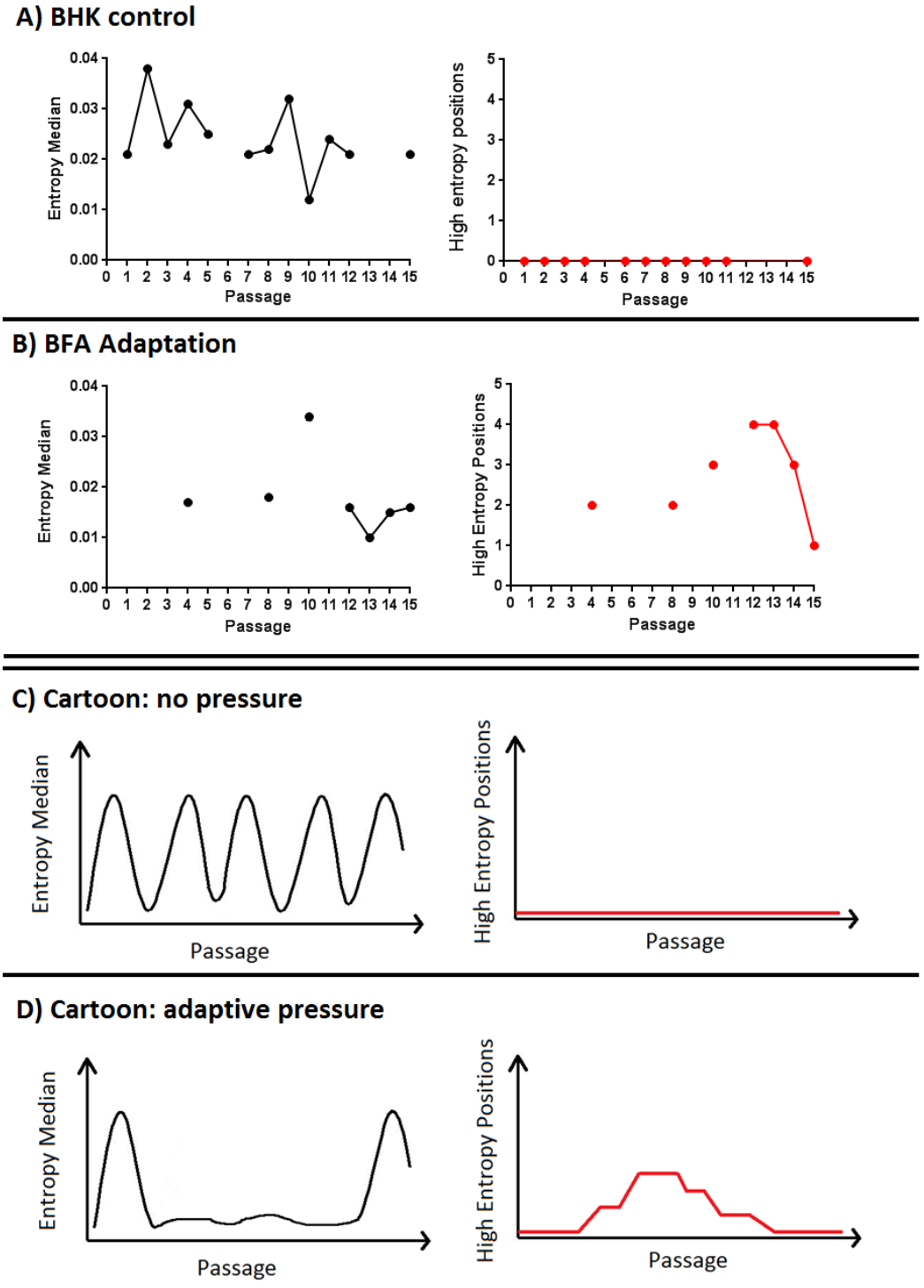


FIGURE 6.18: **Proposed swarm dynamics** A) The mean average entropy of the BHK control passage from BHK1-BHK15 is shown on the left hand side. The right hand side of the panel shows the number of high entropy positions each passage. B) Shows mean entropy score for the BFA adaptive passage on the left hand side of the panel. The associated number of high entropy positions is shown in the right hand panel. A cartoon of this pattern is seen in the lower panel. C) Under no adaptive pressure the swarm cycles between a swarm with minimal variation and swarm with slightly more variation. D) In periods of adaptive pressure there are some highly variable positions as adaptive changes come to fixation but less low level variation

This could be because less genomes will have the capacity to survive in the new environment or evidence of the swarm having the capacity to withstand a limited amount of variation. If the latter explanation proves to be correct this could be some preliminary evidence towards the FMDV swarm being a quasi-species as discussed in the Chapter 8 (Fig. 6.18).

6.5 Summary

Adapted virus shows a different phenotype The adapted virus had improved replication efficiency and appears to cause CPE more efficiently.

Virus adapted to BFAs shows consensus level changes. There were four consensus level changes identified on each in the IRES, VP2, VP3 and 2C. These all appeared following a decrease in viral titre suggesting a bottle neck produced in the blind passage process.

Swarm dynamics This work suggests when a viral swarm is under no continued adaptive pressure the swarm cycles between a state of minimal variation and slightly more variation. These two states are a product of each other (more variation means an increased number of non-viable genomes resulting in less variation). These periods of more variation may allow the virus to adapt when under selective pressure as the increased variation may included a mutation to respond to/overcome the pressure. When under this pressure or adapting to it, high entropy positions appear as mutations and become fixed in the majority of genomes. This is coupled with a lower mean entropy score, either because less genomes can survive under the selective pressure (smaller swarm = less variation) or because the swarm can withstand only a limited amount of variation - a concept that would hint at FMDV acting as a quasi-species.

Chapter 7

Putative packaging signals

7.1 Abstract

Next generation sequencing is now an affordable technique that allows the creation and interrogation of large data sets revealing details about how a virus exists and evolves. These large data sets can be used not only for observing viral dynamics and consensus level sequencing but to consider fundamental biological questions from a new perspective. The following work shows how interrogation of a viral swarm can be used in a more fundamental virological study to investigate a feature that may have previously been camouflaged by the highly variable nature of the virus swarm: RNA packaging in *Picornaviridae* .

FMDV samples before and after virus purification were sequenced on the Illumina Miseq. These samples represented the 'total' population and 'encapsidated' population of the viral swarm respectively. Comparison of Shannon's entropy score at each position of the genome for each population allowed regions of the encapsidated population's genome that were more conserved to be identified. Secondary structure analysis of these regions showed conserved stem loop structures. Repeats of this analysis with the same virus showed that the features are, to an extent, functionally redundant with approximately 50% repeated in both experiments. These potentially functionally redundant secondary structures were confirmed in a second virus from the *Picornaviridae* family; poliovirus. Subsequent molecular work disrupted a subset of these conserved features in FMDV and resulted in a phenotype that achieved full CPE more slowly and produced a log less virus. The mutant virus was shown to replicate and translate comparably to the wildtype virus and the phenotype was therefore attributed to genome packaging.

7.2 Introduction

Although much is known about *Picornaviridae* capsids, and to an extent how the capsid physically assembles, detailed explanation of how the RNA folds and interacts with the capsid proteins is still to be elucidated. The folding of viral genomic material and subsequent packaging into outer capsid proteins offers a target for effective antiviral strategy as has been shown for DNA viruses hepatitis B and herpes simplex virus [259, 260] and RNA viruses HIV and influenza [261, 262]. If a sufficiently conserved element of the picornavirus encapsidation process can be identified an antiviral could be designed.

7.2.1 Encapsidation processes

Diverse mechanisms are used for genome packaging by different virus families. In double stranded DNA viruses, DNA is often packaged into a pre-formed viral capsid through an entrance vertex with the assistance of powerful ATP driven motors [263]. In RNA viruses however virions are inclined to assemble spontaneously around the genome. For this process to be successful the genome must be small enough to fit within the capsid and co-ordination is required for the capsid and RNA to assemble together. Therefore, this strategy for packaging is chiefly dependant on two elements: genome compactness and RNA-protein interactions. The simplest RNA viruses selectively package a single copy of their genome into a capsid derived from a single protein subunit minimising the protein required to produce virions and thus maximising potential output from one infected cell.

7.2.1.1 Genome compactness

Naked viral RNA can be 25% larger than the virion [264]. Furthermore, a compacted structure can be necessary to bind capsid proteins [265]. RNA folding and collapse is therefore a key part of the RNA encapsidation process. The compact, high density form of viral RNA is created by branching and folding. This can be due to binding or proximity of the RNA and capsid protein resulting in a rapid collapse of the viral RNA which would previously have had a hydrodynamic radius far larger than the inside of the capsid.

7.2.1.2 RNA-protein interactions

Viral packaging processes usually involve interactions between the viral genomic material and structural proteins (capsid elements) or non-structural proteins. These interactions can either aid in packaging efficiency or be an essential step in the packaging process.

RNA-protein interactions have been well studied in several viruses and are primarily electrostatic. Electrostatic interaction depends on a non-covalent bond between a positive area on the proteins surface and the negatively charged RNA. Several RNA-protein interactions have been described [266]. For example, in one group, the groove in an RNA-helix binds peptides or proteins. Alternatively single stranded RNA can interact in a sequence specific manner with pockets created in a protein's β-sheet surface. This latter interaction can be enhanced by the RNA structure [267].

Most RNA-capsid interactions are assumed to be nonspecific but more recently, studies have been completed that show certain electrostatic interactions are sequence specific. This has been shown for example in Cowpea mosaic virus (CPMV)[268, 269]. A slightly more complex situation exists in satellite tobacco necrosis virus (STNV). In STNV a combination of sequence specific interactions and electrostatic forces result in successful packaging [270]. STNV capsid monomers have sufficient electrostatic force to prevent them from forming a full capsid due to repulsion. The addition and interaction of the RNA overcomes this [271].

In RNA bacteriophages such as MS2 from the *Leviviridae* family the encapsidation process has been quite fully defined. In this scenario binding of an RNA stem loop to the capsid protein can cause a conformational change in the capsid protein, facilitating dimerization. These dimers bind one another to form the assembled capsid [272]. RNA can also interact with neighbouring dimers, helping to bring together the capsid [273].

Stem loop structures with capsid-protein interactions such as the ones described in MS2 have been found in many viruses although they do not necessarily result in conformational changes in structural proteins. They are often termed packaging signals [271]. In some cases, such as for hepatitis B, one stem loop is sufficient for encapsidation [274]. Recent work completed by Dykeman and colleagues has suggested an assembly mechanism for some viruses which includes multiple degenerate packaging signals throughout the genome that results in a conserved assembly pathway [275].

7.2.2 *Picornaviridae* encapsidation

In viruses belonging to the picornavirus family the encapsidation process is not well understood. Advances in viral structure analysis have allowed for the visualisation of

RNA-protein interaction points. Electronmicroscopy facilitated the structure of the picornavirus Ljungan virus to be resolved at 3.78 in 2015 revealing potential electrostatic interactions between the inside of the capsid shell and the RNA itself [276]. A similar finding was reported from the structure of human parachovirus 1 suggesting a hexanucleotide association close to the 5-fold axis [277]. This suggests that RNA-structural protein interactions may be important in picornavirus encapsidation.

These interactions could be specific. An RNA stem loop packaging signal has been identified at the 5' end of Aichi virus (Genus: *kobuvirus*, Family:*picornaviridae*) [278, 279] which has given rise to investigation of RNA packaging signals as a method for encapsidation in other picornaviruses.

Work with the picornaviruses coxsackie virus and poliovirus has suggested a role for viral non-structural protein 2C in RNA encapsidation. It has been suggested that 2C provides specificity to the RNA during the encapsidation process (for enteroviruses) and scanning mutagenesis has more specifically shown the C-terminus of 2C to be required for RNA encapsidation [255, 280, 281]. Studies have found co-localisation between capsid proteins VP1 and VP3 and 2C further highlighting its involvement in the encapsidation pathway [280, 282].

Other pieces of information are available about packaged picornavirus RNA. It has been shown that poliovirus only encapsidates newly synthesised positive strand RNA [98, 99] that are linked to the genome linked protein (VPg) [283]. This protein acts as a primer during RNA synthesis and is covalently linked to the 5' end of the genome. This has led to the suggestion that viral replication could be linked to encapsidation, thus allowing for specific encapsidation of newly synthesised RNA molecules.

Numerous potential elements of the picornavirus packaging pathway have been identified but the variation between viruses of the same family makes for a highly confusing and contested area of research. Work to understand which of these mechanisms is in use for each of the picornaviruses needs to be completed.

7.2.3 Experimental approaches

Previous studies to consider packaging signals have been completed in a range of viruses. Methods used have included, but are not limited to, deletion analysis, trans-encapsidation studies and SELEX. Each of these methodologies has limitations. In deletion analysis regions of the genome are systematically deleted and packaging of subsequent mutants considered. Similarly, different regions of viral genomes have been cloned into other viruses or non-viral RNA to observe if this confers the ability to package. In retrovirus

Moloney murine leukemica virus, experiments have successfully cloned packaging signals into non-viral mRNA allowing their subsequent encapsidation [284]. Issues with this area of research relate to the difficulty in separating packaging from other genomic functions. Many small RNA viruses have short genomes and it has been predicted that each region of the genome is likely to be multifunctional with its sequence and/or structure being important in several viral processes. Therefore amending a viruses even with synonymous mutations could result in changes in as yet unknown processes. Realistically, this may be unavoidable as in viruses where processes are inherently coupled it would be impossible to alter one aspect without effecting another.

SELEX is the systematic evolution of ligands by exponential enrichment [285]. This process involves the exposure of a protein or ligand to a random library of nucleotide sequences. Those that successfully bind are subsequently amplified and the process is repeated with this new selected library. This process is repeated numerous times to identify the sequences most likely to be bound by the protein of choice. In work considering packaging signals these selected sequences are then aligned to the genome with a mapping algorithm to identify their most appropriate positions and identified regions are statistically and structurally analysed to consider their potential function in packaging [275]. This methodology is preferable over traditional deletion studies because it does not depend on genomic regions being solely responsible for one function. This process requires the bombardment of small fragments of genomic material at capsid pentamers. The capsid pentamers are created by denaturing fully formed empty capsids. Formation of capsids involves the auto-catalytic cleavage of VP0 so any fully formed capsid would have already undergone this cleavage. Consequently, the pentamer subunits produced by capsid denaturation are slightly different from natural pentamer subunits. Secondly the fragment length means that each fragment could align to several regions in a genome, requiring statistical analysis to find positions of best fit.

7.3 Experimental Design

This study uses a novel approach to consider the question of FMDV RNA packaging. Virus populations before and after encapsidation have been sequenced and compared to assess restrictions put on the encapsidated population only (Fig. 7.2). This offers an improvement on previous methodologies as it does not require any amendments to the genome or protein that could obscure or alter identification.

In short, exemplar viruses were grown up in large culture and a proportion of this culture was purified to produce virions.

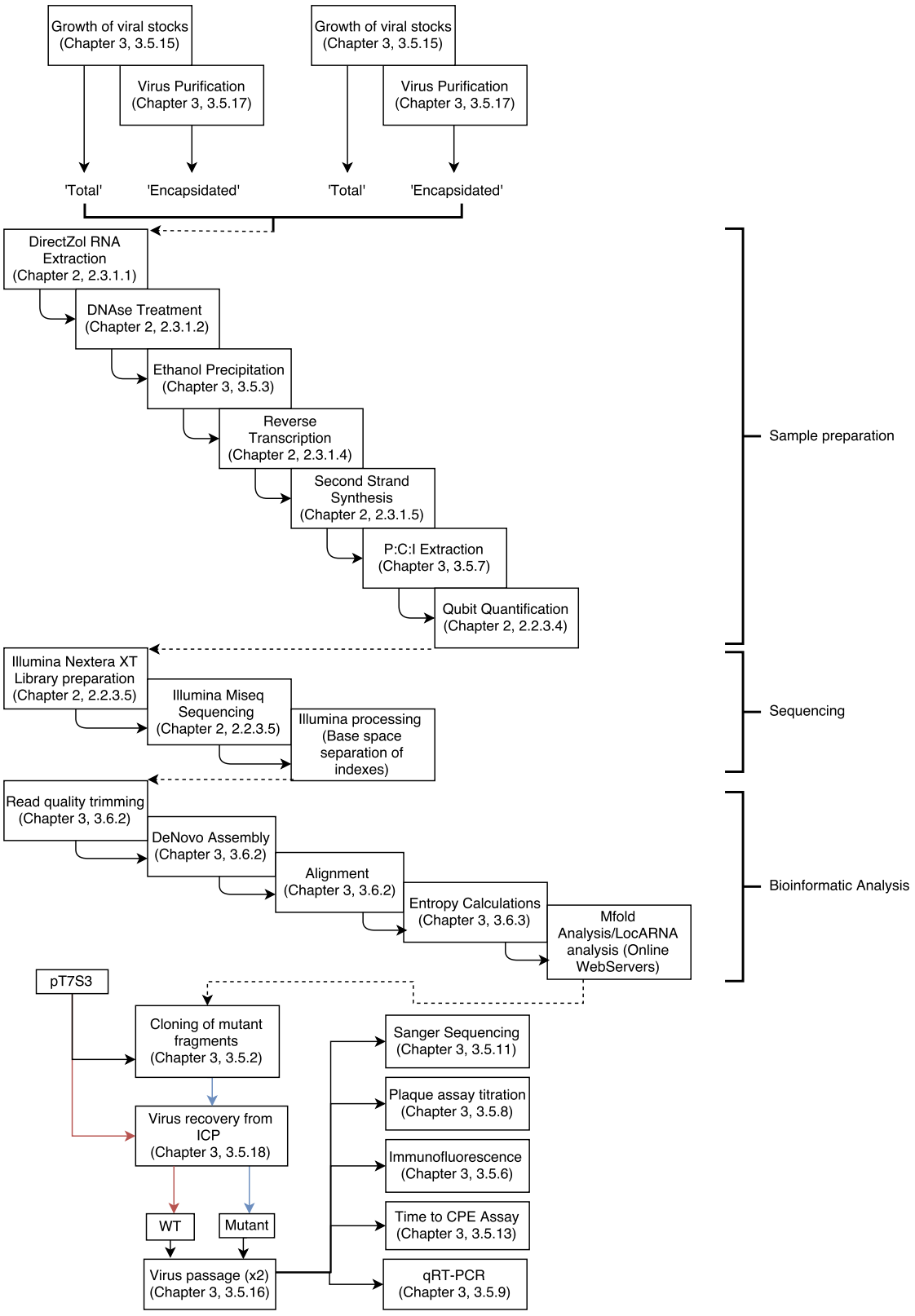


FIGURE 7.1: **Putative packaging signals: flowchart of methods used** Samples of FMDV and poliovirus were purified by Lidia Lasecka and Caroline Wrights or Toby Tuthill and Hannah Wenham respectively. Samples pre and post purification were sequences and entropy scores at each position of the genome compared. Areas more conserved in the packaged population were identified and used to inform the design of a packaging mutant. Geneart fragments with the mutation were cloned into an ICP and comparisons made between the WT and mutant.

The original population (total) and the resulting purified virions (encapsidated) were used in RNA extractions and the genomic material sequenced on the Illumina MiSeq as previously described [170]. At positions with coverage of greater than 1000x the Shannon's entropy score was calculated and the score for the encapsidated population subtracted from the score for the total population.

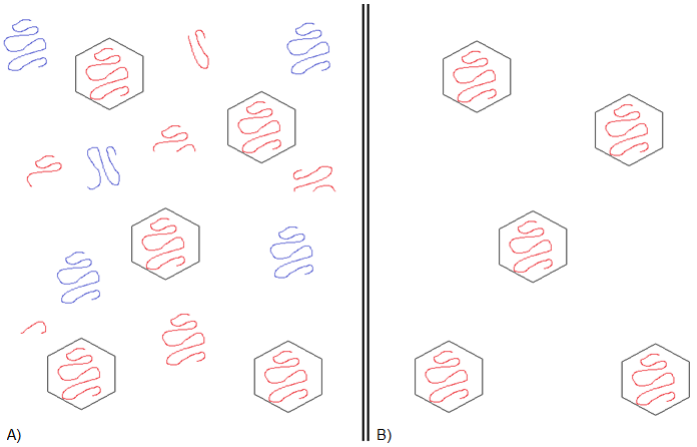


FIGURE 7.2: **Packaging requirements: experimental design** Two populations of RNA were sequenced to consider requirements for sucessful packaging: A) The 'total' population. This contained packaged full length positive sense RNA (red, in capsid), unpackaged full length positive sense RNA (red, no capsid), unpackaged full length negative sense RNA (blue, no capsid) and non full length of both negative and positive sense RNA (blue and red respectively.) B) The 'encapsidated' population. This population included only successfully formed virions, assumed to contained positive sense full length RNA strands.

Positions with positive values were deemed to show less variation in the encapsidated population than the total population and thus likely to be required or at least preferable for successful packaging. This process was completed with FMDV and poliovirus to asses if any packaging elements could be identified that were conserved within the *picornaviridae* family. A mutant virus was then created lacking 3/12 identified regions identified in FMDV. This mutant was then compared to the wildtype. An outline of the methods used in this chapter can be seen in Figure 7.1. Further detail can be found in Chapter 2 and 3.

7.4 Results and Discussion

7.4.1 Bioinformatic analysis

7.4.1.1 The majority of genome positions are more variable in the encapsidated swarm.

A pilot experiment was conducted with O1Kaufbeuren (O1K) to confirm the proposed protocol would have the resolution to identify differences between the total and encapsidated population. The difference in Shannon's entropy scored between the total population and the packaged population was calculated and visualised to identify regions more conserved in the encapsidated population (positive values). 70% (5630 genome positions) had negative scores, 29.1% (2329 genome positions) had positive values, 0.1% (9 genome positions) showed no difference between the two populations and 2.5% (202 genome positions) had insufficient coverage in at least one population preventing accurate comparison of the entropy scores (Fig. 7.3).

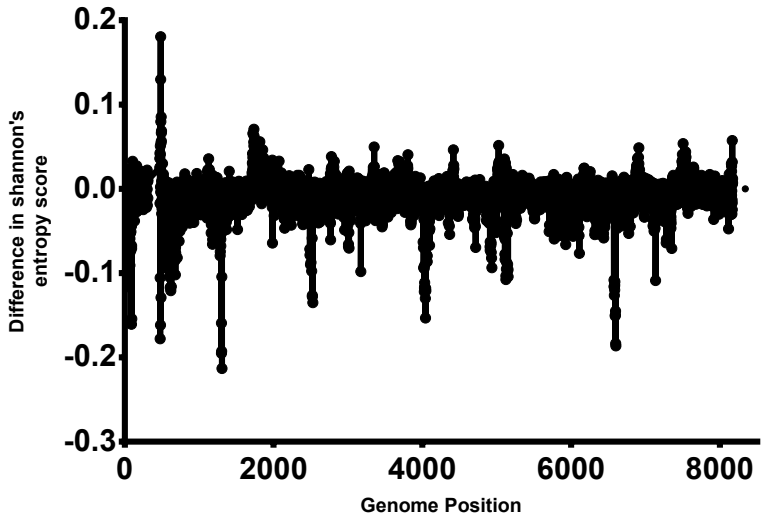


FIGURE 7.3: **Difference in Shannon's entropy scores.** The Shannon's entropy score from the encapsidated populations was subtracted from the score from the total populations and values were plotted against genome positions (Equation. 4.1.)

The majority of positions reported a negative score, that is to say the entropy score from the total population was smaller than that in the encapsidated population. This was an unexpected finding. Considering the encapsidated population is a subset of the total population it was expected that the amount of variation in this population would be mostly lower due to the smaller population size. One possible explanation for this

can be seen in recent work by Schulte *et al.* This work suggests that the traditional 'stamping out' method of viral progeny production may not be accurate. They propose that virions produced from each cell are approximately five generations removed from the virus that entered [286]. A similar point is made by Thebaud *et al* [143]. This provides the opportunity to accumulate variation. If the later replicated genomes are the genomes that are preferentially packaged then the encapsidated population will in fact be made up of the genomes in the population most diverged from the original infecting population. Further work needs to be completed to identify if this explains the pattern seen.

7.4.1.2 Conserved genome positions in the encapsidated swarm cluster.

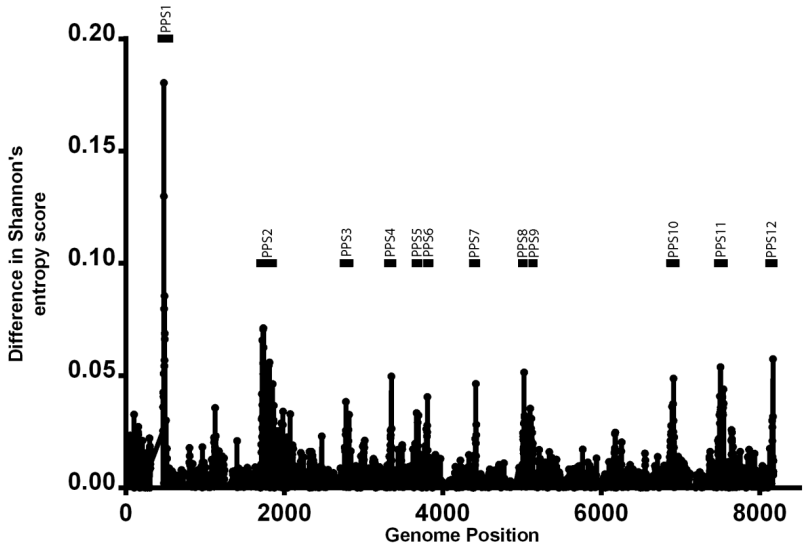


FIGURE 7.4: **Genome positions more conserved in the packaged swarm cluster.** The Shannon's entropy score from the packaged populations was subtracted from the score from the total populations and positive values were plotted (representing less variation in the packaged population) against genome positions. Clusters of positive values representing conserved regions were marked and termed putative packaging signals (PPS).

The difference in entropy score, calculated at each nucleotide position, between the packaged and total population which resulted in a positive value was used to identify 12 regions (termed putative packaging signals (PPS) 1-12) more conserved in the encapsidated swarm in comparison to the total population (Fig. 7.4). These regions were spread throughout the genome. Previously it has been assumed that any packaging signal present in the genome would not be situated in P1 as numerous *picornaviridae* sub-genomic replicons have been created with a reporter gene in place of P1 and some

of these have been shown to encapsidate in transencapsidation studies. This may not be entirely accurate however. Hamiltonian pathway theory suggests that packaging signals may be present throughout the genome and their efficiency may be degenerative. If this is the case, replicons with no P1 may be able to package but with lesser efficiency [275].

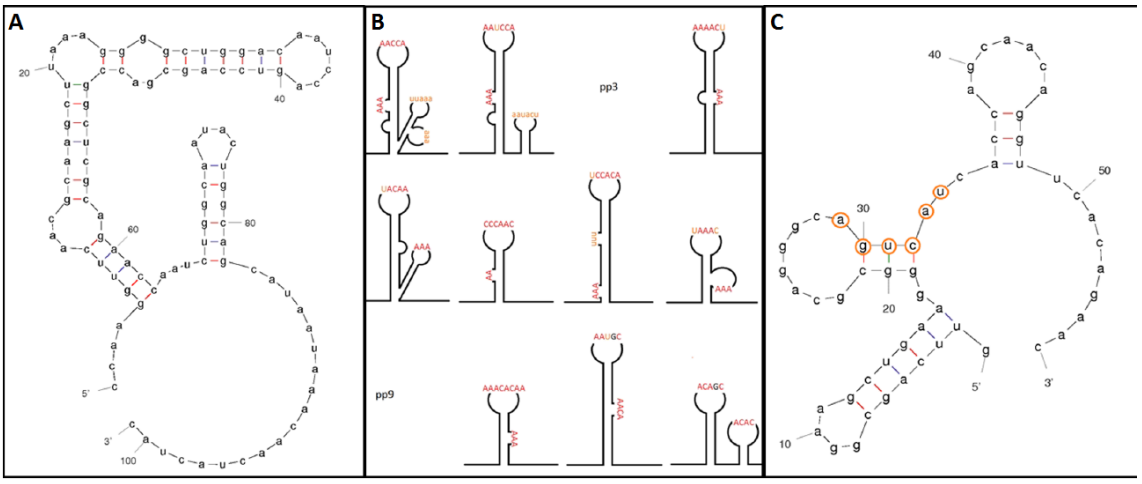


FIGURE 7.5: **The secondary structure of each PPS formed a stem loop with a consistent shape and motif. Synonymous nucleotide changes could be engineered that changed the structure.** A) Each identified conserved region, or PPS, was run through Mfold to consider its secondary structure (PPS2 shown as an example). B) A consistent pattern was found in 10/12 stem loops. This featured an adenine and cytosine rich loop with a smaller adenine bulge or protrusion in close proximity. C) Non-synonymous changes could be made to disrupt both loop region (highlighted in orange) and the structure (PPS2 shown as an example).

7.4.1.3 Conserved clusters form repetitive secondary structures termed putative packaging signals (PPS).

To consider if the less variable regions identified (PPS 1-12) had a repetitive secondary structure the PPS were analysed using Mfold. This revealed a conserved stem loop structure around all the PPS positions with the exception of PPS3 and PPS9. These structures consisted of an adenine and cytosine rich loop complemented by a proximal adenine rich bulge (Fig. 7.5). Although this structure was not obvious in PPS 3 or 9 there were adenine rich regions and thus it could be possible that these structure do exists although they were not immediately obvious.

To asses whether this structure was conserved from isolate to isolate findings were confirmed using locaRNA [287–289]. This software performs a structural alignment to show structures that are conserved, regardless of their nucleotide sequence. This showed conservation of the structures in 30 other type O consensus sequences downloaded from

genbank (Appendix A,Table E.2). This suggests that the structures found could be biologically relevant as there is evidence of them on a serotype wide scale.

7.4.1.4 Approximately half the identified PPS are repeated from passage to passage.

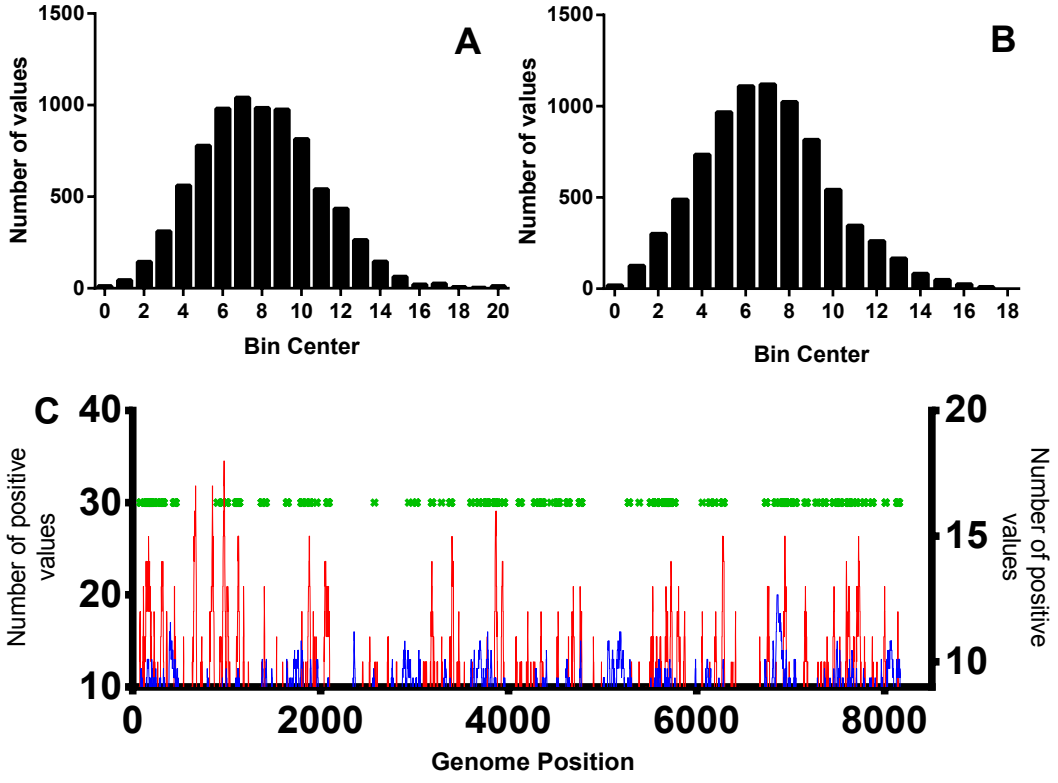


FIGURE 7.6: **FMDV: a subsection of PPS regions are identified in multiple strains and some appear to be viral swarm specific.** A frequency distribution of the number of positive values per 20 nucleotide window was completed for experiment one (A) and experiment two (B). Windows with a score above the 75 percentile (10 and 9 respectively) were visualised (C). Experiment one (red) on the left axis and experiment two (blue) on the right axis had 711 nucleotides in common (green) spread throughout the genome.

Having established a protocol, repeat experiments were performed with a new virus preparation of the same virus (FMDV O1K experiment 2) and two different virus preparations of type 1 poliovirus (Mahoney) (poliovirus experiment 1 and 2). This resulted in four data sets: two FMDV O1K and two poliovirus M1. In the previous experiment regions of conservation were identified visually. This allows for some human introduced bias. An improved analysis techniques was developed to improve repeatability. A sliding window analysis was performed across the genome (with a 20 nucleotide window)

to count the number of positive values within a window. The frequency distribution of these values was considered and windows with a score above the 75 percentile highlighted (Fig. 7.6).

For FMDV O1K there were 711 nucleotides identified in both analysis (Fig. 7.6C). 35 regions were identified in experiment one and 53 in experiment two potentially suggesting improved resolution in the latter experiment. This could be due to slightly higher coverage in experiment two, although this difference is minimal. 51% of the regions identified in experiment one share some overlap with regions highlighted in experiment two, which equates to 45% of the identified regions in experiment two.

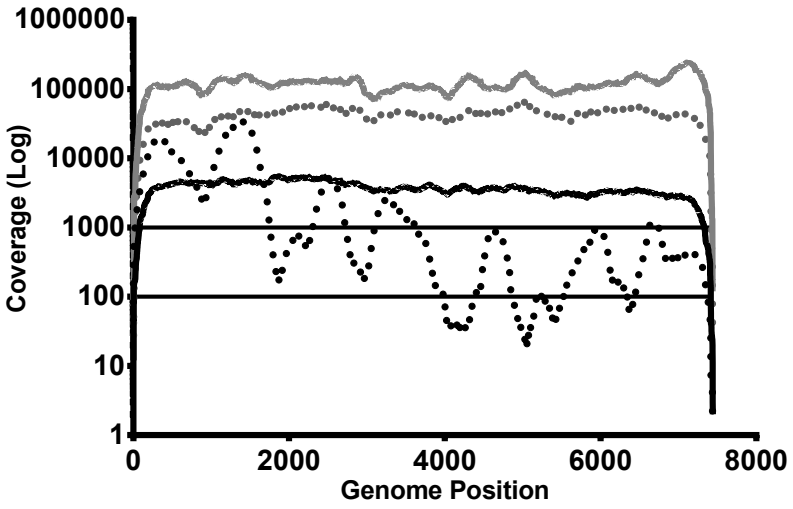


FIGURE 7.7: **Comparable levels of coverage were not achieved between poliovirus repeat experiments.** The coverage across the genome is shown for total samples (solid line) and encapsidated samples (dashed line) for experiment 1 in black and experiment 2 in grey. Coverage cut off is marked at 1000x. Analysis on experiment one was completed on regions above 100x (also marked).

For poliovirus M1 coverage of one of the viral preparations was low (Fig. 7.7). This is likely due to the first sample being in the freezer for a long time and potentially suffering some degradation. The second sample, a passage grown from the first sample, achieved much higher coverage. The coverage cut off for the poliovirus data sets was therefore set lower at 100x. The distribution of positive scores per window for the poliovirus experiment one was slightly different from that of the second poliovirus experiment and the FMDV experiments. The different distribution of values can be explained by the differences in coverage achieved for each of the repeats (Fig. 7.7).

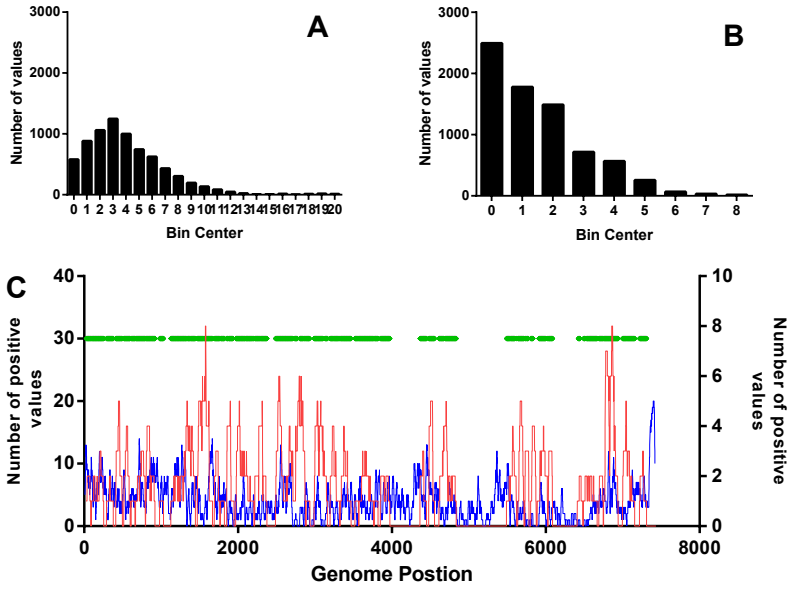


FIGURE 7.8: **Poliovirus: a subsection of PPS regions are identified in multiple strains and some appear to be viral swarm specific.** Sliding window analysis was completed counting the number of positive values in 20nt windows across the genome. The frequency distribution is shown in a histogram for experiment 1 (A) and experiment two (B) with a bin width of 1. The mean of experiment one is 1.510 with a standard deviation of 1.545. The mean of experiment two is 3.999 with a standard deviation of 2.969. C) Experiment one (red) on the left axis and experiment two (blue) on the right axis had 4549 nucleotides in common (green) spread throughout the genome.

Lower levels of coverage in experiment one would have resulted in some low level variants not being represented and thus more positions returning a Shannon's entropy score of 0 (Fig. 7.8). Although the distributions are markedly different the descriptive statistics of the frequency distribution could be used to identify windows with values above the 75% percentile. Using these value as a cut off clusters were identified in experiment 1 (windows with >2 positive values) and experiment 2 (windows with >6 positive values). Regions were merged together if they were within 50bp of one another on the genome revealing 24 regions of interest in experiment one and 26 regions of interest in experiment 2 (Table 7.1). As with FMDV approximately half showed overlap in the other experiment (0.5-0.53).

This data suggests that approximately half of the putative packaging signals are consistent and half vary from isolate to isolate. There could be a number of reasons for this. Given the extensive secondary structure in the FMDV genome, and to a lesser extend in the poliovirus genome, it may be that there are numerous packaging signals that could be utilised throughout the genome but that they are, to an extent, functionally redundant. Therefore a selection of packaging signals will be evident in each passage but they

will not necessarily always be the same selection of packaging signals. Potentially, secondary structures of the genome that have a dual purpose are more common packaging signals. This would lead to a system where genomes with structure required for entry or replication would be preferentially packaged. This fits with the observation that the secondary structure in the IRES has been identified in all experiments. A secondary structure element required for replication was also identified in FMDV experiment one in the form of stem loop one (SL1) but the coverage was insufficient at this location in the second experiment to confirm this observation. Similarly, in poliovirus experiment two, stem loop Y was identified. This is also a secondary structure element required for replication.

Some secondary structures have been found to be functionally redundant in other picornavirus process. For example, two functionally redundant RNA structures (α and β) have previously been described by Song *et al* in poliovirus 3Dpol [290]. These structure have been identified in this analysis also: α sits 6995nt-7042nt highlighted in poliovirus experiment one (7009nt-7087nt) and β sits 7192nt-7267nt highlighted in poliovirus experiment two (7114nt-7209nt). As the viruses used are different passages of the same strain it is conceivable that the majority in the swarm of experiment one use α while the majority in experiment two use β. This couples with the concept that exactly the same combination of packaging signals is not required in every genome, potentially highlighting the slightly flexible nature of this process. Each virus may have a similar number of packaging signals in similar regions but there may be several options in each region.

If α and β represent functionally redundant features that can sometimes act as packaging signals, there are other more consistent structural elements that may always act as packaging signals. Four known secondary structure elements in poliovirus were identified by the analysis in this current study: the IRES [291, 292], 2C *Cre* [293], RNase L competitive inhibitor [294, 295] and Stem loop Y [296]. Each of these features were identified in both experiments except Stem loop Y which lacked sufficient coverage for consideration in experiment one. The fact that previously recognised secondary structures may also have an additional function in packaging is not surprising as small RNA genomes have previously been shown to have multifunctional secondary structure elements [270]. The largest difference in entropy between total and encapsidated population was found in poliovirus experiment two and coincides with stem loop Y. This stem loop is at the 3'end of the open reading frame and has previously been implicated in replication. This being the most constrained stem loop in a swarm would be beneficial from an evolutionary standpoint as the full length genomes would be preferentially packaged thus producing viable virus. Further work would need to be completed to asses if this was a theory reflected by the data.

Experiment 1	Genome region	Experiment 2	Genome region
*204-219	IRES	*14-277	Clover leaf/IRES
416-465	IRES	329-372	IRES
534-568	IRES	513-528	IRES
*742-855	VP4	*699-759	VP4
		*811-1322	VP4/VP2
*1303-1625	VP2	*1404-1467	VP2
1712-1741	VP2	1617-1680	VP2
		1743-1757	VP2
1865-1905	VP3	*2067-2080	VP3
*1980-2080	VP3		
2270-2320	VP3		
*2498-2567	VP3/VP1		
*2663-2867	VP1	*2540-2691	VP1
*2999-3123	VP1	*2884-2886	VP1
3222-3248	VP1	*3030-3041	VP1
*3374-3723	VP1/2A	*3383-3475	VP1/2A
3620-3669	2A	3582-3587	2A
		3785- 4065	2A/2B
4400-4414	2C	4286-4379	2C
*4483-4532	2C	*4441-4505	2C
4658-4720	2C		
		4561-4575	3A
		5117-5126	3A/3B
5566-5590	3C	5346-5406	3C
5653-5866	3C		
5925-6081	3C/3D	5467-5591	3C/3D
*6771-6907	3D	*5933-5938	3D
7009-7087	3D	*6810-6909	3D
7261-7268	3D	7114-7209	3D
		*7267-7284	3D
		7343-7420	Stem loop Y

TABLE 7.1: **Location of PPS within the poliovirus genome.** Regions of the genome identified in the sliding window analysis for experiment 1 and experiment 2 and the regions of the genome in which they lie. Regions identified in both experiments are marked with an asterisk (*). Regions of the genome where coverage was not sufficient in the comparison genome for analysis to be completed are marked in red.

Previous studies have shown that poliovirus lacking the 3'UTR produced up to a log less virus although whether this was due to a packing deficiency was not considered [297].

It should be noted that not all secondary structures previously reported in poliovirus have been found in the current analysis. For example a structure found in 3D by Burill *et al* [298] has not been identified, however this is consistent with their results as they showed knocking out this structure did not affect the viruses ability to package and shows the specificity of the technique described here.

7.4.2 Molecular confirmation of bioinformatic observations.

7.4.2.1 Packaging mutant achieves CPE more slowly than wild type virus.

Using regions identified in FMDV experiment one, synonymous changes were designed to amend the secondary structure and coding loop of a mutant region spanning PPS2-PPS4. These loops were mutated with 5-13 nucleotide changes per region. Mutated regions were ordered from GeneArt and cloned in to the pT7S3 infectious copy plasmid. Mutant virus was recovered and confirmed using Sanger sequencing.

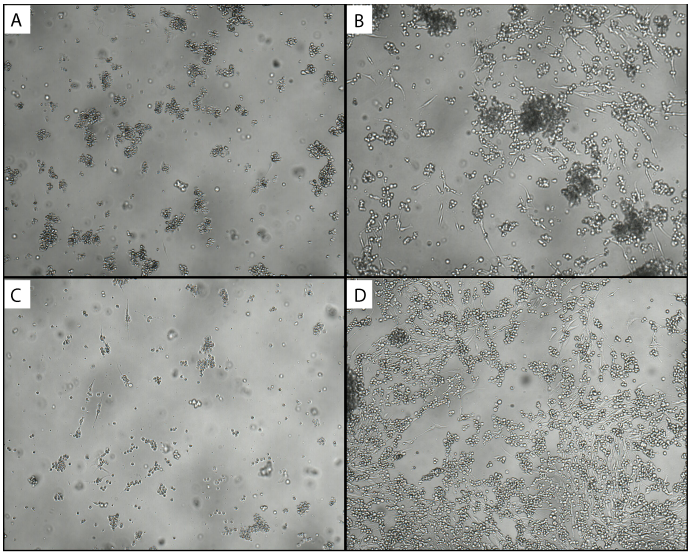


FIGURE 7.9: **In equal MOI infections cell death was more apparent in wild-type than mutant at comparable time points.** BHK-21 cells infected with wild-type or mutant virus at comparable MOIs at differing time points imaged on Juli smart fluorescent cell analyser (Digital Bio). A)MOI1 wildtype virus 30hpi. B) MOI1 mutant virus 30hpi. C) MOI0.01 wildtype virus 45hpi. D) MOI0.01 mutant virus 45hpi.

Equal MOI infections were completed and infected cell monolayers were imaged to compare mutant and wildtype rate of CPE. At thirty hours post infection with an MOI

of 1, cells infected with the wildtype virus were all rounded and mostly detached. In comparison, at the same MOI and time point the cells infected with the mutant showed some cell death with some rounded and detached cells but it also showed some healthy cells either not yet infected or in an early stage of infection. This effect is even more apparent at a lower MOI over a longer period. When infected with an MOI of 0.01 the wildtype had completely eradicated the monolayer by 45 hours post infection. In comparison the mutant virus at the same time point still had a number of viable cells. This suggests that the time taken for the mutant viruses to produce complete CPE is greater than that of the wildtype (Fig. 7.2).

7.4.2.2 Replication and translation efficiency of packaging mutant is comparable to wildtype virus.

To confirm the MOI used was comparable the number of infected cells were calculated at four hours post infection. The Minimax plate reader was used to count the number of cells as determined using ToPro3 nuclear stain. Infected cells were identified with immunofluorescence (IF) to viral non-structural protein 3A. The proportion of cells infected was not statistically significantly different (two-sample T-test, p-value=0.830) between the wildtype and mutant (Fig. 7.10A).

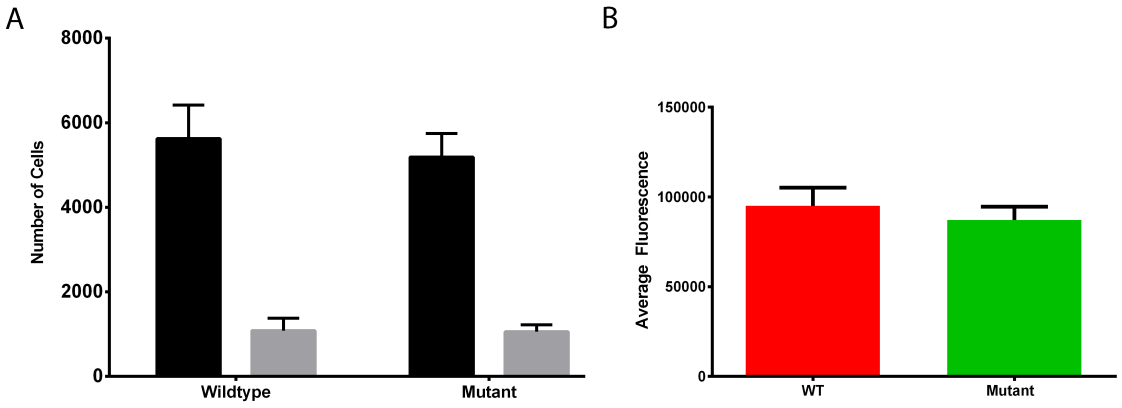


FIGURE 7.10: **Infection and replication efficiency between mutant and wildtype is comparable.** Immunofluorescence images were taken and quantified on the MiniMax. This data represents 8 replicates and is graphed with standard deviation A) ToPro3 staining and subsequent fluorescent object count at 620nm identified the total number of cells (black bar) and 2C2 labelling and fluorescent object count at 425nm determined the number of infected cells (grey) at 4hpi. 18.98% of cells in wild type were infected. 20.51% of cells in mutant were infected. B)The average fluorescence of each infected cells was calculated for wildtype (red) and mutant (green) at 4hpi.

The immunofluorescence data was further used to consider replication efficiency. The average fluorescence for infected cells in mutant and wildtype were compared. This

was used as a measure of replication as the more replication completed, the higher the number of non-structural proteins produced and thus the more labelling visible. The average fluorescence of wildtype infected cells and mutant infected cells is not significantly different (Mann-Whitney, p=0.1930) (Fig. 7.10B). This suggests that the translation efficiency of the two viruses is comparable.

Translation is not a completely accurate measure of replication as the ratio of replicating genomes to translating genomes is not fully understood. To address this potential discrepancy quantitative PCR (qPCR) was completed on samples taken every hour for eight hours for mutant and virus samples infected at equal MOIs. This analysis suggests comparable amounts of RNA are produced by the wildtype and mutant virus in an equal MOI infection (Fig 7.11).

The only point at which this was not consistent was at 2 hours post infection. At two hours post infection there was a dip in the amount of RNA present in the wildtype sample. This could be a product of cellular degradation. Not all RNA that reaches the cell goes on to be successfully translated or replicated, some may be degraded by cellular proteins or remain in the cell but in a non-functional capacity. Why there is a difference between mutant and wildtype at this point is unclear but as the remaining time points are comparable it suggests replication is not affected by the introduced mutations.

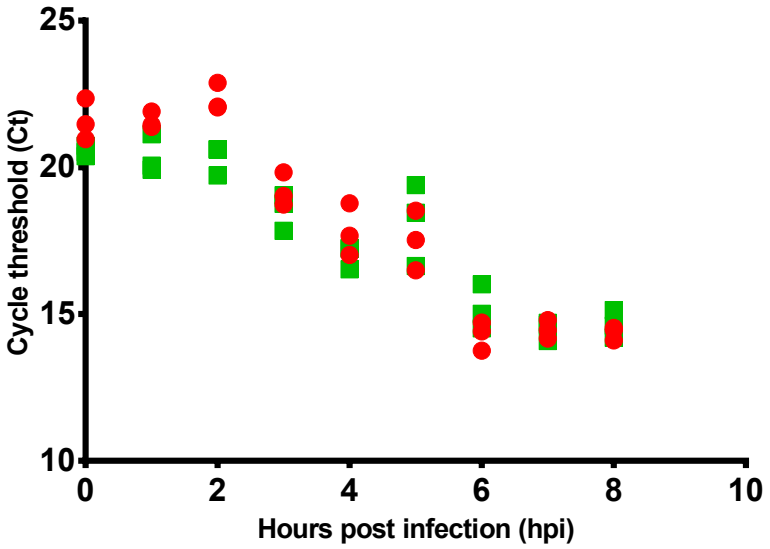


FIGURE 7.11: **Comparable amounts of RNA are produced by an equal MOI infection of wildtype and mutant viruses** RTqPCR was preformed on Directzol extracted RNA samples from an equal MOI time course comparison of wildtype and mutant viruses. Three time courses of wildtype and mutant were completed and RTqPCR was completed on each sample in triplicate. Cycle threshold Ct) averages of the triplicate values are shown for wildtype (red) and mutant (green) plotted against hours post infection.

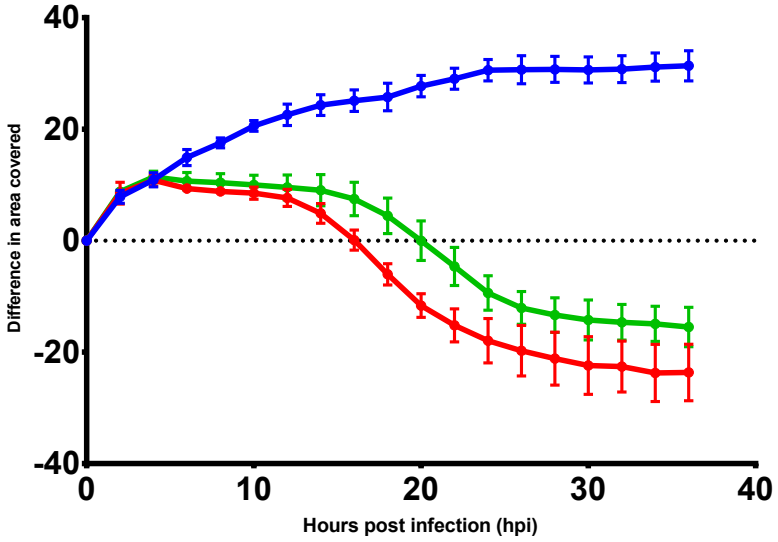


FIGURE 7.12: **The mutant virus depletes the monolayer more slowly than the wildtype virus.** Change in area covered by cell monolayer for uninfected cells (blue), wildtype infected cells (red) and mutant virus infected cells (green) is shown every two hours over a 36 hour period. These values were calculated by the MiniMax plate reader on the transmitted light channel. This data is representative of 8 replicates for the cell only control and mutant and 7 replicates for wildtype and is graphed with standard deviation.

Having confirmed that the infection rate and replication efficiency of the mutant and wildtype viruses were comparable (Fig. 7.11,7.10) an assay was designed on the SpectraMax MiniMax 300 Imaging Cytometer to compare the spread of virus through a cell monolayer. Equal MOI infections were performed in a 96-well plate and monolayers were imaged every two hours over a 36 hour time period.

Both sets of infected monolayers show the same growth as cell only control wells for the first four hours. From this point, the first round of replication is expected to be near complete and the equal number of cells originally infected by both is reflected by the plateauing of both the wildtype and mutant. From 9hpi a decrease in the area of the well covered by cells is apparent in the wildtype sample. An equivalent decrease is not evident in the mutant until 11hpi. Further to this delay the gradient of the wildtype curve appears steeper than that of the mutant (Fig. 7.12). These observations were confirmed statistically (Fig. 7.13) showing that the two curves were different and that the wildtype kills more cells than the mutant and earlier.

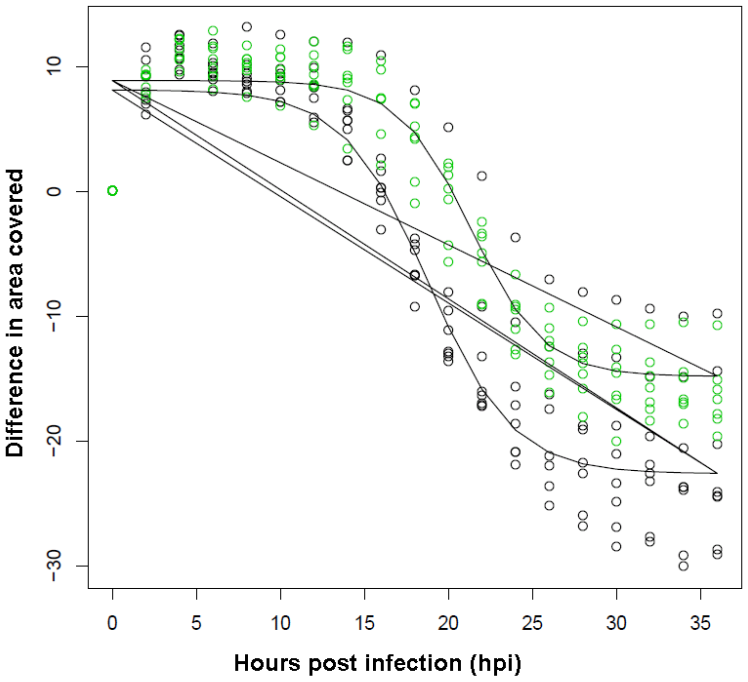


FIGURE 7.13: **The wildtype virus kills more cells, more quickly.** Model fitted by Dr. Simon Gubbins. A logistic (S-shaped) curve was fitted to the cell death data assuming the parameters were the same for both wildtype and mutant, the same process was completed assuming the parameters were different. The fit of the two curves were considered, and the curve with different parameters had a significantly better fit than the curve with common parameters. It can therefore be concluded that the parameters differ. From this model it can be seen that the wildtype (black) kills more cells (k2(wt)<k2(mut)) than the mutant (green) and earlier (d(wt)<d(mut)).

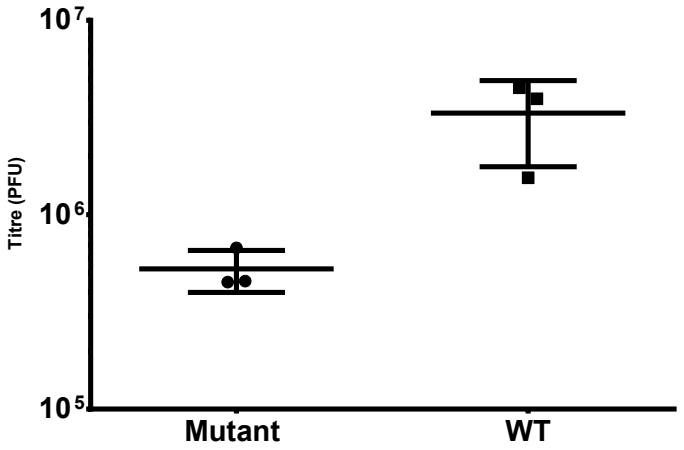


FIGURE 7.14: **The mutant virus produces a log less viable virus than the wildtype.** Triplicate plaque assays were completed on samples from equal MOI infection of mutant and wildtype virus at 5hpi and titre graphed on a log scale.

The delay in the mutant could be attributed to the decrease in packaged viable virus being produced. A plaque assay at five hours post infection showed a statistically significant log reduction in titre produced by the mutant in comparison to the wild-type: 5.3 x 10^5 PFU/mL and 3.3 x10^6 PFU/mL respectively (two-sample T-test, p-value=0.039)(Fig.7.14).

Although the sequence changes used to create the packaging mutant virus were synonymous it is important to consider if any of the genomic changes could have had an effect on known packaging requirements. There are a number of possible requirements on the genome to allow it to package. It has previously been shown that poliovirus only packages genomes linked to VPg. This could also be true for FMD. VPg is covalently linked to the 5'end of the FMDV genome [60]. The 5' end of the genome is unchanged by the mutations introduced. Reduction of VPg genes has a dose dependant effect on the quantity of viral RNA synthesised and virions produced [299]. There are three copies of VPg encoded within the FMDV genome, a unique feature of this virus among animal picornaviruses [300]. The mutations introduced do not sit in the region of the genome encoding the three copies of VPg. This suggests the mutations introduced are unlikely to have affected the linkage or production of VPg specifically. Furthermore, Falk et al [299] show the VPg reduction affects RNA synthesis and therefore virion formation although it is not required for translation [301]. This work has found that the titre is reduced but RNA synthesis remains the same (Fig. 7.14,7.11) suggesting the effect is not due to the genomes interaction and production of VPg.

It has also been shown that poliovirus only encapsidates newly synthesised RNA genomes, if this is also true in FMDV, mutations could have affected the number of newly synthesised RNA produced. Both wildtype and mutant genomes produced comparative amounts of RNA so it is unlikely this can explain the effect seen.

An important factor in virus encapsidation is genome compactness as previously discussed. Tubiana *et al* completed an experiment that modelled the effect of synonymous mutations throughout the genome on maximum ladder distance (MLD), a measure of genome compactness [302, 303]. They found that as mutations were introduced compactness was reduced in a gradual manner. This suggests that changes across the genome have a similar effect rather than certain regions being of maximum importance in terms of compactness. They also found that mutating 5% of genome positions or more was sufficient to eradicate the compactness observed in the wildtype virus. The number of nucleotides changed in the packaging mutant was 34. This accounts for only 0.4% of the genome and thus Tubiana *et al*'s model suggests should not result in a decrease in compactness. The targeted nature of the mutations made in my work (between nucleotide

positions 1700-3700) should not have had an increased effect as no evidence for genome hot spots dramatically affecting the compactness of RNA have been identified [303].

Non-structural protein 2C has previously been implicated in RNA encapsidation in some members of the picornavirus family although not directly demonstrated in FMDV. The mutations introduced are not in the coding region of 2C. However they are in the region encoding capsid proteins that 2C may interact with. As the changes are minimal and non-synonymous it is unlikely they have any effect on potential interactions between 2C and structural proteins.

As the mutations introduced are unlikely to have inadvertently affected known packaging requirements it is reasonable to suggest the previously unidentified predicted secondary structures identified in this study could be acting as packaging signals. In repeat experiments of the same strain in both FMDV and polio the results are not exactly the same with approximately half of the signals being evident in both repeats. I propose a model where a series of degenerate packaging signals exist along the genome. A selection of these are used by a given virus population, and the majority is reflected by conservation in the swarm. There are some that are used with greater prevalence that may have the strongest affinity for the capsid and represent the 50% overlap found in repeat experiment. These structures potentially have dual purpose, being required for processes such as replication, creating a system that allows for preferential packaging of replication competent genomes.

The idea that viral processes are not entirely distinct is not new. Due to the small genome length it is conceivable that RNA structures have dual purposes. For example the internal ribosomal entry site (IRES) of poliovirus has previously been found to have a role in encapsidation [304]. Experiments replacing the poliovirus IRES with the IRES from encephalomyocarditis virus (EMCV) [253], hepatitus C [305], human papilloma virus 14 (HPV14) [253] and coxsackievirus B3 [306] have all been completed previously. In each of these experiments the chimeric viruses showed slowed growth and a small plaque phenotype. This phenomenon has been further dissected by Johansen and Marrow [304] who created a poliovirus replicon with firefly luciferase in place of P1 and EMCV IRES in place of PV IRES. When packaged with vaccina virus expressing P1 the control PV replicon packaged at 37°C with a higher yield when passaged at 33°C. The EMCV-IRES replicon was unable to package efficiently at 37°C and, although possible to passage at 33°C, the yield was lower than the wildtype at either temperature variable. Both replicons were temperature stable and showed comparable amounts of luciferase fluorescence confirming comparable replication. The authors highlighted that the IRES was unlikely to be a singular packaging signal, as the EMCV-IRES replicon was unable to transencapsidate into EMCV capsids, but that the IRES might function as part of

the larger encapsidation process. This corroborates with the packaging signal identified in this region in both poliovirus experiments.

This theory that secondary structures are multifunctional is further validated by the identification of α and β in poliovirus 3D. These elements are functionally redundant but at least one is required for successful replication. In each poliovirus experiment one of these features was identified suggesting that these features are functionally redundant in replication and packaging.

In FMDV the IRES has also been identified in both experiments as a potential part of the packaging story. A further known structure, stem loop 1 (SL1), was also highlighted. SL1 is within the 3' non-coding region of FMDV between the coding region and poly-A tail. This region has previously been shown to be required for replication [307] and has also been identified by our analysis. As with previous studies mutating the stem loops in this region (PPS11) resulted in a replication incompetent virus (data not shown). It is therefore impossible to dissect the necessity of these structure for packaging without also affecting replication, but it is possible, as with the IRES in poliovirus, that these RNA structures are relevant for both processes.

This difficulty in separating other cellular functions from packaging coupled with the variable nature of the packaging signal repertoire explains why this viral process has been so difficult to elucidate. Previous experiments looking for single packaging signals have been misleading and led to the assumption, for example, that P1 has no elements relevant to packaging. The packing signals mutated in this work are all in a region that has been removed by cloning (P1). The effect is subtle, as have been shown by other experiments [304], because only a subset of the PPS are affected. This has potentially led to it being unnoticed in experiments where packaging is not the focus of the study.

7.5 Summary

These experiments have identified regions of the genome that are less conserved in the total population than in the encapsidated virions formed from it. These regions are predicted to form stem loop structures. This has been identified in FMDV and poliovirus. It is hypothesised that these regions are involved in the packaging of viral RNA. To asses this theory, non-synonymous mutations were introduced to create a mutant FMDV where some of these structures had been eradicated. It has been shown that the transfection and replication of both mutant and wildtype viruses is comparable (Fig. 7.10,7.11) but the titre and spread through the culture are lower and slower respectively

in the mutant (Fig.7.14,7.12) consistent with the conjecture that it is the encapsidation and not transcription or translation that was impeded by these mutations.

Chapter 8

Conclusions

8.1 Introduction

The study of RNA virus evolution typically considers the consensus-level viral genomes present in an infection. Understanding how the virus evolves and survives in new environments has focused on reverse genetics. The cloning of genetic features from virus genome to virus genome and observation of resulting phenotype has greatly expanded our understanding of how genomic change can affect virus survival. However, it is widely accepted that RNA viruses exist as a swarm of genetically related variants. This is a product of their generally error prone RNA-dependant RNA polymerases that lack proof reading capabilities, coupled with relatively short generation times and large populations. Much as the effects on fitness caused by changes in one part of the genome may depend on the precise nucleotide sequence (or genotype) in other parts of the genome, the evolution of a virus may be affected by the genetic variants present within the swarm - a form of epistasis. Work to fully consider genetic evolution of RNA viruses can no longer focus solely on the majority genome in the swarm, but must consider the swarm itself and the variants within it. This study has investigated the evolution of foot-and-mouth disease virus (FMDV) from this perspective.

8.2 FMDV swarm dynamics appear to be limited

Work was completed to consider how the dynamics of the swarm can impact fitness and phenotype. Previous experiments focusing on poliovirus have shown that mutations that increase polymerase fidelity, decrease viral virulence and fitness [308]. Equally, mutagens that decrease fidelity, such as ribavirin, result in error catastrophe [137]. This suggests that the mutant spectrum of an RNA virus like poliovirus is limited. This agrees with

the findings of this thesis that each isolate sequenced has relatively consistent cumulative entropy score (defined as the sum of the entropy score of each position of a genome). In this work, the range in this value was less than 100. Considering the approximate length of an FMDV genome is 8500nt, and entropy scores at each positions can range from 0 to 2, cumulative entropy can theoretically range from 0 (where every position is completely conserved) to 17000 (where each positions has an equal distribution of all four nucleotides). This suggests that although there is the capacity for variation in this measure, each virus swarm is able to withstand a comparable amount of variation within it.

Distributions of entropy scores contributing to this cumulative entropy score was investigated in Chapter 6. An experiment to observe how the swarm evolves under adaptive pressure of a new cell line *in vitro* was completed to understand how this swarm structure exists as a consequence of viral evolution. Analysis of the swarm structure showed that as consensus level changes appeared, a population under adaptive pressure and a comparative control population had a similar amount of variation within them. When considering this at a finer resolution it became apparent that although the overall level of variation was comparable (as represented by cumulative entropy) the distribution of entropy scores varied. One interpretation of these findings is that as variants increase and become fixed in the population, variation elsewhere in the genome is restricted. This means that although there are more high level entropy positions there are less mid-range entropy positions (with midrange positions now showing low entropy). This makes logical sense; due to the adaptive pressure, less of the genomes produced may be able to survive efficiently, restricting the overall size and variation of the swarm. Conversely, under neutral conditions without selection pressures, there are numerous genomes able to survive; consequently there are less high entropy positions associated with consensus level changes, but more midrange positions that contribute to a larger swarm. In addition, the swarm structure without adaptive pressure appears to cycle between two distributions of entropy scores; a tight distribution with a low mean and a broader distribution with a higher mean. Again this makes logical sense; as the virus replicates it produces numerous genomes some of which will be viable some will not. This large swarm of variation will undergo purifying selection and those that survive will produce a smaller less variable swarm of viruses.

Further work could be completed to understand the restrictions on the populations more fully. It would be interesting to observe if a more diverse swarm (with a larger mutant spectrum) evolved more efficiently than a less diverse swarm: i.e. does the size of the mutant spectrum directly affect a virus' evolvability? This could be achieved with experiments similar to the poliovirus polymerase experiment previously described [137, 308]. It is important to determine if the restriction on the amount of variation the

swarm can withstand is purely mechanical, as described above, or if the swarm has a limit of variation as a whole which would hint at it acting as a quasi-species which is discussed in more detail below.

8.3 The FMDV swarm may not be a quasi-species

The swarm within which FMDV exists is often termed a quasi-species [309]. However, whether the FMDV viral swarm actually acts as a quasi-species has yet to be determined. A 'molecular quasi-species' was first defined by Eigen, McCaskill and Schuster [150] who proposed a model of evolution that is strikingly different from Darwinian survival of the fittest. Their evolutionary theory highlights that it is not necessarily the individual that is evolving but the swarm of genetically related variants. Therefore, the term of 'fitness' is no longer relevant to an individual virus within a swarm but the mutation spectrum as a whole. Eigen *et al* suggest that the entire quasi-species is the target of selection. Swarm structure can be affected by a complex number of factors including environment and external parameters as well as competing neighbouring mutants in the swarm. Furthermore Eigen *et al*'s hypothesis that due to their large population sizes but small genome sizes viral swarms may include all possible options in the neutral space surrounding the master sequence. This suggests that because the sequences space is fully explored, the population does not undergo genetic drift into low fitness regions but instead evolves along high fitness 'ridges' associated with increased fitness. Jenkins *et al* explore this concept through modelling [310]. Their model argues against the concept of FMD viral swarms being refered to as a quasi-species. They suggest that the neutral space FMDV has the capacity to cover exceeds the populations size of a viral swarm suggesting it cannot form the quasi-species distribution Eigen *et al* describe. They propose instead a model they referred to as the 'random walk' model which fit more closely with traditional population genetics where each genome evolves independently.

The quasi-species being the target of selection rather than solely the genomes within it is difficult to separate. A study by Domingo *et al* was completed using the RNA phage Qβ and used T1 fingerprinting to show that comparing multiple passages showed as similar distribution of mutants, and these mutants were diverged from the majority by only one or two nucleotide changes. They suggested that these populations had a constant equilibrium, a factor that would be consistent with it behaving as a quasi-species [311]. A similar experiment in vesicular stomatitis virus (VSV) showed that after 523 low MOI infections the T1 oligonucleotide map for each passage remained identical [312]. This 'equilibrium' is perhaps hinted at by the consistent cumulative entropy score identified in this thesis. The consistent pattern seen in VSV in a low MOI

infection was very different from similar experiments at a high MOI. High MOI infections prompted rapid consensus level changes in and vastly more genetically different viruses within the population. This suggest perhaps that the equilibrium in the swarm found in these experiments is only apparent when no obvious adaptive pressure is present and would potentially be less evident in an *in vivo* infection. Equally, the high MOI infection increases swarm size introduced, and thus the potential for variation within it, and could explain the difference between the low and high MOI results. The experiments in this thesis have found however that the cumulative entropy score is consistent both with and without adaptive pressure.

8.4 Selection acts on the swarm at a sub-consensus level

Although the work in this thesis cannot demonstrate that selection acts on the swarm as a whole, there is evidence of selection acting on several different levels. In Chapter 4 work was completed to characterise the FMD viral swarm. It was found that selection could be identified at a within swarm level. Interestingly this level of selection differed at some genome positions from selection at the consensus level. This identifies the concept of FMDV proteins having functional roles that are important for replication (and observed at the consensus level) and for higher level processes which are only apparent at the within swarm level. Further work needs to be completed to use sub-consensus level data to understand the selection acting on the swarm more clearly. One protein, 2A, is thought to be highly conserved at the consensus level (as shown by Carillo *et al* [221]) but is highly variable within the swarm based on the data in this thesis. Understanding what level of selection is acting on this protein to produce this affect could help in understanding fundamental viral processes.

8.5 Sub-consensus level analysis can help reveal how diversity is created

Identification of these sub-consensus level features via deep-sequencing is a developing area of research. Methods are being designed to allow for accurate deep sequencing of RNA-swarms. These methods include, but are not limited to, the method outlined in this thesis [170], CirSeq (designed by Acevedo *et al* [157]) and barcoding methodologies [163, 164]. These techniques have allowed for the accurate creation of a mutation landscape for a poliovirus swarm [157] and the identification of low level drug resistant variants [154, 160, 161]. The identification of these low level drug resistant variants highlights that understanding the mutations present in a viral swarm can allow you to understand

what genomes may evolve from it. For example, in Chapter 5, sub-consensus level immune escape variants were identified in diverse SAT populations. Previously, the SAT serotypes have been considered genetically and antigenically more variable than their European counterparts (type A, type O and type C) [224, 225]. However, this thesis offers evidence that this variability may not necessarily be only a virus specific feature but a product of the host from which the isolate was derived. This finding changes the way in which SAT variability should be viewed. Furthermore, this work suggests the maintenance host, African buffalo, can sustain multiple infections simultaneously of closely related viral genomes. This agrees with a serological study of African buffalo in Zambia, completed by Sikombe *et al*, which found 68.6% of African buffalo that tested positive for FMDV harboured mixed infections [313]. However, serological data is not directly comparable due to this longevity of the immune response and cross-reactivity. This could provide an environment for between serotype recombination. Further work needs to be completed to understand why buffalo are able to maintain multiple infections and how this contributes to FMDV epidemiology in this region. Understanding how these infections are established is important to determine if it is a product of co-infection or evolution of a secondary population from the originally infecting population. Work to understand how this increased variability is passed from buffalo to cattle should be completed to determine if a mixed population is originally contracted but cannot be maintained in the bovine host and whether this more variable swarm affects disease progression in the bovine host.

8.6 Sub-consensus level analysis can inform fundamental virology studies

These types of deep sequencing experiments have the capacity to define the limits a virus must evolve within. This has been done more generally with error catastrophe experiments, but can be used more specifically to investigate functional requirements for viral processes. For example, in this thesis, interrogating NGS data of the viral swarm has been used to determine whether successfully encapsidated genomes are under any form of genomic restriction. Picornavirus packaging constraints are a contested area of research. Findings in different viruses within the family suggest different packaging processes. In this thesis a novel approach was considered to understand packaging constraints without amending the genetic code through deletion analysis or cloning. This allowed for the unbiased identification of relevant features without the risk of affecting other viral processes. Sequencing of the successfully encapsidated viral population in comparison to the total population from which it was derived revealed clusters of constrained nucleotides. Comparison of these regions found they contained conserved

stem loops with conservation of structures through multiple isolates. These structures appeared to be, to an extent, functionally redundant with approximately 50% present in replicates of the analysis. Sometimes the identified stem loops had already been found to be necessary for replication suggesting a system where replication competent genomes are preferentially packaged. Molecular removal of a subset of these structure through the introduction of synonymous changes produced a mutant virus which was replication competent but produced less viable virus and spread through the culture more slowly than the wildtype comparison. This was potentially due to a packaging defect. Further work can be completed to understand if an increasing number of the structures being removed results in a more obvious phenotype. Identification of the coding loop of the structure and how conserved this needs to be to function also needs to be completed. Work to clone these features into non-viral RNA to prompt packaging would be a relatively conclusive demonstration of the concept. Although the bioinformatic analysis was also completed in poliovirus, concurrent molecular experiments have not been completed, it would be interesting to establish if a similar effect would be evident in a poliovirus mutant. Finally it would also be interesting to establish if the proposed packaging signals are a general feature of more picornaviruses and to understand if this works as part of a larger packaging process involving other previously identified features such as protein chaperones. Work in the picornavirus Aichi virus has shown one stem loop at the beginning of the genome beneficial to packaging. However, the authors of this work have suggested this may not be the sole determinant [278, 279]. This would agree with what has been found in this thesis for FMDV. This novel technique allowed for the identification of regions not necessarily conserved in sequence (and thus not represented in the consensus) but conserved in structure.

8.7 Concluding remarks

In total this work highlights the importance of considering the viral swarm not just the consensus genome within it. From a diagnostic perspective the swarm can disguise immune escape mutants. Subconsensus level sequencing would allow for the identification of these mutants and subconsensus level strains to better inform vaccine strategy. Using swarm data in epidemiological studies may eventually allow for predictive models explaining virus evolution, as understanding the variation present within the swarm allows you to understand what can evolve from it. From a more fundamental perspective, analysis of the swarm allows the investigation of selective pressure acting on the virus at a subconesensus level. This has the capacity to help resolve fundamental virus traits and understand protein functions that may not yet be known.

Appendix A

CirSeq Comparison

An experiment was designed to compare CirSeq to the method used in this thesis (Table A.1).

CirSeq	Published Protocol
TRIzol Extraction	
DNase Treatment	
Ethanol Precipitation	
Fragmentation	
Elution	
Ligation	
Ethanol Precipitation	
Rolling Circle RT	RT
Second Strand Synthesis	
P:C:I Extraction	

TABLE A.1: **CirSeq comparison experimental design** An experiment was designed to compare elements of the CirSeq protocol with our published protocol [170]. The CirSeq protocol differs by including fragmentation of the genome, selection of the correctly sized fragments and their elution. Ligation of the fragments into circularised sections a clean up and subsequent rolling circle reverse transcription reaction.

The output from the two compared protocols (Table A.1) were sequenced on the Illumina MiSeq using the Nextera XT library preparation kit. The samples represented different percentages of the overall reads (CirSeq=4.11%, Published Protocol=5.54%). As 24 samples were multiplexed on this run each sample would expect to have ≈4.16% of the reads attributed to it. These two figures were within this range and relatively comparable.

Quality of the reads was considered using FastQC which provides a graphical output of the average quality score at each position along the read (Babraham Bioinformatics).

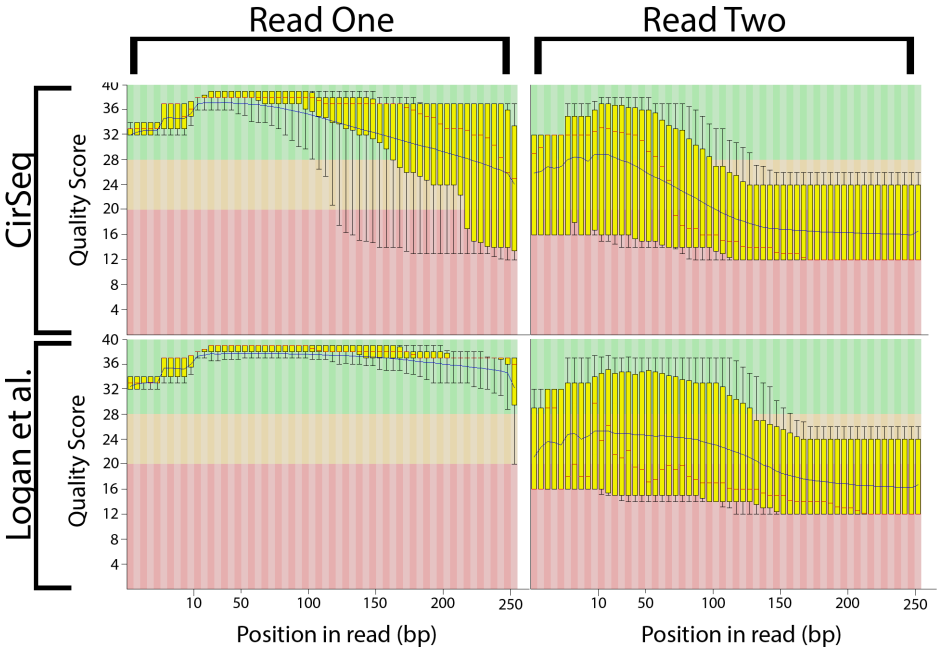


FIGURE A.1: **Quality score across the read** The average quality score at each position along the read for Read One (left hand side) and Read Two (right hand side) for CirSeq and the published protocol. The y axis represents quality score. Each base is represented by a BoxWhisker Plot showing the median value (red line), the inter-quartile range (yellow box) and the mean quality (blue line). Reads with a median Illumina quality score of 30 or above are consider to be of good quality, shaded in green. The x axis is separated into shaded bar representing 1 base for the first 10 bar and 5 bases for the subsequent bars.

The quality of read one is better than that of read two. At least 150 positions of read one have a 'good' mean and median quality score (Q>30) in both protocols. In comparison neither protocol have positions with 'good scores' in read two. The second run usually shows lower quality scores than the first but poor quality on this run was notable. After communication with Illumina it became apparent that the chemistry in the sequencing kit we were sent was not functional and as such the rotation of the fragments between reads was unsuccessful. Due to this, data from the second read was not used in subsequent analysis. The lower quality the latter end of read one of the CirSeq protocol could be attributed to shorter cDNA fragments due to a less efficient RT reaction caused by the circular nature of the template. Due to the fragmentation and subsequent rolling circle RT, long reads are dependant on the success of the RT reaction.

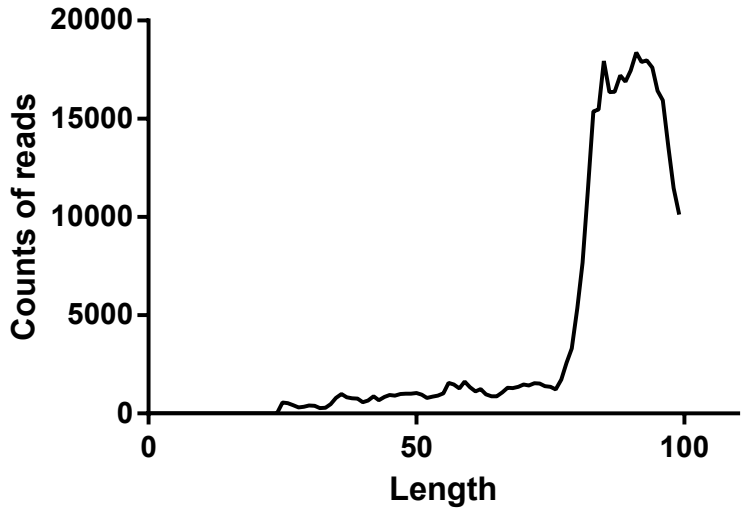


FIGURE A.2: **Fragment length produced by CirSeq Protocol** The CirSeq analysis software has an output that details the counts of reads with repeats from 25 - 99 bases long. This distribution can aid in diagnosis of issues with sequencing library preparation. The counts of reads (y-axis) is show in comparsion to read length (x-axis). This represents not actual length of the read but the length of the tandem repeat.

Maximum coverage was 375x and 10,226x for CirSeq and the published protocol respectively. This coverage for the CirSeq protocol was too low for the published software to make accurate observations regarding mutation frequency and transition/transversion ratios. In the work completed by Acevedo *et al* the samples run with the CirSeq protocol were not multiplexed [157]. In this work each sample is multiplexed in a total of 24 samples. As in CirSeq each read represents tandem repeats, the coverage that can be achieved is automatically 3x less than it could be. It could be that to achieve sufficient coverage for subsequent analysis samples need to be run non-multiplexed. The reads created in the protocol could also have been too short for the analysis. In order for the CirSeq analysis to be used each read must represent three repeats of the same region otherwise the read is discarded. As the quality scores suggest that some of the reads could be shorter than 250bp they would not contain three complete repeats and thus be discarded. Samples selected should have been between 80-90nt. The protocol did output some reads of the correct length however (Fig. A.2) and this was not entirely unsuccessful.

Appendix B

Un-used protocol optimisation steps

B.0.0.1 Removing host material

FMDV is a polyadenylated virus. The poly(A)tail at the 3' end of the genome is the site of one of the FMDV specific reverse transcription primers (Rev6) described in the published methodology paper by Logan *et al* [170]. Ribosomal RNA (rRNA) also has a poly(A)tail which acts to protect the RNA from degradation and is involved in it's functionality. Due to these common motifs there is some back ground level of rRNA in the sequencing data produced from the published protocol. To address this issue and improve the percentage of reads aligning to the FMDV genome an rRNA depleting kit was tested. RiboMinus (ThermoScientific) selectively depletes ribosomal RNA. RiboMinus involves a probe that specifically targets rRNA. This probe is linked to biotin allowing for rRNA to be removed using steptavadin bound magnetic beads. This kit has, as yet, only been validated against mouse and human ribosomal RNA. Therefore it was tested for it's specificity of bovine rRNA. RiboMinus depletion was completed as per manufacturers instructions and depletion of ribosomal RNA confirmed by analysing the sample before and after via gel electrophoresis on the Agilent bioanalyser.

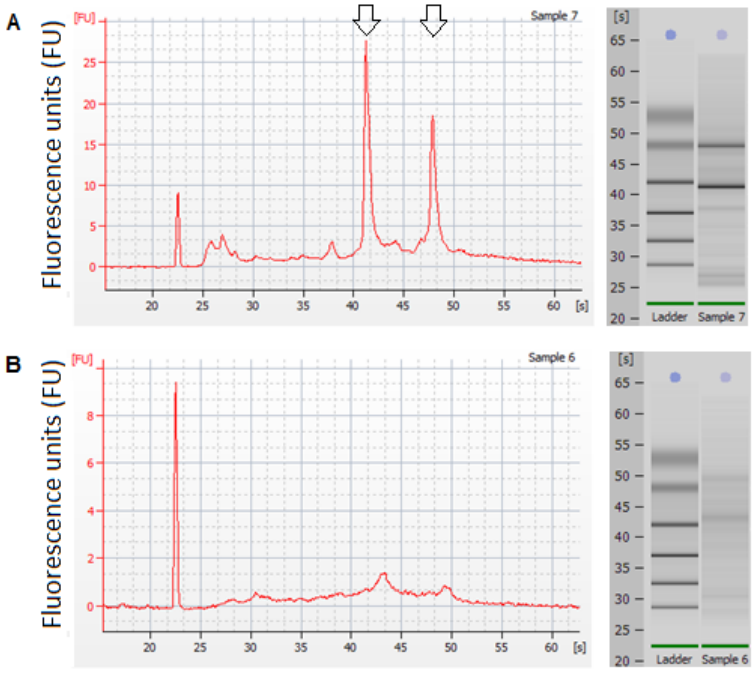


FIGURE B.1: **RiboMinus efficiently depletes bovine rRNA** Bioanalyser traces of Extracted RNA is shown A) before and B) after RiboMinus depletion. The trace shows in graphical form the fluorescence units (FU) produced at each bp over time (s). This is duplicated in a gel format. Experiment completed by Eleanor M. Cottam.

The first peak is a marker input during sample preparation. The small peaks present just after the marker in sample A represent 5S and 5.8S subunits, tRNA and small RNA fragments. The next two large peaks represent the 18S and 18S ribosomal peaks. These are clearly visible in the bioanalyser trace before treatment (Fig. B.1 A) but are depleted after RiboMinus treatment (Fig. B.1 B). This suggests riboMinus treatment efficiently reduces the ribosomal RNA present in the sample.

Having confirmed the kits ability to deplete bovine rRNA, RiboMinus was used in a direct sequencing comparison. In this comparison a 17 fold increase was seen in the number of reads aligning to the FMDV genome. Ribominus treatment also resulted in more even coverage and more of the genome being represented (only 21nt unrepresented compared to 148 in the non-treated sample). The increased amount of the genome covered and more even coverage is likely a product of the higher number of FMDV reads present.

Although this product provided marked improvement in number of reads aligned to the genome the cost was deemed impractical on the PhD budget. RiboMinus treatment costs approximately £95 per sample.

B.0.0.2 Improving end coverage

As detailed in the published method, coverage at the ends of the genome is lower or non-existent. This is particularly apparent at the 5' end of the genome. In an attempt to overcome this bias work was completed to circularise the genome.

To aid in the ligation of the genome an oligo 'bridge' was designed. This oligo was complimentary to the FMDV genome ends holding them closer together to increase the chance of ligation. This bridge was biotinylated to allow for specific selection of bound RNA using streptavidin beads. Primers were also designed to allow for PCR confirmation of the ligation of the genome ends. These oligs are detailed in Table B.1.

Oligo Name	Sequence (5'→3')
Oligo Bridge	ATATGCGATCGCCCTTRCGCCCCYTTTCAATTTTTTTTTTT TTTTTTT
Circ Primer 1.4	CGCGGTCCCGTGAGTCCAG
Circ Primer 2.1	TCCAGGCTACAGATCACTTTACCTGC
O-1C272F	TBGCRGGNCTYGCCCAGTACTAC
EUR-2B52R	GACATGTCCTCCTGCATCTGGTTGAT

TABLE B.1: **Oligos created for genome circularisation** Oligos were designed to bridge the ends of the genomes together to allow for efficient ligation (Oligo bridge). PCR primers were then designed to confirm this ligation had been successful (Circ Primer 1 and 2). Primers for VP1 were also used as a control (O-1C272 forward and EUR-2B52 reverse)

After TRIzol extraction and DNase treatment an annealing and ligation step was added to the protocol. An annealing solution was created by combining 100μL of 5M Sodium Chloride (NaCl), 1μL of 0.5M EDTA, 50μL of 1M Tris HCL and 4849μL of nfH_2O. An annealing reaction was assembled by combing 20μL of annealing solution (as described above) with 10μL of RNA and 5μL of Bridge Oligo (10μM). This reaction was heated to 72 degrees for three minutes to allow for the bridge to anneal to the RNA and then incubated on ice for three minutes to stop the reaction. To the annealing reaction 5μL of ligation buffer, 1 unit of RNase Ligase (NEB), 8μL of nfH_2O and 1 unit of RNase OUT(ThermoScientific) was added and the reaction was heated to 37°for 15 minutes and boiled for two minutes to terminate.

Two separate RT-PCR reaction was completed using the Qiagen One-step RT-PCR kit to check for the presence of FMDV genomes (VP1 primers O-1C272 forward and EUR-2B52 reverse, Table B.1) and check for successful circulariastion of the genome (Circ Primers, Table B.1).

The reaction mix included 8μL of nfH_2O, 5μL of 5x buffer, 1μL of dNTPs, 2.5μL of the forward and reverse primer required, 1 unit of enzyme and 2.5μL of RNA. Samples had an initial RT step of 50°for 30 minutes followed by a PCR activation step of 95°for 15 minutes. 45 cycles of denaturation (95°- 15 seconds), annealing (55°- 15 seconds) and extension (72°- 60 seconds) before a final extension step at 72°for 5 minutes.

PCR products were subjected to electrophoresis through a 1 % agarose gel containing ethidium bromide and visualised by trans illumination with UV light. The presence of a PCR product of 1000bp in the reaction containing VP1 primers indicated the presence of FMDV genome in the starting sample (Fig. B.2). The presence of product of 400bp in the reaction containing bridge primers indicated the sucessful circularisation of the genome. There was evidence of some non specific priming in the latter reaction as shown in the fainter bands representing different lengths of product.

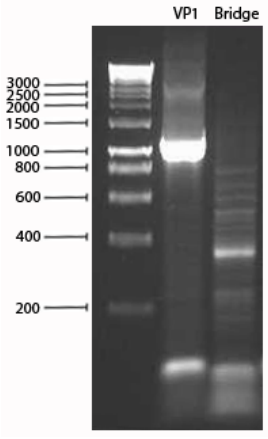


FIGURE B.2: **PCR of genome circularisation** Agarose gel electrophoresis products generated. Lane 1 contains HyperLadder 1, Lane two contains PCR product representing VP1 (O-K 272F and OK 272R) and lane three contains PCR product representing the bridge between the beginning and end of the genome(Circ Primer 1.4 and Circ Primer 2.1)

Circularised samples were used in RT and second strand synthesis reactions described above. PCI clean-up and subsequent Nextera XT prep was completed. Coverage of the genome termini was compared between a sample with and without genome circularisation.

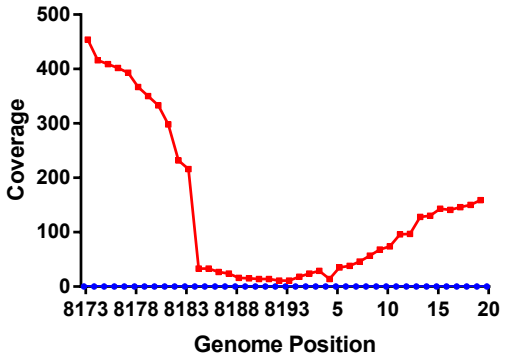


FIGURE B.3: **Circularised genomes offer better end coverage** Sample preparation including genome circularisation steps (red) and lacking genome circularisation steps (blue) was sequenced and aligned to a reference genome as detailed in the published protocol [170]. Coverage (y-axis) was established from the alignment using bed tools and graphed in comparison to genome position (x-axis) for the 20nt at each end of the genome. In the circularised genome all positions achieved some coverage (11-454x) in the non circularised sample no coverage was achieved for this region.

Without circularisation essentially no coverage of the first and last 20nt of the genome was achieved. With circularisation all of these region were represented with coverage ranging from 11x-454x (Fig. B.3). Therefore coverage achieved for the genome termini was improved with genome circularisation. However, there is still a noticeable dip. This suggests only a small proportion of genomes were successfully circularised. This is corroborated by the increased VP1 product in comparison to the bridge product (Fig. B.2).

Although this methodology clearly improves end coverage it introduces a notable bias. This selection of genomes bound to the bridge oligo results in the selection of only full length genomes. This would not be fully representative of the viral genomes present in the swarm as the large number of truncated genomes would not be included. Even without streptavaidin selection of bridge bound particles the extra annealing step decreases the copies of RNA present dramatically ($1.99x10^{11}$ copies lost). The high temperature termination step is likely to be resulting in degradation of genomes.

As the work in this thesis is considering the viral swarm rather than the genome termini genome circularisation was not used in general sample preparation. However, it should be noted as a useful asset when trying to achieve full genome sequences.

Appendix C

Adaptive passage low coverage investigation

C.0.0.1 Some samples had low overall coverage

Forty one samples were sequenced on the Illumina Miseq (11x population expansion, 15x BFA adaptation and 15x BHK control). In an attempt to achieve high coverage only 24 samples were multiplexed on each run instead of the maximum 96. This resulted in the samples being sequenced on two separate runs. Run one included inoculum, e1-10 and H1-8. Run two included H9-15 and F1-15. Both runs had acceptable quality scores with 80.50% of the reads in run one having an average quality score of 30 or above and 85.18 % of the reads in run two. The percentage of clusters on the flow cell that passed filter (%PF) was also good; 92.56% and 95.52% respectively. This is a measure of signal purity from each cluster and shows that clusters are not close enough together to produce problems with fluorescent signal overlap. The second run having higher values for each of these measures is likely due to it being under clustered. During the normalisation step of sample preparation for read two the shaker being used failed. This is reflected in the output. The incorrect normalisation procedure resulted in underclustering of the flow cell. Cluster density is an important parameter that can greatly affects run quality. Although under clustering does not affect the quality of the reads or the proportion of clusters passing filter, it does result in a lower data output. Run one produced 7.87GB or data compared to only 4.46GB of data for run two. As there were less clusters representing each sample, less reads were produced and therefore these samples had lower coverage.

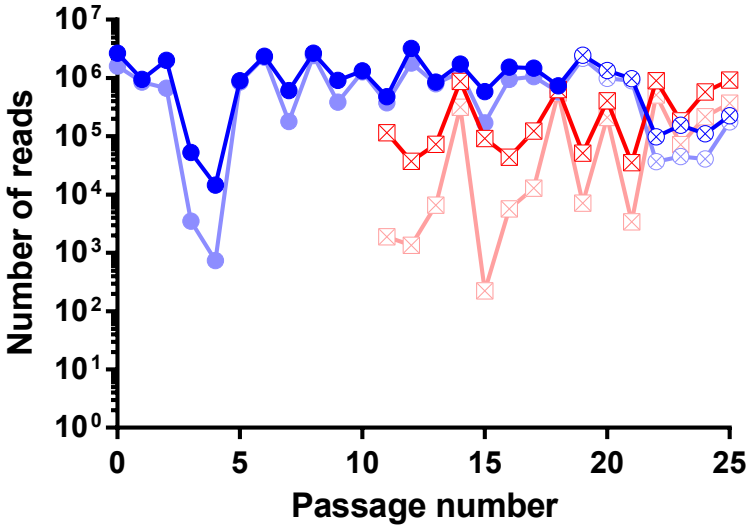


FIGURE C.1: **Samples with less reads attributed to them contain a lower percentage of FMDV reads.** The total number of reads attributed to each sequenced sample for virus passaged in BHK cells (dark blue, circle) and virus passaged in BFA (dark red, square). The number of reads that aligned to the FMDV genome is also shown for virus passaged in BHK cells (light blue, circle) and BFA cells (light red, square.) Passages marked with an 'x' were on run 2.

The majority of samples sequenced on the first run have approximately 1000000 reads attributed to them (average 1318024). The only two samples on this run that have less than this are e3 and e4 which have 52965 and 14516 reads respectively. The second read are in the same range as these two low samples with an average total number of reads as 473292.5. The number of reads that align to the FMDV genome is slightly less than than the total reads. This is to be expected as some reads associated with the bovine host are likely. This difference between the total reads and FMDV reads seems to be exaggerated the lower the number of total reads (Fig. C.1).

C.0.0.2 Low coverage equates to a lower percentage of reads being FMDV

Correlation analysis was completed to see if a decrease in the number of total reads resulted in a decrease in the proportion of the reads that aligned ot the FMDV genome.

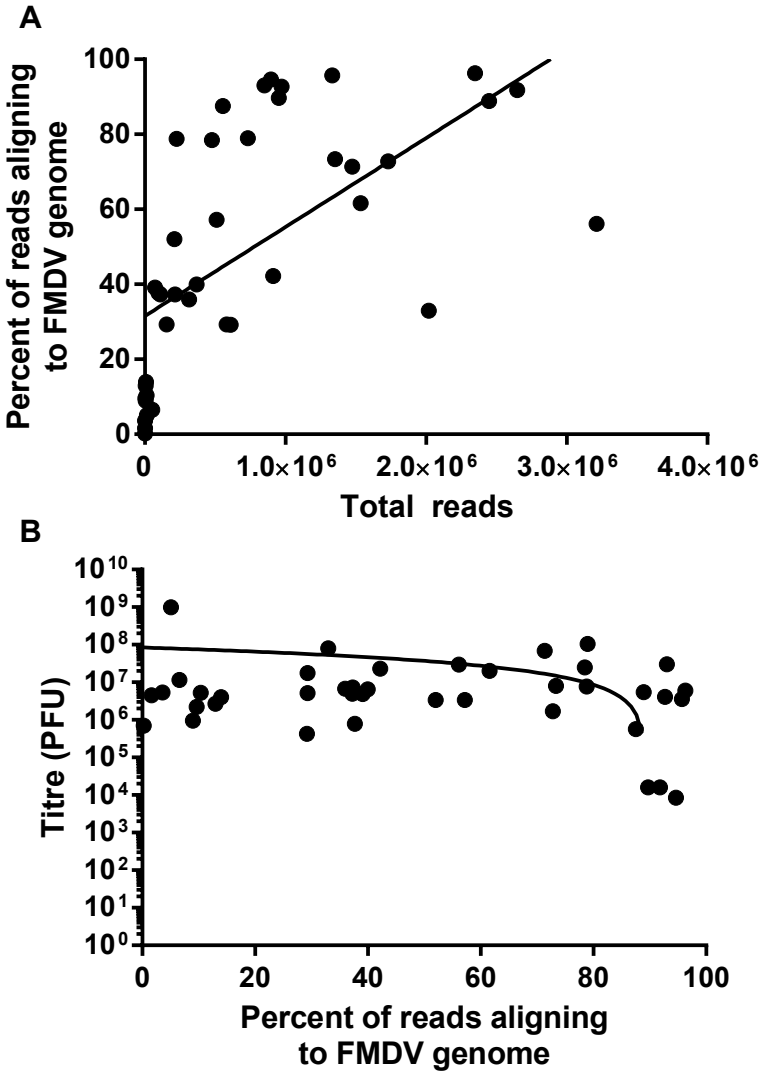


FIGURE C.2: **Decreased total number of reads correlates with the proportion of reads that are FMDV** A)Total reads against the percent of total reads that align to the FMDV genome is shown. The data is not normally distributed and thus a non-parametric test was completed to consider correlation. A two tailed Spearman correlation was completed with a 95% confidence interval. This showed there was significant correlation between the total number of reads and the percentage of these reads that aligned to the FMDV genome ($p<0.0001$). B) No significant correlation was found between viral titre and the number of reads aligning to the FMDV genome (Spearman's correlation, $p=0.9437$)

Comparing the percentage of reads identified as FMDV and the total number of reads showed a monotonic relationship. Correlation analysis showed a significant correlation between these two parameters (Spearmans, p=<0.0001)(Fig. C.2A).

This phenomenon is difficult to explain. One cause could be some level of inefficiency in the virus specific primers. This would decrease both the overall efficiency of the reverse

transcription (RT) reaction and the proportion of sample that were FMDV. What makes this unlikely is that each set of samples was processed as a batch of 5 i.e. F1-5, F6-10 and F11-15. In every instance where possible (i.e. the RT reaction) master mixes were made and thus it would be expected that any preparation induced issue would be evident in all five samples processed together which is not the case here. Each sample was Qubit quantified and diluted to 0.2ng/μL before library preparation so the amount of double stranded DNA at this point should be comparable. Qubit quantification relies upon fluorescent dies that are specific to the target of interest. These dyes are designed to emit only if bound to this molecule. This therefore does not differentiate between full length dsDNA and fragments or partially degraded DNA. Therefore it is possible that although the quantity of dsDNA was comparable the length of the DNA was not. Again this would have to have been caused by sample preparation steps but remains confusing as not all samples in each batch showed the same low percentage of FMDV reads. Gel electrophoresis of the sample at each stage of the process to asses nucleic acid length would help to asses if an issue arises during sample preparation but could result in lower overall yield due to loss of sample to each gel.

This could also have been caused by low titre of input virus. However when considering virus titre there appears to be no correlation between titre and percentage of reads aligning to FMDV (Spearmans, p=0.2439) (Fig. C.2B).

Due to these issues with normalisation and the proportion of reads aligning to the FMDV genome not all samples had a sufficient quantity of their genome represented by >1000x coverage for further analysis.

Appendix D

Scripts Used

Listing D.1 and D.2 were written by Rocio Enriques-Gasca at TGAC. The shell script nt count.sh runs an Rscript eleano.r which counts the number of each nucleotide represented in the sequencing data at each position along the genome and outputs this information in tabular format.

```
#!/bin/sh


i=$1

#source R-2.15.0
awk '{print $1"\t"$2"\t"$3"\t"$4"\t"$5}' $i | sed 's/+.\w
   \+/+/g' | sed 's/-.\w\+/-/g' | sed 's/\^.//g' | sed 's/\
   $//g' > filtered_${i}
Rscript eleano.r filtered_${i}
rm filtered_${i}
```

LISTING D.1: nt count.sh -Written by Rocio Enriquez-Gasca

```
args<-commandArgs(TRUE)

pos<-read.table(args[1], sep="\t", header=FALSE)

cero<-rnorm((nrow(pos)*9), 0, 0)
val<-matrix(cero, nrow(pos), 9)
#a=1, c=2,g=3,t=4
```

```
states<-c("A", "C", "G", "T", "\\+", "\\-")

for(i in 1:(nrow(pos))){
        bases<-strsplit(as.character(pos$V5[i]), split="")
        val[i, 1]<-pos$V1[i]
        val[i, 2]<-pos$V2[i]
        val[i, 3]<-pos$V3[i]
                if(pos$V3[i]=="A"){
                val[i, 4]<-length(grep("\\.", bases[[1]]))+
   length(grep(",", bases[[1]]))
                val[i, 5]<-(length(grep("c", bases[[1]]))+
   length(grep("C", bases[[1]])))
                val[i, 6]<-(length(grep("g", bases[[1]]))+
   length(grep("G", bases[[1]])))
                val[i, 7]<-(length(grep("t", bases[[1]]))+
   length(grep("T", bases[[1]])))
                val[i, 8]<-(length(grep("\\-", bases[[1]]))
   )
                val[i, 9]<-(length(grep("\\+", bases[[1]]))
   )
        }else if(pos$V3[i]=="C"){
                val[i, 4]<-(length(grep("a", bases[[1]]))+
   length(grep("A", bases[[1]])))
                val[i, 5]<-length(grep("\\.", bases[[1]]))+
   length(grep(",", bases[[1]]))
                val[i, 6]<-(length(grep("g", bases[[1]]))+
   length(grep("G", bases[[1]])))
                val[i, 7]<-(length(grep("t", bases[[1]]))+
   length(grep("T", bases[[1]])))
                val[i, 8]<-(length(grep("\\-", bases[[1]]))
   )
                val[i, 9]<-(length(grep("\\+", bases[[1]]))
   )
        }else if(pos$V3[i]=="G"){
                val[i, 4]<-(length(grep("a", bases[[1]]))+
   length(grep("A", bases[[1]])))
                val[i, 5]<-(length(grep("c", bases[[1]]))+
   length(grep("C", bases[[1]])))
```

```
                val[i, 6]<-length(grep("\\.", bases[[1]]))+
    length(grep(",", bases[[1]]))
                val[i, 7]<-(length(grep("t", bases[[1]]))+
    length(grep("T", bases[[1]])))
                val[i, 8]<-(length(grep("\\-", bases[[1]]))
    )
                val[i, 9]<-(length(grep("\\+", bases[[1]]))
    )
        }else if(pos$V3[i]=="T"){
                val[i, 4]<-(length(grep("a", bases[[1]]))+
    length(grep("A", bases[[1]])))
                val[i, 5]<-(length(grep("c", bases[[1]]))+
    length(grep("C", bases[[1]])))
                val[i, 6]<-(length(grep("g", bases[[1]]))+
    length(grep("G", bases[[1]])))
                val[i, 7]<-length(grep("\\.", bases[[1]]))+
    length(grep(",", bases[[1]]))
                val[i, 8]<-(length(grep("\\-", bases[[1]]))
    )
                val[i, 9]<-(length(grep("\\+", bases[[1]]))
    )
        }
}
write.table(val, "bases.txt", quote=FALSE, sep="\t", row.
    names=FALSE, col.names=c("Seq Name", "Position", "Base
    in the reference", "A", "C", "G", "T", "Insertions", "
    Deletions"))
```

LISTING D.2: eleano.r - Written by Rocio Enriquez-Gasca

EntropyCalc.py (listing D.3) is a python script that uses R to calculate Shannon's entropy as out put from the nt count script included above.

```
#!/usr/bin/python

##GracieLogan##
'''usage - python EntropyCa.py inputfile.txt'''

import sys
```

```
from rpy import r

bases=(sys.argv[1]) #import file

f=open('Entropy.txt','w') #make new file called Entropy
f.write('EntropyScore')
f.write('\n')
f.close()

with open (bases) as fi: #calling the open file f
        for line in fi: #for each line in f
                try:
#turn each item into a float (as split by split)
                        numbers_float = map(float,line.
    split())
#creating a sum by adding total of ACG and T together
                        sum = ((numbers_float [3])+(
    numbers_float [4])+(numbers_float [5])+(numbers_float
    [6]))
                        try:
                                Ae1 = str((((numbers_float
    [3])/sum)*(r.log2((numbers_float [3])/sum))))
#A entropy (turned into a string)
                                if Ae1 == 'nan':
#if A entropy is not a number
                                        Ae1 = 0
#report it as zero
                                else:
                                        Ae1 = Ae1
                        except ZeroDivisionError:
                                Ae1=0
                        try:
                                Ce1 = str((((numbers_float
    [4])/sum)*(r.log2((numbers_float [4])/sum))))
                                if Ce1 == 'nan':
                                        Ce1 = 0 #report it
    as zero
                                else:
                                        Ce1 = Ce1
```

```
                        except ZeroDivisionError:
                                Ce1=0
                        try:
                                Ge1 = str((((numbers_float
    [5])/sum)*(r.log2((numbers_float [5])/sum))))
                                if Ge1 == 'nan':
                                        Ge1 = 0
                                else:
                                        Ge1 = Ge1
                        except ZeroDivisionError:
                                Ge1=0
                        try:
                                Te1 = str((((numbers_float
    [6])/sum)*(r.log2((numbers_float [6])/sum))))
                                if Te1 == 'nan':
                                        Te1 = 0
                                else:
                                        Te1 = Te1
                        except ZeroDivisionError:
                                Te1=0
#changing string back into a float
                        Ae = float(Ae1)
                        Ce = float(Ce1)
                        Ge = float(Ge1)
                        Te = float(Te1)
#adding it together
                        EntropyScore = -(Ae+Ce+Ge+Te)
                        f1=open('Entropy.txt','a')
                        f1.write(str(EntropyScore))
                        f1.write('\n')
                        f1.close()
                except ValueError:
                        pass
```

LISTING D.3: EntropyCalc.py

NS mpileup cutoff.1.0.2.py is a python script written by Nick Sanderson that implements a coverage cut-off on an mpileup file.

```
#!/usr/bin/env python
import sys
# usage pipe to -> | python NS_mpileup_cuttoff.py limit |
# where limit equals your coverage cut-off
target = sys.stdin # data from pipe |
limit = int(sys.argv[1])


def ignoreIndel(seq, qual):  # function to get limit of
   each
    n = 0        #  total number of reads
    s = 0        # total number in string
    seq2 = []    # list to add sequences strig to
    while n < limit:  # while total n reads is below limit
        if seq[s] in OK_list:    # increase n if bona-fide
   read in OK_list
            s = s + 1
            n = n + 1
        elif seq[s] in insertion:   ## this adds teh exact
   amount dependent on the insertion
            s = s + 2
            s = s + int(seq[s-1])
#        elif seq[s] in deletion:   ## this adds the exact
   amount dependent on the deletion
#            s = s - 1
#            n = n + 1
#            s = s + int(seq[s-1])
        else:
            s = s + 1      # go to next in string
    for r in range(s):     # iterate through total string
   range which contains n reads
        seq2.append(seq[r])  # add to list
    seq3 = "".join(seq2)     # join list to create string
    return seq3      # return string from function


insertion = set()
#deletion = set()
insertion.add("+")
```

```
insertion.add("-")
#deletion.add("*")


OK_list = set() # create OK_list set, list of allowed
    sequence

OK_list.add(".")
OK_list.add(",")
OK_list.add("a")
OK_list.add("A")
OK_list.add("c")
OK_list.add("C")
OK_list.add("t")
OK_list.add("T")
OK_list.add("g")
OK_list.add("G")
OK_list.add("*")
#OK_list.add("$")  # these don't actually count as reads,
    just signal the end of a read, F and K represent quality
     too
#OK_list.add("^")

for line in target: # for each line from the pipe
    l = line.split("\t")    # split line on the tab to make
    list l
    if int(l[3]) > limit:   # if greater than limit than do
        sequence = str(l[4])
        quality = str(l[5])
        qual4 = quality[0:limit]
        seq4 = ignoreIndel(sequence, quality)  # run
    function with sequence string to generate cut off seq4
        print str(l[0]) + "\t" + str(l[1]) + "\t" + str(l
    [2]) + "\t" + str(limit) + "\t" + seq4  + "\t" + qual4
    else:                   # or do this
        print line.replace("\n","")
```

Listing D.4: NS mpileup cutoff.1.0.2.py - Written by Nick Sanderson

Appendix E

Comparison Genomes

Samples used in MLT Fig.5.7.

Genbank Number	Isolate
HM067706.1	SAT1/Buff/17/QE
HM067705.1	SAT2/Buff/10/QE
HM067704.1	SAT2/Buff/6/QE
AF540910.1	SAT2/ZIM/7/83
KM268901.1	SAT3/ZIM/6/91
KM268900.1	SAT2/TAN/5/2012
KM268899.1	SAT1/TAN/22/2012
KJ820999.1	SAT3/UGA/1/13
KC440884.1	SAT2/EGY/3/2012
JX014256.1	SAT2/PAT/1/2012
JX014255.1	SAT2/EGY/9/2012
KU821592.1	SAT2/ZAM18/2009
JF749864.1	SAT2/ZIM/22/2003
JF749862.1	SAT2/UGA/002/2002
JF749861.1	SAT2/KEN/002/2002
JF749860.1	SAT1/KEN/004/2002
FJ461346.1	SAT2/MFNP
KU821590.1	SAT1/NAM01/2010
KR108950.1	SAT3/KNP/10/90
KR108949.1	SAT2/KNP/19/89
KR108948.1	SAT1/KNP/196/91/1
AY593853.1	SAT3/4bech/iso23
AY593852.1	SAT3/3kenya/iso22

AY593851.1	SAT3/3bech/iso29
AY593850.1	SAT3/2sa/iso27
AY593849.1	SAT2/3kenya/21
AY593848.1	SAT2/iso25
AY593847.1	SAT2/1rhod/iso26
AY593846.1	SAT1/rhod/iso33
AY593845.1	SAT1/bot/iso47
AY593844.1	SAT1/7isrl/iso12
AY593843.1	SAT1/6swa/iso16
AY593842.1	SAT1/5sa/iso13
AY593841.1	SAT1/4srhod/iso24
AY593840.1	SAT1/3swa/iso14
AY593839.1	SAT1/20/iso11
AY593838.1	SAT1/1bech/iso30
AJ251473.1	SAT2/KEN/3/57
Unpublished	O1K
Unpublished	SAT3/BOT/109/66
Unpublished	SAT3/BOT/BUFF/13/70
Unpublished	SAT3/RHO/7/74
Unpublished	SAT3/RV/7/34
Unpublished	SAT3/UGA/BUFF/27/70
Unpublished	SAT3/ZIM/2/84
Unpublished	SAT3/ZIM/P6/83/BUFF/19
Unpublished	SAT1/UGA/47/71
Unpublished	SAT2/ZIM/8/89
Unpublished	SAT2/ZAM/BUFF/18/74
Unpublished	SAT2/UGA/BUFF/24/70
Unpublished	SAT2/UGA/BUF/12/70
Unpublished	SAT2/SRHO/1/65
Unpublished	SAT1/NR/1/64
Unpublished	SAT2/ETH/1/90
Unpublished	SAT2/BOT/BUFF/170/74
Unpublished	SAT2/BOT/BUFF/107/72
Unpublished	SAT2/BOT/BUFF/17/69
Unpublished	SAT2/BOT/BUFF/7/72
Unpublished	SAT2/BOT/BUFF/2/69
Unpublished	SAT2/BOT/BUFF/2/68
Unpublished	SAT1/UGA/BUF/10/70
Unpublished	SAT1/SWA/2/89

Unpublished	SAT1/MOZ/5/81
Unpublished	SAT1/K303/83
Unpublished	SAT1/BOT/BUFF/62/74
	QuRe80.21477356739344
	QuRe18.349784802998503
	QuRe1.4354416296080705

TABLE E.1: **Genomes used to create MLT**

Samples used in LocaRNA analysis (Chapter 7).

Genbank Number	Isolate
AY593826.1	type O2brescia, iso17, 1947
AY593825.1	type O1valle, iso64, 1953
AY593812.1	type O10phil, iso76, 1958
AY593830.1	type O7polland, iso49, 1959
AY593813.1	type O1indonesia, iso52, 1962
AY593837.1	type Ouraguay, iso51, 1963
AY593820.1	type Ocanefa, iso59, 1964
AY593814.1	type O1argentina, iso5, 1965
AY593834.1	type Orey-iran, iso53, 1966
AY593815.1	type O1bfs, iso18, 1967
AY593823.1	type O1manisa, iso87, 1969
AY593827.1	type O3venezuela, iso15, 1971
AY593817.1	type O1brugge, iso79, 1973
AY593835.1	type Otaiwan, iso106/112, 1997
AY593833.1	type Openghu, iso108, 1999
AJ539139.1	type O, strain SKR/2000
AJ633821.1	type O/FRA/1/2001
EF614457.1	type O/SKR/14/02
HQ632770.1	type O MAY/1/2004
HQ632771.1	type O MAY/8/2005
FJ175666.1	type Oisrael 07-6387
GU384682.1	type O/PAK/44/2008
JN998085.1	type O/BY/CHA/2010

TABLE E.2: **Genomes used for LocARNA Analysis**

Appendix F

Additional Coverage Graphs

F.1 Coverage Zam/Nan/11

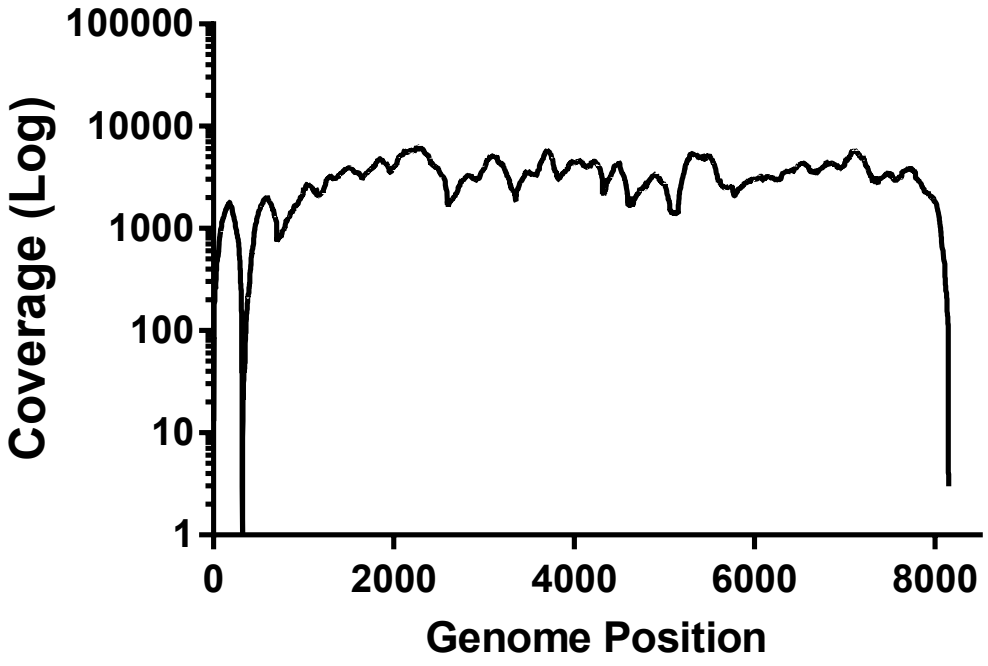


FIGURE F.1: **Coverage of Zam/Nan/11 (Log scale)** The coverage of each position a long the genome is shown

Zam/Nan/11 showed an unusual distribution of nucleotides in Chapter 5, Fig. 5.3 at 6000-7000nt. This does not equate to a dramatic fluctuation in coverage Fig. F.1. There is some variation in the coverage (Fig. F.2) but this is evident from 2000-7000nt. The

unusual distribution of nucleotides does not span this entire area suggesting the two are not related.

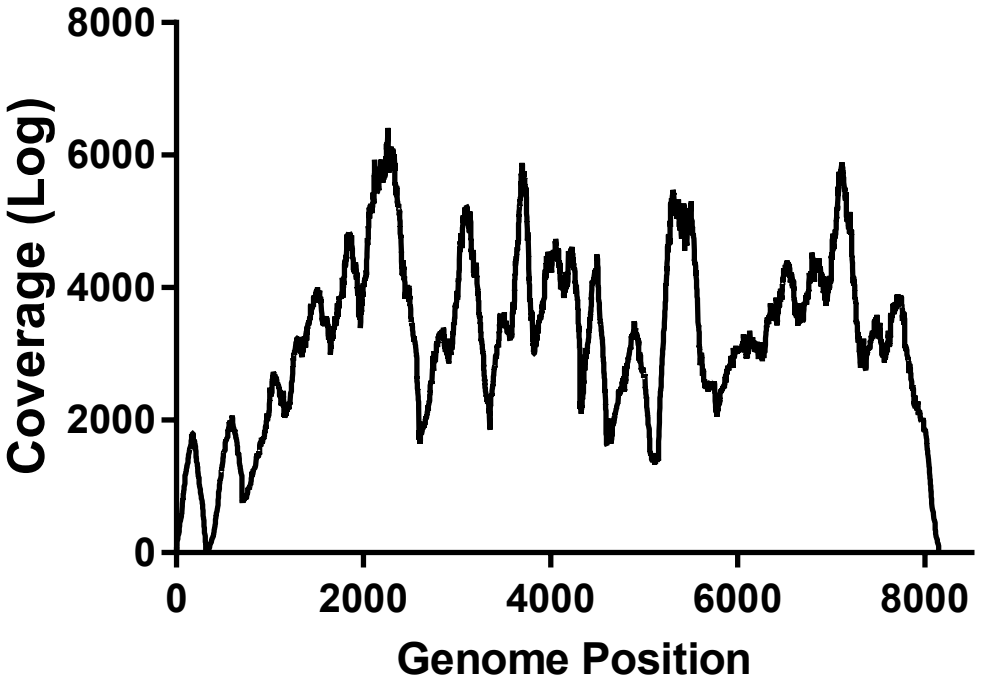


FIGURE F.2: **Coverage of Zam/Nan/11** The coverage of each position a long the genome is shown

F.2 Coverage KNP/196/91

SAT KNP/196/91 was used in a haplotype reconstruction in Chapter 5. Typically QuRe is less successful in low coverage genome regions. The point at which the reconstruction fails (after the structural region) does not coincide with drop in coverage.

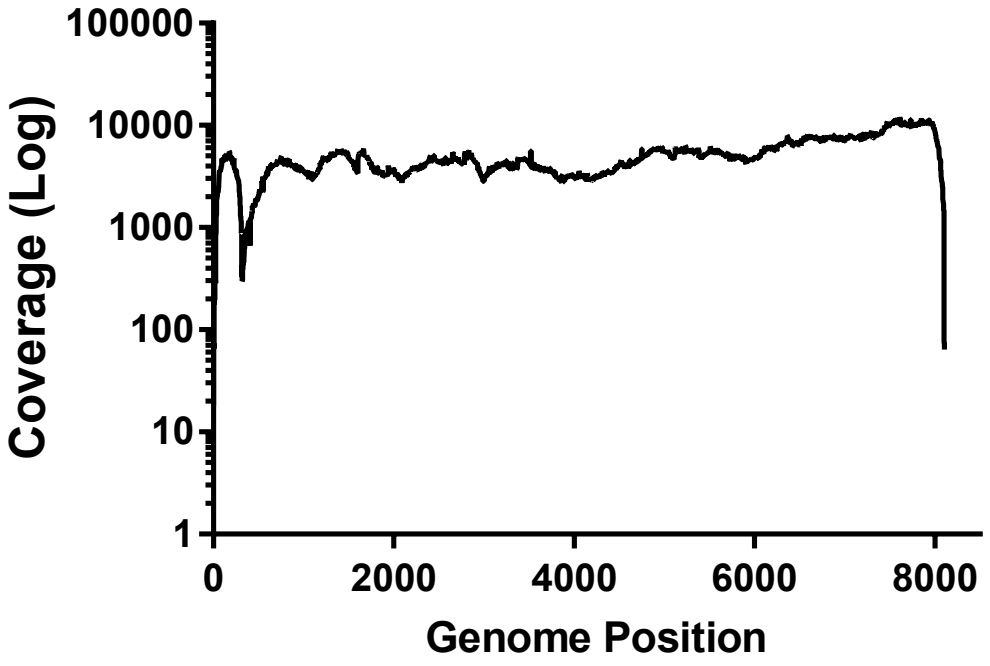


FIGURE F.3: **Coverage of KNP/196/91 (Log scale** The coverage of each position a long the genome is shown

Bibliography

[1] H. Valle and H. Carr. Sur le pluralit des virus aphteusess. *Comput. Rend. Acad. Sci.*, 174:207–208, 1922.

[2] O. Waldmann and K. Trautwein. Experimentelle Untersuchungen ber die Pluralitt des Maul-und Klauenseuchevirus. *Berl. Tierrztl. Wochenschrift*, 42:569–571, 1926.

[3] M.R. Dhanda, V.R. Gopalakrishnan, and H.S. Dhillon. Note on the occurrence of atypical strains of foot-and-mouth diseases virus in India. *Ind. J. Vet. Sci.*, 27: 79–84, 1957.

[4] J. B. Brooksby and J. Rogers. Methods used in typing the virus of foot-and-mouth disease at Pirbright, 195055. *Methods of typing and cultivation of foot-and-mouth disease virus, European Productivity Agency of the Organisation for European Economic Co-operation*, 1957.

[5] J. B. Brooksby. The virus of foot-and-mouth disease. *Adv. Virus Res.*, 5:1–37, 1958.

[6] OIE (WORLD ORGANISATION FOR ANIMAL HEALTH). Oie member countries' official fmd status map, 2016. URL http://www.oie.int/en/animal-health-in-the-world/official-disease-status/fmd/en-fmd-carte/.

[7] R. P. Kitching. A recent history of foot-and-mouth disease. *J. Comp. Pathol.*, 118 (2):89–108, Feb 1998.

[8] P.L. Roeder and N.J. Knowles. Foot-and-mouth disease virus type c situation: the first target for eradication? the global control of fmd tools, ideas and ideals, 2008. URL http://www.fao.org/ag/againfo/commissions/docs/research_group/erice/APPENDIX_07.pdf.

[9] N. J. Knowles and A. R. Samuel. Molecular epidemiology of foot-and-mouth disease virus. *Virus Res*, 91, 2003.

[10] D. M. Ansell, A. R. Samuel, W. C. Carpenter, and N. J. Knowles. Genetic relationships between foot-and-mouth disease type Asia 1 viruses. *Epidemiol. Infect.*, 112(1):213–224, Feb 1994.

[11] The Royal Society. *Infectious diseases in livestock - policy document 19/02*. The Royal Society, London, 2002. ISBN 085403580X.

[12] DEFRA Publications. The Strategy for achieving Officially Bovine Tuberculosis Free status for England, 2014. URL https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/300447/pb14088-bovine-tb-strategy-140328.pdf.

[13] S. E. Roche, M. G. Garner, R. L. Sanson, C. Cook, C. Birch, J. A. Backer, C. Dube, K. A. Patyk, M. A. Stevenson, Z. D. Yu, T. G. Rawdon, and F. Gauntlett. Evaluating vaccination strategies to control foot-and-mouth disease: a model comparison study. *Epidemiol. Infect.*, 143(6):1256–1275, Apr 2015.

[14] Council of the European Union. Council directive 2003/85/ec on community measures for the control of foot-and-mouth disease repealing directive 85/511/eec and decisions 89/531/eec and 91/665/eec and amending directive 92/46/eec., 2003. URL http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:306:0001:0087:EN:PDF.

[15] D. Thompson, P. Muriel, D. Russell, P. Osborne, A. Bromley, M. Rowland, S. Creigh-Tyte, and C. Brown. Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001. *Rev. - Off. Int. Epizoot.*, 21(3):675–687, Dec 2002.

[16] J. Rushton. *The economics of animal health and production*. Centre for Agriculture and Biosciences International, Egham, UK, 2008. ISBN 9781845931940.

[17] T. J. Knight-Jones and J. Rushton. The economic impacts of foot and mouth disease - what are they, how big are they and where do they occur? *Prev. Vet. Med.*, 112(3-4):161–173, Nov 2013.

[18] G.R. Thomson and A.D.S. Bastos. *Infectious Diseases of Livestock*, volume 2. Oxford University Press, Oxford, 3 edition, 2004. ISBN 9780195761719. Ed. J.A.W. Coetzer and R.C. Tustin.

[19] G. R. Thomson, W. Vosloo, and A. D. Bastos. Foot and mouth disease in wildlife. *Virus Res.*, 91(1):145–161, Jan 2003.

[20] P. Sutmoller and A. Gaggero. Foot-and mouth diseases carriers. *Vet. Rec.*, 77(33): 968–969, Aug 1965.

[21] P. Moonen, L. Jacobs, A. Crienen, and A. Dekker. Detection of carriers of foot-and-mouth disease virus among vaccinated cattle. *Vet. Microbiol.*, 103(3-4):151–160, Nov 2004.

[22] S. Alexandersen, Z. Zhang, and A. I. Donaldson. Aspects of the persistence of foot-and-mouth disease virus in animals–the carrier problem. *Microbes Infect.*, 4 (10):1099–1110, Aug 2002.

[23] J. Arzt, J. M. Pacheco, and L. L. Rodriguez. The early pathogenesis of foot-and-mouth disease in cattle after aerosol inoculation. Identification of the nasopharynx as the primary site of infection. *Vet. Pathol.*, 47(6):1048–1063, Nov 2010.

[24] C. C. Brown, M. E. Piccone, P. W. Mason, T. S. McKenna, and M. J. Grubman. Pathogenesis of wild-type and leaderless foot-and-mouth disease virus in cattle. *J. Virol.*, 70(8):5638–5641, Aug 1996.

[25] C. C. Brown, R. F. Meyer, H. J. Olander, C. House, and C. A. Mebus. A pathogenesis study of foot-and-mouth disease in cattle, using in situ hybridization. *Can. J. Vet. Res.*, 56(3):189–193, Jul 1992.

[26] S. Alexandersen, M. B. Oleksiewicz, and A. I. Donaldson. The early pathogenesis of foot-and-mouth disease in pigs infected by contact: a quantitative time-course study using TaqMan RT-PCR. *J. Gen. Virol.*, 82(Pt 4):747–755, Apr 2001.

[27] R. P. Kitching and S. Alexandersen. Clinical variation in foot and mouth disease: pigs. *Rev. - Off. Int. Epizoot.*, 21(3):513–518, Dec 2002.

[28] F. Sobrino and E. Domingo. *Foot and mouth disease : current perspectives*. CRC Press, Boca Raton, Florida, 2004. ISBN 9780849329517.

[29] W. Vosloo, L. M. de Klerk, C. I. Boshoff, B. Botha, R. M. Dwarka, D. Keet, and D. T. Haydon. Characterisation of a SAT-1 outbreak of foot-and-mouth disease in captive African buffalo (Syncerus caffer): clinical symptoms, genetic characterisation and phylogenetic comparison of outbreak isolates. *Vet. Microbiol.*, 120(3-4):226–240, Mar 2007.

[30] D. J. Paton and D. P. King. Diagnosis of foot-and-mouth disease. *Dev Biol (Basel)*, 135:117–123, 2013.

[31] P. Sutmoller, S. S. Barteling, R. C. Olascoaga, and K. J. Sumption. Control and eradication of foot-and-mouth disease. *Virus Res.*, 91(1):101–144, Jan 2003.

[32] E. Ehrenfeld, E. Domingo, and R. P. Roos. *The picornaviruses*. American Society for Microbiology Press, Washington DC, 2010. ISBN 9781555816032.

[33] S. Alexandersen, Z. Zhang, A. I. Donaldson, and A. J. Garland. The pathogenesis and diagnosis of foot-and-mouth disease. *J. Comp. Pathol.*, 129(1):1–36, Jul 2003.

[34] S. Alexandersen and A. I. Donaldson. Further studies to quantify the dose of natural aerosols of foot-and-mouth disease virus for pigs. *Epidemiol. Infect.*, 128 (2):313–323, Apr 2002.

[35] OIE (WORLD ORGANISATION FOR ANIMAL HEALTH). *Manual of Diagnostic Tests and Vaccines for Terrestrial Animals*. World Organisation for Animal Health, Paris, 2012.

[36] R. Reeve, B. Blignaut, J. J. Esterhuysen, P. Opperman, L. Matthews, E. E. Fry, T. A. de Beer, J. Theron, E. Rieder, W. Vosloo, H. G. O'Neill, D. T. Haydon, and F. F. Maree. Sequence-based prediction for vaccine strain selection and identification of antigenic variability in foot-and-mouth disease virus. *PLoS Comput. Biol.*, 6(12):e1001027, 2010.

[37] N. Mattion, N. Goris, T. Willems, B. Robiolo, E. Maradei, C. P. Beascoechea, A. Perez, E. Smitsaart, N. Fondevila, E. Palma, K. De Clercq, and J. La Torre. Some guidelines for determining foot-and-mouth disease vaccine strain matching by serology. *Vaccine*, 27(5):741–747, Jan 2009.

[38] R. Ranjan, M. Kangayan, S. Subramaniam, J. K. Mohapatra, J. K. Biswal, G. K. Sharma, A. Sanyal, and B. Pattnaik. Development and evaluation of a one step reverse transcription-loop mediated isothermal amplification assay (RT-LAMP) for rapid detection of foot and mouth disease virus in India. *Virusdisease*, 25(3): 358–364, 2014.

[39] K. Morioka, K. Fukai, K. Yoshida, R. Kitano, R. Yamazoe, M. Yamada, T. Nishi, and T. Kanno. Development and Evaluation of a Rapid Antigen Detection and Serotyping Lateral Flow Antigen Detection System for Foot-and-Mouth Disease Virus. *PLoS ONE*, 10(8):e0134931, 2015.

[40] The center for food security and public health. Vaccines: Foot and mouth disease, 2016. URL http://www.cfsph.iastate.edu/Vaccines/disease_list.php/disease_list.php?disease=foot-and-mouth-disease&lang=en.

[41] T. R. Doel. FMD vaccines. *Virus Res.*, 91(1):81–99, Jan 2003.

[42] C. Porta, A. Kotecha, A. Burman, T. Jackson, J. Ren, S. Loureiro, I. M. Jones, E. E. Fry, D. I. Stuart, and B. Charleston. Rational engineering of recombinant picornavirus capsids to produce safe, protective vaccine antigen. *PLoS Pathog.*, 9 (3):e1003255, Mar 2013.

[43] A. Kotecha, J. Seago, K. Scott, A. Burman, S. Loureiro, J. Ren, C. Porta, H. M. Ginn, T. Jackson, E. Perez-Martin, C. A. Siebert, G. Paul, J. T. Huiskonen, I. M. Jones, R. M. Esnouf, E. E. Fry, F. F. Maree, B. Charleston, and D. I. Stuart. Structure-based energetics of protein interfaces guides foot-and-mouth disease virus vaccine design. *Nat. Struct. Mol. Biol.*, 22(10):788–794, Oct 2015.

[44] D. Aarthi, K. Ananda Rao, R. Robinson, and V. A. Srinivasan. Validation of binary ethyleneimine (BEI) used as an inactivant for foot and mouth disease tissue culture vaccine. *Biologicals*, 32(3):153–156, Sep 2004.

[45] N. J. Knowles. Picornaviridae, 2016. URL http://www.picornaviridae.com/.

[46] Polio gobal eradication initiative. Global eradication of wild poliovirus type 2 declared, 2015. URL http://www.polioeradication.org/mediaroom/newsstories/Global-eradication-of-wild-poliovirus-type-2-declared/tabid/526/news/1289/Default.aspx.

[47] N. Lewis-Rogers and K. A. Crandall. Evolution of Picornaviridae: an examination of phylogenetic relationships and cophylogeny. *Mol. Phylogenet. Evol.*, 54(3):995–1005, Mar 2010.

[48] R. Basavappa, R. Syed, O. Flore, J. P. Icenogle, D. J. Filman, and J. M. Hogle. Role and mechanism of the maturation cleavage of VP0 in poliovirus assembly: structure of the empty capsid assembly intermediate at 2.9 A resolution. *Protein Sci.*, 3(10):1651–1669, Oct 1994.

[49] R. Acharya, E. Fry, D. Stuart, G. Fox, D. Rowlands, and F. Brown. The three-dimensional structure of foot-and-mouth disease virus at 2.9 A resolution. *Nature*, 337(6209):709–716, Feb 1989.

[50] E. E. Fry, J. W. Newman, S. Curry, S. Najjam, T. Jackson, W. Blakemore, S. M. Lea, L. Miller, A. Burman, A. M. King, and D. I. Stuart. Structure of Foot-and-mouth disease virus serotype A10 61 alone and complexed with oligosaccharide receptor: receptor conservation in the face of antigenic variation. *J. Gen. Virol.*, 86(Pt 7):1909–1920, Jul 2005.

[51] A. Kotecha. Vaccine design at the atomic level, 2013. URL http://www.labnews.co.uk/features/vaccine-design-at-the-atomic-level-3-11-07-2013/.

[52] T. Jackson, D. Sheppard, M. Denyer, W. Blakemore, and A. M. King. The epithelial integrin alphavbeta6 is a receptor for foot-and-mouth disease virus. *J. Virol.*, 74(11):4949–4956, Jun 2000.

[53] T. Jackson, A. P. Mould, D. Sheppard, and A. M. King. Integrin alphavbeta1 is a receptor for foot-and-mouth disease virus. *J. Virol.*, 76(3):935–941, Feb 2002.

[54] T. Jackson, S. Clark, S. Berryman, A. Burman, S. Cambier, D. Mu, S. Nishimura, and A. M. King. Integrin alphavbeta8 functions as a receptor for foot-and-mouth disease virus: role of the beta-chain cytodomain in integrin-mediated infection. *J. Virol.*, 78(9):4533–4540, May 2004.

[55] A. Burman, S. Clark, N. G. Abrescia, E. E. Fry, D. I. Stuart, and T. Jackson. Specificity of the VP1 GH loop of Foot-and-Mouth Disease virus for alphav integrins. *J. Virol.*, 80(19):9798–9810, Oct 2006.

[56] K. Chamberlain, V. L. Fowler, P. V. Barnett, S. Gold, J. Wadsworth, N. J. Knowles, and T. Jackson. Identification of a novel cell culture adaptation site on the capsid of foot-and-mouth disease virus. *J. Gen. Virol.*, 96(9):2684–2692, Sep 2015.

[57] S. Berryman, S. Clark, N. K. Kakker, R. Silk, J. Seago, J. Wadsworth, K. Chamberlain, N. J. Knowles, and T. Jackson. Positively charged residues at the five-fold symmetry axis of cell culture-adapted foot-and-mouth disease virus permit novel receptor interactions. *J. Virol.*, 87(15):8735–8744, Aug 2013.

[58] F. Brown and B. Cartwright. Dissociation of foot-and-mouth disease virus into its nucleic acid and protein components. *Nature*, 192:1163–1164, Dec 1961.

[59] S. Forss, K. Strebel, E. Beck, and H. Schaller. Nucleotide sequence and genome organization of foot-and-mouth disease virus. *Nucleic Acids Res.*, 12(16):6587–6601, Aug 1984.

[60] D. V. Sangar, D. J. Rowlands, T. J. Harris, and F. Brown. Protein covalently linked to foot-and-mouth disease virus RNA. *Nature*, 268(5621):648–650, Aug 1977.

[61] T. Lyons, K. E. Murray, A. W. Roberts, and D. J. Barton. Poliovirus 5'-terminal cloverleaf RNA is required in cis for VPg uridylylation and the initiation of negative-strand RNA synthesis. *J. Virol.*, 75(22):10696–10708, Nov 2001.

[62] W. Xiang, K. S. Harris, L. Alexander, and E. Wimmer. Interaction between the 5'-terminal cloverleaf and 3AB/3CDpro of poliovirus is essential for RNA replication. *J. Virol.*, 69(6):3658–3667, Jun 1995.

[63] H. Toyoda, D. Franco, K. Fujita, A. V. Paul, and E. Wimmer. Replication of poliovirus requires binding of the poly(rC) binding protein to the cloverleaf as

well as to the adjacent C-rich spacer sequence between the cloverleaf and the internal ribosomal entry site. *J. Virol.*, 81(18):10017–10028, Sep 2007.

[64] E. Rieder, T. Bunch, F. Brown, and P. W. Mason. Genetically engineered foot-and-mouth disease viruses with poly(C) tracts of two nucleotides are virulent in mice. *J. Virol.*, 67(9):5139–5145, Sep 1993.

[65] C. Escarmis, J. Dopazo, M. Davila, E. L. Palma, and E. Domingo. Large deletions in the 5'-untranslated region of foot-and-mouth disease virus of serotype C. *Virus Res.*, 35(2):155–167, Feb 1995.

[66] M. A. Devaney, V. N. Vakharia, R. E. Lloyd, E. Ehrenfeld, and M. J. Grubman. Leader protein of foot-and-mouth disease virus is required for cleavage of the p220 component of the cap-binding protein complex. *J. Virol.*, 62(11):4407–4409, Nov 1988.

[67] D. Wang, L. Fang, P. Li, L. Sun, J. Fan, Q. Zhang, R. Luo, X. Liu, K. Li, H. Chen, Z. Chen, and S. Xiao. The leader proteinase of foot-and-mouth disease virus negatively regulates the type I interferon pathway by acting as a viral deubiquitinase. *J. Virol.*, 85(8):3758–3766, Apr 2011.

[68] M. Medina, E. Domingo, J. K. Brangwyn, and G. J. Belsham. The two species of the foot-and-mouth disease virus leader protein, expressed individually, exhibit the same activities. *Virology*, 194(1):355–359, May 1993.

[69] M. L. Donnelly, G. Luke, A. Mehrotra, X. Li, L. E. Hughes, D. Gani, and M. D. Ryan. Analysis of the aphthovirus 2A/2B polyprotein 'cleavage' mechanism indicates not a proteolytic reaction, but a novel translational effect: a putative ribosomal 'skip'. *J. Gen. Virol.*, 82(Pt 5):1013–1025, May 2001.

[70] D. Ao, H. C. Guo, S. Q. Sun, D. H. Sun, T. S. Fung, Y. Q. Wei, S. C. Han, X. P. Yao, S. Z. Cao, D. X. Liu, and X. T. Liu. Viroporin Activity of the Foot-and-Mouth Disease Virus Non-Structural 2B Protein. *PLoS ONE*, 10(5):e0125828, 2015.

[71] N. L. Teterina, A. E. Gorbalenya, D. Egger, K. Bienz, M. S. Rinaudo, and E. Ehrenfeld. Testing the modularity of the N-terminal amphipathic helix conserved in picornavirus 2C proteins and hepatitis C NS5A protein. *Virology*, 344 (2):453–467, Jan 2006.

[72] T. R. Sweeney, V. Cisnetto, D. Bose, M. Bailey, J. R. Wilson, X. Zhang, G. J. Belsham, and S. Curry. Foot-and-mouth disease virus 2C is a hexameric AAA+ protein with a coordinated ATP hydrolysis mechanism. *J. Biol. Chem.*, 285(32): 24347–24359, Aug 2010.

[73] D. P. Gladue, V. O'Donnell, R. Baker-Branstetter, L. G. Holinka, J. M. Pacheco, I. Fernandez-Sainz, Z. Lu, E. Brocchi, B. Baxt, M. E. Piccone, L. Rodriguez, and M. V. Borca. Foot-and-mouth disease virus nonstructural protein 2C interacts with Beclin1, modulating virus replication. *J. Virol.*, 86(22):12080–12090, Nov 2012.

[74] Z. Cheng, J. Yang, H. Xia, Y. Qiu, Z. Wang, Y. Han, X. Xia, C. F. Qin, Y. Hu, and X. Zhou. The nonstructural protein 2C of a Picorna-like virus displays nucleic acid helix destabilizing activity that can be functionally separated from its ATPase activity. *J. Virol.*, 87(9):5205–5218, May 2013.

[75] V. Maroudam, S. B. Nagendrakumar, P. N. Rangarajan, D. Thiagarajan, and V. A. Srinivasan. Genetic characterization of Indian type O FMD virus 3A region in context with host cell preference. *Infect. Genet. Evol.*, 10(5):703–709, Jul 2010.

[76] J. M. Pacheco, T. M. Henry, V. K. O'Donnell, J. B. Gregory, and P. W. Mason. Role of nonstructural proteins 3A and 3B in host range and pathogenicity of foot-and-mouth disease virus. *J. Virol.*, 77(24):13017–13027, Dec 2003.

[77] J. M. Pacheco and P. W. Mason. Evaluation of infectivity and transmission of different Asian foot-and-mouth disease viruses in swine. *J. Vet. Sci.*, 11(2):133–142, Jun 2010.

[78] X. Ma, P. Li, P. Sun, Z. Lu, H. Bao, X. Bai, Y. Fu, Y. Cao, D. Li, Y. Chen, Z. Qiao, and Z. Liu. Genome sequence of foot-and-mouth disease virus outside the 3A region is also responsible for virus replication in bovine cells. *Virus Res.*, 220:64–69, Jul 2016.

[79] V. N. Vakharia, M. A. Devaney, D. M. Moore, J. J. Dunn, and M. J. Grubman. Proteolytic processing of foot-and-mouth disease virus polyproteins expressed in a cell-free system from clone-derived transcripts. *J. Virol.*, 61(10):3199–3207, Oct 1987.

[80] V. R. Wells, S. J. Plotch, and J. J. DeStefano. Determination of the mutation rate of poliovirus RNA-dependent RNA polymerase. *Virus Res.*, 74(1-2):119–132, Apr 2001.

[81] C. Ferrer-Orta, A. Arias, R. Perez-Luque, C. Escarmis, E. Domingo, and N. Verdaguer. Structure of foot-and-mouth disease virus RNA-dependent RNA polymerase and its complex with a template-primer RNA. *J. Biol. Chem.*, 279(45):47212–47221, Nov 2004.

[82] V. O'Donnell, M. LaRocco, H. Duque, and B. Baxt. Analysis of foot-and-mouth disease virus internalization events in cultured cells. *J. Virol.*, 79(13):8506–8518, Jul 2005.

[83] F. M. Ellard, J. Drew, W. E. Blakemore, D. I. Stuart, and A. M. King. Evidence for the role of His-142 of protein 1C in the acid-induced disassembly of foot-and-mouth disease virus capsids. *J. Gen. Virol.*, 80 (Pt 8):1911–1918, Aug 1999.

[84] A. Panjwani, M. Strauss, S. Gold, H. Wenham, T. Jackson, J. J. Chou, D. J. Rowlands, N. J. Stonehouse, J. M. Hogle, and T. J. Tuthill. Capsid protein VP4 of human rhinovirus induces membrane permeability by the formation of a size-selective multimeric pore. *PLoS Pathog.*, 10(8):e1004294, Aug 2014.

[85] G. J. Belsham and C. J. Bostock. Studies on the infectivity of foot-and-mouth disease virus RNA using microinjection. *J. Gen. Virol.*, 69 (Pt 2):265–274, Feb 1988.

[86] N. Luz and E. Beck. A cellular 57 kDa protein binds to two regions of the internal translation initiation site of foot-and-mouth disease virus. *FEBS Lett.*, 269(2): 311–314, Sep 1990.

[87] E. V. Pilipenko, T. V. Pestova, V. G. Kolupaeva, E. V. Khitrina, A. N. Poperechnaya, V. I. Agol, and C. U. Hellen. A cell cycle-dependent protein serves as a template-specific translation initiation factor. *Genes Dev.*, 14(16):2028–2045, Aug 2000.

[88] L. Saleh, R. C. Rust, R. Fullkrug, E. Beck, G. Bassili, K. Ochs, and M. Niepmann. Functional interaction of translation initiation factor eIF4G with the foot-and-mouth disease virus internal ribosome entry site. *J. Gen. Virol.*, 82(Pt 4):757–763, Apr 2001.

[89] G. J. Belsham, G. M. McInerney, and N. Ross-Smith. Foot-and-mouth disease virus 3C protease induces cleavage of translation initiation factors eIF4A and eIF4G within infected cells. *J. Virol.*, 74(1):272–280, Jan 2000.

[90] A. V. Gamarnik and R. Andino. Switch from translation to RNA replication in a positive-stranded RNA virus. *Genes Dev.*, 12(15):2293–2304, Aug 1998.

[91] G. A. Belov and E. Sztul. Rewiring of cellular membrane homeostasis by picornaviruses. *J. Virol.*, 88(17):9478–9489, Sep 2014.

[92] P. Monaghan, H. Cook, T. Jackson, M. Ryan, and T. Wileman. The ultrastructure of the developing replication site in foot-and-mouth disease virus-infected BHK-38 cells. *J. Gen. Virol.*, 85(Pt 4):933–946, Apr 2004.

[93] A. V. Paul, E. Rieder, D. W. Kim, J. H. van Boom, and E. Wimmer. Identification of an RNA hairpin in poliovirus RNA that serves as the primary template in the in vitro uridylylation of VPg. *J. Virol.*, 74(22):10359–10370, Nov 2000.

[94] J. Herold and R. Andino. Poliovirus RNA replication requires genome circularization through a protein-protein bridge. *Mol. Cell*, 7(3):581–591, Mar 2001.

[95] K. E. Murray and D. J. Barton. Poliovirus CRE-dependent VPg uridylylation is required for positive-strand RNA synthesis but not for negative-strand RNA synthesis. *J. Virol.*, 77(8):4739–4750, Apr 2003.

[96] S. J. Flint, L. W. Enquist, V. R. Racaniello, and A. M. Skalka. *Principles of Virology: Molecular Biology, Pathogenesis, and Control of Animal Viruses*. ASM Press, Washington D.C., 2003. ISBN 9781555812591.

[97] A. E. Gorbalenya and E. V. Koonin. Viral proteins containing the purine NTP-binding sequence pattern. *Nucleic Acids Res.*, 17(21):8413–8440, Nov 1989.

[98] J. E. Novak and K. Kirkegaard. Improved method for detecting poliovirus negative strands used to demonstrate specificity of positive-strand encapsidation and the ratio of positive to negative strands in infected cells. *J. Virol.*, 65(6):3384–3387, Jun 1991.

[99] C. I. Nugent, K. L. Johnson, P. Sarnow, and K. Kirkegaard. Functional coupling between replication and packaging of poliovirus replicon RNA. *J. Virol.*, 73(1): 427–435, Jan 1999.

[100] W. A. Hagan and Bruner. D. W. *Hagan and Bruner's Microbiology and Infectious Diseases of Domestic Animals*. Comstock Publishing Associates, New York, USA, isbn = 0801418968, 1988.

[101] T. Yilma, J. W. McVicar, and S. S. Breese. Pre-lytic release of foot-and-mouth disease virus in cytoplasmic blebs. *J. Gen. Virol.*, 41(1):105–114, Oct 1978.

[102] Y. H. Chen, W. Du, M. C. Hagemeijer, P. M. Takvorian, C. Pau, A. Cali, C. A. Brantner, E. S. Stempinski, P. S. Connelly, H. C. Ma, P. Jiang, E. Wimmer, G. Altan-Bonnet, and N. Altan-Bonnet. Phosphatidylserine vesicles enable efficient en bloc transmission of enteroviruses. *Cell*, 160(4):619–630, Feb 2015.

[103] M. J. Grubman and B. Baxt. Foot-and-mouth disease. *Clin. Microbiol. Rev.*, 17 (2):465–493, Apr 2004.

[104] M. A. de La Vega, D. Stein, and G. P. Kobinger. Ebolavirus Evolution: Past and Present. *PLoS Pathog.*, 11(11):e1005221, 2015.

[105] P. J. Gerrish and J. G. Garcia-Lerma. Mutation rate and the efficacy of antimicrobial drug treatment. *Lancet Infect Dis*, 3(1):28–32, Jan 2003.

[106] M. Eigen. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10):465–523, Oct 1971.

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

[108] P. Farci, A. Shimoda, A. Coiana, G. Diaz, G. Peddis, J. C. Melpolder, A. Strazzera, D. Y. Chien, S. J. Munoz, A. Balestrieri, R. H. Purcell, and H. J. Alter. The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science*, 288(5464):339–344, Apr 2000.

[109] A. S. Lauring and R. Andino. Exploring the fitness landscape of an RNA virus by using a universal barcode microarray. *J. Virol.*, 85(8):3780–3791, Apr 2011.

[110] E. A. Duarte, I. S. Novella, S. C. Weaver, E. Domingo, S. Wain-Hobson, D. K. Clarke, A. Moya, S. F. Elena, J. C. de la Torre, and J. J. Holland. RNA virus quasispecies: significance for viral disease and epidemiology. *Infect Agents Dis*, 3 (4):201–214, Aug 1994.

[111] S. J. Martin, A. C. Highfield, L. Brettell, E. M. Villalobos, G. E. Budge, M. Powell, S. Nikaido, and D. C. Schroeder. Global honey bee viral landscape altered by a parasitic mite. *Science*, 336(6086):1304–1306, Jun 2012.

[112] M. Vignuzzi, J. K. Stone, J. J. Arnold, C. E. Cameron, and R. Andino. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature*, 439(7074):344–348, Jan 2006.

[113] J. Dopazo, F. Sobrino, E. L. Palma, E. Domingo, and A. Moya. Gene encoding capsid protein VP1 of foot-and-mouth disease virus: a quasispecies model of molecular evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 85(18):6811–6815, Sep 1988.

[114] C. R. Pringle. Evidence of genetic recombination in foot-and-mouth disease virus. *Virology*, 25:48–54, Jan 1965.

[115] E. M. Cottam, D. T. Haydon, D. J. Paton, J. Gloster, J. W. Wilesmith, N. P. Ferris, G. H. Hutchings, and D. P. King. Molecular epidemiology of the foot-and-mouth disease virus outbreak in the united kingdom in 2001. *J Virol*, 80, 2006.

[116] M. J. Morelli, C. F. Wright, N. J. Knowles, N. Juleff, D. J. Paton, D. P. King, and D. T. Haydon. Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Vet. Res.*, 44:12, 2013.

[117] B. Gaschen, J. Taylor, K. Yusim, B. Foley, F. Gao, D. Lang, V. Novitsky, B. Haynes, B. H. Hahn, T. Bhattacharya, and B. Korber. Diversity considerations in HIV-1 vaccine selection. *Science*, 296(5577):2354–2360, Jun 2002.

[118] E. Domingo. Rapid evolution of viral RNA genomes. *J. Nutr.*, 127(5 Suppl): 958S–961S, May 1997.

[119] E. Domingo, E. Baranowski, J. I. Nunez, C. M. Ruiz-Jarabo, S. Sierra, N. Molina, and F. Sobrino. [Quasispecies and molecular evolution of viruses]. *Rev. - Off. Int. Epizoot.*, 19(1):55–63, Apr 2000.

[120] E. Domingo, E. Baranowski, C. M. Ruiz-Jarabo, A. M. Martin-Hernandez, J. C. Saiz, and C. Escarmis. Quasispecies structure and persistence of RNA viruses. *Emerging Infect. Dis.*, 4(4):521–527, 1998.

[121] E. Domingo, C. Escarmis, N. Sevilla, A. Moya, S. F. Elena, J. Quer, I. S. Novella, and J. J. Holland. Basic concepts in RNA virus evolution. *FASEB J.*, 10(8): 859–864, Jun 1996.

[122] C. Ferrer-Orta, A. Arias, R. Perez-Luque, C. Escarmis, E. Domingo, and N. Verdaguer. Sequential structures provide insights into the fidelity of RNA replication. *Proc. Natl. Acad. Sci. U.S.A.*, 104(22):9463–9468, May 2007.

[123] M. Kozak. Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. *Mol. Cell. Biol.*, 9(11):5134–5142, Nov 1989.

[124] L. A. Wagner, R. B. Weiss, R. Driscoll, D. S. Dunn, and R. F. Gesteland. Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. *Nucleic Acids Res.*, 18(12):3529–3535, Jun 1990.

[125] D. A. Steinhauer and J. J. Holland. Rapid evolution of RNA viruses. *Annu. Rev. Microbiol.*, 41:409–433, 1987.

[126] J. W. Drake. A constant rate of spontaneous mutation in DNA-based microbes. *Proc. Natl. Acad. Sci. U.S.A.*, 88(16):7160–7164, Aug 1991.

[127] J. W. Drake. Rates of spontaneous mutation among RNA viruses. *Proc. Natl. Acad. Sci. U.S.A.*, 90(9):4171–4175, May 1993.

[128] J. W. Drake. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Ann. N. Y. Acad. Sci.*, 870:100–107, May 1999.

[129] D. T. Haydon, A. R. Samuel, and N. J. Knowles. The generation and persistence of genetic variation in foot-and-mouth disease virus. *Prev. Vet. Med.*, 51(1-2): 111–124, Sep 2001.

[130] M. A. Martinez, J. Dopazo, J. Hernandez, M. G. Mateu, F. Sobrino, E. Domingo, and N. J. Knowles. Evolution of the capsid protein genes of foot-and-mouth disease virus: antigenic variation without accumulation of amino acid substitutions over six decades. *J. Virol.*, 66(6):3557–3565, Jun 1992.

[131] N. F. Abdul-Hamid, M. F?rat-Sarac, A. D. Radford, N. J. Knowles, and D. P. King. Comparative sequence analysis of representative foot-and-mouth disease virus genomes from Southeast Asia. *Virus Genes*, 43(1):41–45, Aug 2011.

[132] E. Domingo and J. J. Holland. RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.*, 51:151–178, 1997.

[133] M. Eigen. Error catastrophe and antiviral strategy. *Proc. Natl. Acad. Sci. U.S.A.*, 99(21):13374–13376, Oct 2002.

[134] S. Crotty, C. Cameron, and R. Andino. Ribavirin's antiviral mechanism of action: lethal mutagenesis? *J. Mol. Med.*, 80(2):86–95, Feb 2002.

[135] S. Crotty, C. E. Cameron, and R. Andino. RNA virus error catastrophe: direct molecular test by using ribavirin. *Proc. Natl. Acad. Sci. U.S.A.*, 98(12):6895–6900, Jun 2001.

[136] C. E. Cameron and C. Castro. The mechanism of action of ribavirin: lethal mutagenesis of RNA virus genomes mediated by the viral RNA-dependent RNA polymerase. *Curr. Opin. Infect. Dis.*, 14(6):757–764, Dec 2001.

[137] M. Vignuzzi, J. K. Stone, and R. Andino. Ribavirin and lethal mutagenesis of poliovirus: molecular mechanisms, resistance and biological implications. *Virus Res.*, 107(2):173–181, Feb 2005.

[138] L. Heath, E. van der Walt, A. Varsani, and D. P. Martin. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J. Virol.*, 80(23): 11827–11832, Dec 2006.

[139] A. L. Jackson, H. O'Neill, F. Maree, B. Blignaut, C. Carrillo, L. Rodriguez, and D. T. Haydon. Mosaic structure of foot-and-mouth disease virus genomes. *J. Gen. Virol.*, 88(Pt 2):487–492, Feb 2007.

[140] T. Rhodes, H. Wargo, and W. S. Hu. High rates of human immunodeficiency virus type 1 recombination: near-random segregation of markers one kilobase apart in one round of viral replication. *J. Virol.*, 77(20):11193–11200, Oct 2003.

[141] A. R. Wargo and G. Kurath. Viral fitness: definitions, measurement, and current insights. *Curr Opin Virol*, 2(5):538–545, Oct 2012.

[142] E. C. Holmes. Error thresholds and the constraints to RNA virus evolution. *Trends Microbiol.*, 11(12):543–546, Dec 2003.

[143] G. Thebaud, J. Chadoeuf, M. J. Morelli, J. W. McCauley, and D. T. Haydon. The relationship between mutation frequency and replication strategy in positive-sense single-stranded RNA viruses. *Proc. Biol. Sci.*, 277(1682):809–817, Mar 2010.

[144] A. S. Lauring and R. Andino. Quasispecies theory and the behavior of RNA viruses. *PLoS Pathog.*, 6(7):e1001005, 2010.

[145] C. O. Wilke. Quasispecies theory in the context of population genetics. *BMC Evol. Biol.*, 5:44, 2005.

[146] H. A. Orr. Fitness and its role in evolutionary genetics. *Nat. Rev. Genet.*, 10(8): 531–539, Aug 2009.

[147] James S.F. Barker. *Adaptation and Fitness in Animal Populations: Evolutionary and Breeding Perspectives on Genetic Resource Management: Defining Fitness in Natural and Domesticated Populations*. Springer Netherlands, Dordrech, 2009. ISBN 9781402090059.

[148] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, Jul 2001.

[149] F. M. Codoner, J. A. Daros, R. V. Sole, and S. F. Elena. The fittest versus the flattest: experimental confirmation of the quasispecies effect with subviral pathogens. *PLoS Pathog.*, 2(12):e136, Dec 2006.

[150] Manfred Eigen, John McCaskill, and Peter Schuster. Molecular quasi-species. *The Journal of Physical Chemistry*, 92(24):6881–6891, 1988.

[151] H. Li and M. J. Roossinck. Genetic bottlenecks reduce population variation in an experimental RNA virus population. *J. Virol.*, 78(19):10582–10587, Oct 2004.

[152] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351, May 2016.

[153] illumina. Sequencing by synthesis (sbs) technology, 2013. URL http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.html.

[154] C. B. Jabara, C. D. Jones, J. Roach, J. A. Anderson, and R. Swanstrom. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl. Acad. Sci. U.S.A.*, 108(50):20166–20171, Dec 2011.

[155] N. Beerenwinkel, H. F. Gunthard, V. Roth, and K. J. Metzner. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol*, 3:329, 2012.

[156] A. D. Radford, D. Chapman, L. Dixon, J. Chantrey, A. C. Darby, and N. Hall. Application of next-generation sequencing technologies in virology. *J. Gen. Virol.*, 93(Pt 9):1853–1868, Sep 2012.

[157] A. Acevedo, L. Brodsky, and R. Andino. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature*, 505(7485):686–690, Jan 2014.

[158] D. I. Lou, J. A. Hussmann, R. M. McBee, A. Acevedo, R. Andino, W. H. Press, and S. L. Sawyer. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 110(49): 19872–19877, Dec 2013.

[159] A. Varble, R. A. Albrecht, S. Backes, M. Crumiller, N. M. Bouvier, D. Sachs, A. Garcia-Sastre, and B. R. tenOever. Influenza A virus transmission bottlenecks are defined by infection route and recipient host. *Cell Host Microbe*, 16(5):691–700, Nov 2014.

[160] S. Zhou, C. Jones, P. Mieczkowski, and R. Swanstrom. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next-Generation Sequencing of HIV-1 Genomic RNA Populations. *J. Virol.*, 89(16):8540–8555, Aug 2015.

[161] J. R. Keys, S. Zhou, J. A. Anderson, J. J. Eron, L. A. Rackoff, C. Jabara, and R. Swanstrom. Primer ID Informs Next-Generation Sequencing Platforms and Reveals Preexisting Drug Resistance Mutations in the HIV-1 Reverse Transcriptase Coding Domain. *AIDS Res. Hum. Retroviruses*, 31(6):658–668, Jun 2015.

[162] M. Hamady, J. J. Walker, J. K. Harris, N. J. Gold, and R. Knight. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat. Methods*, 5(3):235–237, Mar 2008.

[163] C. T. Deakin, J. J. Deakin, S. L. Ginn, P. Young, D. Humphreys, C. M. Suter, I. E. Alexander, and C. V. Hallwirth. Impact of next-generation sequencing error on analysis of barcoded plasmid libraries of known complexity and sequence. *Nucleic Acids Res.*, 42(16):e129, 2014.

[164] K. Mir, K. Neuhaus, M. Bossert, and S. Schober. Short barcodes for next generation sequencing. *PLoS ONE*, 8(12):e82933, 2013.

[165] L. Mamanova, A. J. Coffey, C.E. Scott, I. Kozarewa, E.H. Turner, A. Kumar, E. Howard, J. Shendure, and D.J. Turner. Target-enrichment strategies for next-generation sequencing. *Nature Methods*, 7(2):111–118, Feb 2010.

[166] R. J. Orton, C. F. Wright, M. J. Morelli, D. J. King, D. J. Paton, D. P. King, and D. T. Haydon. Distinguishing low frequency mutations from rt-pcr and sequence errors in viral deep sequencing data. *BMC Genomics*, 16(1):1–15, March 2015.

[167] E. M. Cottam, J. Wadsworth, A. E. Shaw, R. J. Rowlands, L. Goatley, S. Maan, N. S. Maan, P. P. Mertens, K. Ebert, Y. Li, E. D. Ryan, N. Juleff, N. P. Ferris, J. W. Wilesmith, D. T. Haydon, D. P. King, D. J. Paton, and N. J. Knowles. Transmission pathways of foot-and-mouth disease virus in the united kingdom in 2007. *PLoS Pathog*, 4, 2008.

[168] C. F. Wright, M. J. Morelli, G. Thebaud, N. J. Knowles, P. Herzyk, D. J. Paton, D. T. Haydon, and D. P. King. Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J Virol*, 85, 2011.

[169] C. F. Wright, N. J. Knowles, A. Di Nardo, D. J. Paton, D. T. Haydon, and D. P. King. Reconstructing the origin and transmission dynamics of the 1967-68 foot-and-mouth disease epidemic in the United Kingdom. *Infect. Genet. Evol.*, 20: 230–238, Dec 2013.

[170] G. Logan, G. L. Freimanis, D. J. King, B. Valdazo-Gonzalez, K. Bachanek-Bankowska, N. D. Sanderson, N. J. Knowles, D. P. King, and E. M. Cottam. A universal protocol to generate consensus level genome sequences for foot-and-mouth disease virus and other positive-sense polyadenylated RNA viruses using the Illumina MiSeq. *BMC Genomics*, 15:828, 2014.

[171] J. Arzt, N. Juleff, Z. Zhang, and L. L. Rodriguez. The pathogenesis of foot-and-mouth disease i: viral pathways in cattle. *Transbound Emerg Dis*, 58, 2011.

[172] A. Di Nardo, N. J. Knowles, and D. J. Paton. Combining livestock trade patterns with phylogenetics to help understand the spread of foot and mouth disease in sub-saharan africa, the middle east and southeast asia. *Rev Sci Tech*, 30, 2011.

[173] N. F. Abdul-Hamid, M. Firat-Sarac, A. D. Radford, N. J. Knowles, and D. P. King. Comparative sequence analysis of representative foot-and-mouth disease virus genomes from southeast asia. *Virus Genes*, 43, 2011.

[174] *OIE Manual of Diagnostic Tests and Vaccines for Terrestrial Animals 2013*. 2012.

[175] A. Escobar-Gutierrez, M. Vazquez-Pichardo, M. Cruz-Rivera, P. Rivera-Osorio, J. C. Carpio-Pedroza, J. A. Ruiz-Pacheco, K. Ruiz-Tovar, and G. Vaughan. Identification of hepatitis c virus transmission using a next-generation sequencing approach. *J Clin Microbiol*, 50, 2012.

[176] T. H. Wong, B. L. Dearlove, J. Hedge, A. P. Giess, P. Piazza, A. Trebes, J. Paul, E. Smit, E. G. Smith, J. K. Sutton, M. H. Wilcox, K. E. Dingle, T. E. Peto, D. W. Crook, D. J. Wilson, and D. H. Wyllie. Whole genome sequencing and de novo assembly identifies sydney-like variant noroviruses and recombinants during the winter 2012/2013 outbreak in england. *Virol J*, 10, 2013.

[177] L. Barzon, V. Militello, E. Lavezzo, E. Franchin, E. Peta, L. Squarzon, M. Trevisan, S. Pagni, F. Dal Bello, S. Toppo, and G. Palù. Human papillomavirus genotyping by 454 next generation sequencing technology. *J Clin Virol*, 52, 2011.

[178] A. Topfer, D. Hoper, S. Blome, M. Beer, N. Beerenwinkel, N. Ruggli, and I. Leifer. Sequencing approach to analyze the role of quasispecies for classical swine fever. *Virology*, 438, 2013.

[179] D. N. Black, P. Stephenson, D. J. Rowlands, and F. Brown. Sequence and location of the poly c tract in aphtho- and cardiovirus rna. *Nucleic Acids Res*, 6, 1979.

[180] L. Kasambula, G. J. Belsham, H. R. Siegismund, V. B. Muwanika, A. R. Ademun-Okurut, and C. Masembe. Serotype identification and vp1 coding sequence analysis of foot-and-mouth disease viruses from outbreaks in eastern and northern uganda in 2008/9. *Transbound Emerg Dis*, 59, 2012.

[181] A. R. Samuel and N. J. Knowles. Foot-and-mouth disease type o viruses exhibit genetically and geographically distinct evolutionary lineages (topotypes). *J Gen Virol*, 82, 2001.

[182] B. Valdazo-González, L. Polihronova, T. Alexandrov, P. Normann, N. J. Knowles, J. M. Hammond, G. K. Georgiev, F. Özyörük, K. J. Sumption, G. J. Belsham, and D. P. King. Reconstruction of the transmission history of rna virus outbreaks using full genome sequences: foot-and-mouth disease virus in bulgaria in 2011. *PLoS One*, 7, 2012.

[183] T. A. Leski, B. Lin, A. P. Malanoski, and D. A. Stenger. Application of resequencing microarrays in microbial detection and characterization. *Future Microbiol*, 7, 2012.

[184] B. Mullan, E. Kenny-Walsh, J. K. Collins, F. Shanahan, and L. J. Fanning. Inferred hepatitis c virus quasispecies diversity is influenced by choice of dna polymerase in reverse transcriptase-polymerase chain reactions. *Anal Biochem*, 289, 2001.

[185] B. Mullan, P. Sheehy, F. Shanahan, and L. Fanning. Do taq-generated rt-pcr products from rna viruses accurately reflect viral genetic heterogeneity? *J Viral Hepat*, 11, 2004.

[186] E. L. van Dijk, Y. Jaszczyszyn, and C. Thermes. Library preparation methods for next-generation sequencing: Tone down the bias. *Exp Cell Res*, 322, 2014.

[187] G. M. Daly, N. Bexfield, J. Heaney, S. Stubbs, A. P. Mayer, A. Palser, P. Kellam, N. Drou, M. Caccamo, L. Tiley, G. J. Alexander, W. Bernal, and J. L. Heeney. A viral discovery methodology for clinical biopsy samples utilising massively parallel next generation sequencing. *PLoS One*, 6, 2011.

[188] D. W. Eyre, T. Golubchik, N. C. Gordon, R. Bowden, P. Piazza, E. M. Batty, C. L. Ip, D. J. Wilson, X. Didelot, L. O'Connor, R. Lay, D. Buck, A. M. Kearns, A. Shaw, J. Paul, M. H. Wilcox, P. J. Donnelly, T. E. Peto, A. S. Walker, and D. W. Crook. A pilot study of rapid benchtop sequencing of staphylococcus aureus and clostridium difficile for outbreak detection and surveillance. *BMJ Open*, 2, 2012.

[189] T. M. Walker, C. L. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. J. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, A. S. Walker, R. Bowden, P. Monk, E. G. Smith, and T. E. Peto. Whole-genome sequencing to delineate mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis*, 13, 2013.

[190] D. J. Wilson. Insights from genomics into bacterial pathogen populations. *PLoS Pathog*, 8, 2012.

[191] S. M. Reid, S. Parida, D. P. King, G. H. Hutchings, A. E. Shaw, N. P. Ferris, Z. Zhang, J. E. Hillerton, and D. J. Paton. Utility of automated real-time rt-pcr for the detection of foot-and-mouth disease virus excreted in milk. *Vet Res*, 37, 2006.

[192] J. D. Callahan, F. Brown, F. A. Osorio, J. H. Sur, E. Kramer, G. W. Long, J. Lubroth, S. J. Ellis, K. S. Shoulars, K. L. Gaffney, D. L. Rock, and W. M. Nelson. Use of a portable real-time reverse transcriptase-polymerase chain reaction assay for rapid detection of foot-and-mouth disease virus. *J Am Vet Med Assoc*, 220, 2002.

[193] N. J. Knowles and A. R. Samuel. *Polymerase chain reaction amplification and cycle sequencing of the 1D (VP1) gene of foot and mouth disease viruses*. Vienna, Austria, 1995.

[194] N. J. Knowles, N. D. Dickinson, G. Wilsden, E. Carra, E. Brocchi, and F. De Simone. Molecular analysis of encephalomyocarditis viruses isolated from pigs and rodents in italy. *Virus Res*, 57, 1998.

[195] U. Wernery, N. J. Knowles, C. Hamblin, R. Wernery, S. Joseph, J. Kinne, and P. Nagy. Abortions in dromedaries (camelus dromedarius) caused by equine rhinitis a virus. *J Gen Virol*, 89, 2008.

[196] Fastqc: A quality control tool for high throughput sequence data. [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/], . URL http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[197] Sickle: A sliding-window, adaptive, quality-based trimming tool for fastq files. [https://github.com/najoshi/sickle], . URL https://github.com/najoshi/sickle.

[198] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18, 2008.

[199] Basic local alignment search tool (blast). [http://blast.ncbi.nlm.nih.gov/blast.cgi], . URL http://blast.ncbi.nlm.nih.gov/Blast.cgi.

[200] T. A. Hall. Bioedit: a user-friendly biological sequence editor and analysis program for windows 95/98/nt. *Nucl Acids Symp Ser*, 41, 1999.

[201] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10, 2009.

[202] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25, 2009.

[203] I. Milne, G. Stephen, M. Bayer, P. J. Cock, L. Pritchard, L. Cardle, P. D. Shaw, and D. Marshall. Using tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*, 14, 2013.

[204] A. R. Quinlan and I. M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 2010.

[205] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nat Methods*, 9, 2012.

[206] C. Kilkenny, W. J. Browne, I. C. Cuthill, M. Emerson, and D. G. Altman. Improving bioscience research reporting: The arrive guidelines for reporting animal research. *J Pharmacol Pharmacotherapeutics*, 1, 2010.

[207] D. A. Marston, L. M. McElhinney, R. J. Ellis, D. L. Horton, E. L. Wise, S. L. Leech, D. David, X. de Lamballerie, and A. R. Fooks. Next generation sequencing of viral rna genomes. *BMC Genomics*, 14, 2013.

[208] E. M. Batty, T. H. Wong, A. Trebes, K. Argoud, M. Attar, D. Buck, C. L. Ip, T. Golubchik, M. Cule, R. Bowden, C. Manganis, P. Klenerman, E. Barnes, A. S. Walker, D. H. Wyllie, D. J. Wilson, K. E. Dingle, T. E. Peto, D. W. Crook, and P. Piazza. A modified rna-seq approach for whole genome sequencing of rna viruses from faecal and blood samples. *PLoS One*, 8, 2013.

[209] M. A. Quail, T. D. Otto, Y. Gu, S. R. Harris, T. F. Skelly, J. A. McQuillan, H. P. Swerdlow, and S. O. Oyola. Optimal enzymes for amplifying sequencing libraries. *Nat Methods*, 9, 2012.

[210] B. Valdazo-Gonzalez, N. J. Knowles, J. Hammond, and D. P. King. Genome sequences of sat 2 foot-and-mouth disease viruses from egypt and palestinian autonomous territories (gaza strip). *J Virol*, 86, 2012.

[211] B. Valdazo-Gonzalez, A. Timina, A. Scherbakov, N. F. Abdul-Hamid, N. J. Knowles, and D. P. King. Multiple introductions of serotype o foot-and-mouth disease viruses into east asia in 2010–2011. *Vet Res*, 44, 2013.

[212] B. Valdazo-Gonzalez, N. J. Knowles, and D. P. King. Genome sequences of foot-and-mouth disease virus o/me-sa/ind-2001 lineage from outbreaks in libya, saudi arabia, and bhutan during 2013. *Genome Announc*, 2, 2014.

[213] X. Xiang, D. Qiu, R. D. Hegele, and W. C. Tan. Comparison of different methods of total RNA extraction for viral detection in sputum. *J. Virol. Methods*, 94(1-2): 129–135, May 2001.

[214] L.Z. Santiago-Vazques, L.K. Ranzer, and R.G. Kerr. Comparison of two total RNA extraction protocols using the marine gorgonian coral Pseudopterogorgia elisabethae and its symbiont Symbiodinium sp. Technical report, Florida Atlantic University, Department of Chemistry and Biochemistry, May 2006.

[215] M. Y. Deng, H. Wang, G. B. Ward, T. R. Beckham, and T. S. McKenna. Comparison of six RNA extraction methods for the detection of classical swine fever virus

by real-time and conventional reverse transcription-PCR. *J. Vet. Diagn. Invest.*, 17(6):574–578, Nov 2005.

[216] ThermoScientific. Assessment of Nucleic Acid Purity. Technical report, ThermoScientific.

[217] Illumina. NexteraXT Library Prep: Tips and Troubleshooting. Technical report, Illumina, March 2015.

[218] R. Orton and J. H. Hughes. deversiTools, 2015. URL http://josephhughes.github.io/btctools/. Accessed: 2016-05-26.

[219] J. Yang, E. N. Leen, F. F. Maree, and S. Curry. Crystal structure of the 3C protease from Southern African Territories type 2 foot-and-mouth disease virus. *PeerJ*, 4:e1964, 2016.

[220] J. R. Birtley, S. R. Knox, A. M. Jaulent, P. Brick, R. J. Leatherbarrow, and S. Curry. Crystal structure of foot-and-mouth disease virus 3C protease. New insights into catalytic mechanism and cleavage specificity. *J. Biol. Chem.*, 280 (12):11520–11527, Mar 2005.

[221] C. Carrillo, E. R. Tulman, G. Delhon, Z. Lu, A. Carreno, A. Vagnozzi, G. F. Kutish, and D. L. Rock. Comparative genomics of foot-and-mouth disease virus. *J. Virol.*, 79(10):6487–6504, May 2005.

[222] G. R. Thomson. Overview of foot and mouth disease in southern Africa. *Rev. - Off. Int. Epizoot.*, 14(3):503–520, Sep 1995.

[223] F. F. Maree, B. Blignaut, L. Aschenbrenner, T. Burrage, and E. Rieder. Analysis of SAT1 type foot-and-mouth disease virus capsid proteins: influence of receptor usage on the properties of virus particles. *Virus Res.*, 155(2):462–472, Feb 2011.

[224] W. Vosloo, E. Kirkbride, R. G. Bengis, D. F. Keet, and G. R. Thomson. Genome variation in the SAT types of foot-and-mouth disease viruses prevalent in buffalo (Syncerus caffer) in the Kruger National Park and other regions of southern Africa, 1986-93. *Epidemiol. Infect.*, 114(1):203–218, Feb 1995.

[225] A.C. Palmenberg. Sequence alignments of picornaviral capsid proteins. In B.L. Semler and E. Ehrenfieldd, editors, *Molecular aspects of picornavirus infection and detection*, chapter 13, pages 211–241. American Society of Microbiology, Washington D.C., 1989.

[226] R. S. Hedger, J. B. Condy, and S. M. Golding. Infection of some species of African wild life with foot-and-mouth disease virus. *J. Comp. Pathol.*, 82(4):455–461, Oct 1972.

[227] E. C. Anderson, J. Anderson, W. J. Doughty, and S. Drevmo. The pathogenicity of bovine strains of foot and mouth disease virus for impala and wildebeest. *J. Wildl. Dis.*, 11(2):248–255, Apr 1975.

[228] E. C. Anderson, C. Foggin, M. Atkinson, K. J. Sorensen, R. L. Madekurozva, and J. Nqindi. The role of wild animals, other than buffalo, in the current epidemiology of foot-and-mouth disease in Zimbabwe. *Epidemiol. Infect.*, 111(3):559–563, Dec 1993.

[229] W. Vosloo, A. D. Bastos, E. Kirkbride, J. J. Esterhuysen, D. J. van Rensburg, R. G. Bengis, D. W. Keet, and G. R. Thomson. Persistent infection of African buffalo (Syncerus caffer) with SAT-type foot-and-mouth disease viruses: rate of fixation of mutations, antigenic change and interspecies transmission. *J. Gen. Virol.*, 77 (Pt 7):1457–1467, Jul 1996.

[230] B. S. Phologane, R. M. Dwarka, D. T. Haydon, L. J. Gerber, and W. Vosloo. Molecular characterization of SAT-2 foot-and-mouth disease virus isolates obtained from cattle during a four-month period in 2001 in Limpopo Province, South Africa. *Onderstepoort J. Vet. Res.*, 75(4):267–277, Dec 2008.

[231] F. Maree, L. M. de Klerk-Lorist, S. Gubbins, F. Zhang, J. Seago, E. Perez-Martin, L. Reid, K. Scott, L. van Schalkwyk, R. Bengis, B. Charleston, and N. Juleff. Differential Persistence of Foot-and-Mouth Disease Virus in African Buffalo Is Related to Virus Virulence. *J. Virol.*, 90(10):5132–5140, May 2016.

[232] G. R. Thomson, W. Vosloo, J. J. Esterhuysen, and R. G. Bengis. Maintenance of foot and mouth disease viruses in buffalo (Syncerus caffer Sparrman, 1779) in southern Africa. *Rev. - Off. Int. Epizoot.*, 11(4):1097–1107, Dec 1992.

[233] C. Ayebazibwe, F. N. Mwiine, K. Tj?rneh?j, S. N. Balinda, V. B. Muwanika, A. R. Ademun Okurut, G. J. Belsham, P. Normann, H. R. Siegismund, and S. Alexandersen. The role of African buffalos (Syncerus caffer) in the maintenance of foot-and-mouth disease in Uganda. *BMC Vet. Res.*, 6:54, 2010.

[234] J. B. Condy, R. S. Hedger, C. Hamblin, and I. T. Barnett. The duration of the foot-and-mouth disease virus carrier state in African buffalo (i) in the individual animal and (ii) in a free-living herd. *Comp. Immunol. Microbiol. Infect. Dis.*, 8 (3-4):259–265, 1985.

[235] W. Vosloo, P. N. Thompson, B. Botha, R. G. Bengis, and G. R. Thomson. Longitudinal study to investigate the role of impala (Aepyceros melampus) in foot-and-mouth disease maintenance in the Kruger National Park, South Africa. *Transbound Emerg Dis*, 56(1-2):18–30, Mar 2009.

[236] A. D. Bastos, C. I. Boshoff, D. F. Keet, R. G. Bengis, and G. R. Thomson. Natural transmission of foot-and-mouth disease virus between African buffalo (Syncerus caffer) and impala (Aepyceros melampus) in the Kruger National Park, South Africa. *Epidemiol. Infect.*, 124(3):591–598, Jun 2000.

[237] S. K. Hargreaves, C. M. Foggin, E. C. Anderson, A. D. Bastos, G. R. Thomson, N. P. Ferris, and N. J. Knowles. An investigation into the source and spread of foot and mouth disease virus from a wildlife conservancy in Zimbabwe. *Rev. - Off. Int. Epizoot.*, 23(3):783–790, Dec 2004.

[238] A. Plauzolles, M. Lucas, and S. Gaudieri. Influence of host resistance on viral adaptation: hepatitis C virus as a case study. *Infect Drug Resist*, 8:63–74, 2015.

[239] G. E. Price, R. Ou, H. Jiang, L. Huang, and D. Moskophidis. Viral escape by selection of cytotoxic T cell-resistant variants in influenza A virus pneumonia. *J. Exp. Med.*, 191(11):1853–1867, Jun 2000.

[240] O. Zagordi, A. Bhattacharya, N. Eriksson, and N. Beerenwinkel. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics*, 12:119, 2011.

[241] M. C. Prosperi and M. Salemi. QuRe: software for viral quasispecies reconstruction from next-generation sequencing data. *Bioinformatics*, 28(1):132–133, Jan 2012.

[242] S. Prabhakaran, M. Rey, O. Zagordi, N. Beerenwinkel, and V. Roth. HIV Haplotype Inference Using a Propagating Dirichlet Process Mixture Model. *IEEE/ACM Trans Comput Biol Bioinform*, 11(1):182–191, 2014.

[243] M. C. Prosperi, L. Yin, D. J. Nolan, A. D. Lowe, M. M. Goodenow, and M. Salemi. Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. *Sci Rep*, 3:2837, 2013.

[244] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10(3):512–526, May 1993.

[245] K. Tamura, G. Stecher, D. Peterson, A. Filipski, and S. Kumar. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.*, 30(12):2725–2729, Dec 2013.

[246] C. C. Kok and P. C. McMinn. Picornavirus RNA-dependent RNA polymerase. *Int. J. Biochem. Cell Biol.*, 41(3):498–502, Mar 2009.

[247] A. Berinstein, M. Roivainen, T. Hovi, P. W. Mason, and B. Baxt. Antibodies to the vitronectin receptor (integrin alpha V beta 3) inhibit binding and infection of foot-and-mouth disease virus to cultured cells. *J. Virol.*, 69(4):2664–2666, Apr 1995.

[248] S. Berryman, S. Clark, P. Monaghan, and T. Jackson. Early events in integrin alphavbeta6-mediated cell entry of foot-and-mouth disease virus. *J. Virol.*, 79(13): 8519–8534, Jul 2005.

[249] E. E. Fry, S. M. Lea, T. Jackson, J. W. Newman, F. M. Ellard, W. E. Blakemore, R. Abu-Ghazaleh, A. Samuel, A. M. King, and D. I. Stuart. The structure and function of a foot-and-mouth disease virus-oligosaccharide receptor complex. *EMBO J.*, 18(3):543–554, Feb 1999.

[250] C. Sun, D. Yang, R. Gao, T. Liang, H. Wang, G. Zhou, and L. Yu. Modification of the internal ribosome entry site element impairs the growth of foot-and-mouth disease virus in porcine-derived cells. *J. Gen. Virol.*, 97(4):901–911, Apr 2016.

[251] A. M. Borman, P. Le Mercier, M. Girard, and K. M. Kean. Comparison of picornaviral IRES-driven internal initiation of translation in cultured cells of different origins. *Nucleic Acids Res.*, 25(5):925–932, Mar 1997.

[252] A. Yanagiya, S. Ohka, N. Hashida, M. Okamura, C. Taya, N. Kamoshita, K. Iwasaki, Y. Sasaki, H. Yonekawa, and A. Nomoto. Tissue-specific replicating capacity of a chimeric poliovirus that carries the internal ribosome entry site of hepatitis C virus in a new mouse model transgenic for the human poliovirus receptor. *J. Virol.*, 77(19):10479–10487, Oct 2003.

[253] M. Gromeier, L. Alexander, and E. Wimmer. Internal ribosomal entry site substitution eliminates neurovirulence in intergeneric poliovirus recombinants. *Proc. Natl. Acad. Sci. U.S.A.*, 93(6):2370–2375, Mar 1996.

[254] S. Ohka and A. Nomoto. The molecular basis of poliovirus neurovirulence. *Dev Biol (Basel)*, 105:51–58, 2001.

[255] C. Wang, P. Jiang, C. Sand, A. V. Paul, and E. Wimmer. Alanine scanning of poliovirus 2CATPase reveals new genetic evidence that capsid protein/2CATPase interactions are essential for morphogenesis. *J. Virol.*, 86(18):9964–9975, Sep 2012.

[256] S. Lopez de Quinto and E. Martinez-Salas. Interaction of the eIF4G initiation factor with the aphthovirus IRES is essential for internal translation initiation in vivo. *RNA*, 6(10):1380–1392, Oct 2000.

[257] V. G. Kolupaeva, T. V. Pestova, C. U. Hellen, and I. N. Shatsky. Translation eukaryotic initiation factor 4G recognizes a specific structural element within the internal ribosome entry site of encephalomyocarditis virus RNA. *J. Biol. Chem.*, 273(29):18599–18604, Jul 1998.

[258] I. B. Lomakin, C. U. Hellen, and T. V. Pestova. Physical association of eukaryotic initiation factor 4G (eIF4G) with eIF4A strongly enhances binding of eIF4G to the internal ribosomal entry site of encephalomyocarditis virus and is required for internal initiation of translation. *Mol. Cell. Biol.*, 20(16):6019–6029, Aug 2000.

[259] K. Deres, C. H. Schroder, A. Paessens, S. Goldmann, H. J. Hacker, O. Weber, T. Kramer, U. Niewohner, U. Pleiss, J. Stoltefuss, E. Graef, D. Koletzki, R. N. Masantschek, A. Reimann, R. Jaeger, R. Gross, B. Beckermann, K. H. Schlemmer, D. Haebich, and H. Rubsamen-Waigmann. Inhibition of hepatitis B virus replication by drug-induced depletion of nucleocapsids. *Science*, 299(5608):893–896, Feb 2003.

[260] J. D. Baines. Herpes simplex virus capsid assembly and DNA packaging: a present and future antiviral drug target. *Trends Microbiol.*, 19(12):606–613, Dec 2011.

[261] J. Dietz, J. Koch, A. Kaur, C. Raja, S. Stein, M. Grez, A. Pustowka, S. Mensch, J. Ferner, L. Moller, N. Bannert, R. Tampe, G. Divita, Y. Mely, H. Schwalbe, and U. Dietrich. Inhibition of HIV-1 by a peptide ligand of the genomic RNA packaging signal Psi. *ChemMedChem*, 3(5):749–755, May 2008.

[262] K. Watanabe, F. Momose, H. Handa, and K. Nagata. Interaction between influenza virus proteins and pine cone antitumor substance that inhibits the virus multiplication. *Biochem. Biophys. Res. Commun.*, 214(2):318–323, Sep 1995.

[263] M. Feiss and V. B. Rao. The bacteriophage DNA packaging machine. *Adv. Exp. Med. Biol.*, 726:489–509, 2012.

[264] A. Ben-Shaul and W. M. Gelbart. Viral ssRNAs are indeed compact. *Biophys. J.*, 108(1):14–16, Jan 2015.

[265] A. Borodavka, R. Tuma, and P. G. Stockley. Evidence that viral RNAs have evolved for efficient, two-stage packaging. *Proc. Natl. Acad. Sci. U.S.A.*, 109(39): 15769–15774, Sep 2012.

[266] S. Shazman and Y. Mandel-Gutfreund. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.*, 4(8):e1000146, 2008.

[267] D. E. Draper. Themes in RNA-protein recognition. *J. Mol. Biol.*, 293(2):255–270, Oct 1999.

[268] E. L. Hesketh, Y. Meshcheriakova, K. C. Dent, P. Saxena, R. F. Thompson, J. J. Cockburn, G. P. Lomonossoff, and N. A. Ranson. Mechanisms of assembly and genome packaging in an RNA virus revealed by high-resolution cryo-EM. *Nat Commun*, 6:10113, Dec 2015.

[269] R. F. Garmann, M. Comas-Garcia, M. S. Koay, J. J. Cornelissen, C. M. Knobler, and W. M. Gelbart. Role of electrostatics in the assembly pathway of a single-stranded RNA virus. *J. Virol.*, 88(18):10472–10479, Sep 2014.

[270] N. Patel, E. C. Dykeman, R. H. Coutts, G. P. Lomonossoff, D. J. Rowlands, S. E. Phillips, N. Ranson, R. Twarock, R. Tuma, and P. G. Stockley. Revealing the density of encoded functions in a viral RNA. *Proc. Natl. Acad. Sci. U.S.A.*, 112 (7):2227–2232, Feb 2015.

[271] P. G. Stockley, R. Twarock, S. E. Bakker, A. M. Barker, A. Borodavka, E. Dykeman, R. J. Ford, A. R. Pearson, S. E. V. Phillips, N. A. Ranson, and R. Tuma. Packaging signals in single-stranded RNA viruses: nature's alternative to a purely electrostatic assembly mechanism. *J. Biol. Phys.*, 39(2):277–287, 2013.

[272] P. G. Stockley, O. Rolfsson, G. S. Thompson, G. Basnak, S. Francese, N. J. Stonehouse, S. W. Homans, and A. E. Ashcroft. A simple, RNA-mediated allosteric switch controls the pathway to formation of a T=3 viral capsid. *J. Mol. Biol.*, 369 (2):541–552, Jun 2007.

[273] P. Plevka, A. Kazaks, T. Voronkova, S. Kotelovica, A. Dishlers, L. Liljas, and K. Tars. The structure of bacteriophage phiCb5 reveals a role of the RNA genome and metal ions in particle stability and assembly. *J. Mol. Biol.*, 391(3):635–647, Aug 2009.

[274] M. Junker-Niepmann, R. Bartenschlager, and H. Schaller. A short cis-acting sequence is required for hepatitis B virus pregenome encapsidation and sufficient for packaging of foreign RNA. *EMBO J.*, 9(10):3389–3396, Oct 1990.

[275] E. C. Dykeman, P. G. Stockley, and R. Twarock. Packaging signals in two single-stranded RNA viruses imply a conserved assembly mechanism and geometry of the packaged genome. *J. Mol. Biol.*, 425(17):3235–3249, Sep 2013.

[276] L. Zhu, X. Wang, J. Ren, C. Porta, H. Wenham, J. O. Ekstrom, A. Panjwani, N. J. Knowles, A. Kotecha, C. A. Siebert, A. M. Lindberg, E. E. Fry, Z. Rao, T. J. Tuthill, and D. I. Stuart. Structure of Ljungan virus provides insight into genome packaging of this picornavirus. *Nat Commun*, 6:8316, 2015.

[277] S. Kalynych, L. Palkova, and P. Plevka. The Structure of Human Parechovirus 1 Reveals an Association of the RNA Genome with the Capsid. *J. Virol.*, 90(3): 1377–1386, 2015.

[278] J. Sasaki, S. Nagashima, and K. Taniguchi. Aichi virus leader protein is involved in viral RNA replication and encapsidation. *J. Virol.*, 77(20):10799–10807, Oct 2003.

[279] J. Sasaki and K. Taniguchi. The 5'-end sequence of the genome of Aichi virus, a picornavirus, contains an element critical for viral RNA encapsidation. *J. Virol.*, 77(6):3542–3548, Mar 2003.

[280] Y. Liu, C. Wang, S. Mueller, A. V. Paul, E. Wimmer, and P. Jiang. Direct interaction between two viral proteins, the nonstructural protein 2C and the capsid protein VP3, is required for enterovirus morphogenesis. *PLoS Pathog.*, 6(8): e1001066, 2010.

[281] C. Wang, H. C. Ma, E. Wimmer, P. Jiang, and A. V. Paul. A C-terminal, cysteine-rich site in poliovirus 2C(ATPase) is required for morphogenesis. *J. Gen. Virol.*, 95(Pt 6):1255–1265, Jun 2014.

[282] Y. R. Lee, P. S. Wang, J. R. Wang, and H. S. Liu. Enterovirus 71-induced autophagy increases viral replication and pathogenesis in a suckling mouse model. *J. Biomed. Sci.*, 21:80, 2014.

[283] Q. Reuer, R. J. Kuhn, and E. Wimmer. Characterization of poliovirus clones containing lethal and nonlethal mutations in the genome-linked protein VPg. *J. Virol.*, 64(6):2967–2975, Jun 1990.

[284] M. A. Adam and A. D. Miller. Identification of a signal in a murine retrovirus that is sufficient for packaging of nonretroviral RNA into virions. *J. Virol.*, 62(10): 3802–3806, Oct 1988.

[285] C Tuerk and L Gold. Systematic evolution of ligands by exponential enrichment:RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968): 505–510, 1990.

[286] M. B. Schulte, J. A. Draghi, J. B. Plotkin, and R. Andino. Experimentally guided models reveal replication principles that shape the mutation distribution of RNA viruses. *Elife*, 4, 2015.

[287] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18(5):900–914, May 2012.

[288] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3(4):e65, Apr 2007.

[289] C. Smith, S. Heyne, A. S. Richter, S. Will, and R. Backofen. Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. *Nucleic Acids Res.*, 38(Web Server issue):W373–377, Jul 2010.

[290] Y. Song, Y. Liu, C. B. Ward, S. Mueller, B. Futcher, S. Skiena, A. V. Paul, and E. Wimmer. Identification of two functionally redundant RNA elements in the coding sequence of poliovirus using computer-generated design. *Proc. Natl. Acad. Sci. U.S.A.*, 109(36):14301–14307, Sep 2012.

[291] J. Pelletier, G. Kaplan, V. R. Racaniello, and N. Sonenberg. Cap-independent translation of poliovirus mRNA is conferred by sequence elements within the 5' noncoding region. *Mol. Cell. Biol.*, 8(3):1103–1112, Mar 1988.

[292] J. Pelletier and N. Sonenberg. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature*, 334(6180): 320–325, Jul 1988.

[293] I. Goodfellow, Y. Chaudhry, A. Richardson, J. Meredith, J. W. Almond, W. Barclay, and D. J. Evans. Identification of a cis-acting replication element within the poliovirus coding region. *J. Virol.*, 74(10):4590–4600, May 2000.

[294] J. Q. Han, H. L. Townsend, B. K. Jha, J. M. Paranjape, R. H. Silverman, and D. J. Barton. A phylogenetically conserved RNA structure in the poliovirus open reading frame inhibits the antiviral endoribonuclease RNase L. *J. Virol.*, 81(11): 5561–5572, Jun 2007.

[295] H. L. Townsend, B. K. Jha, J. Q. Han, N. K. Maluf, R. H. Silverman, and D. J. Barton. A viral RNA competitively inhibits the antiviral endoribonuclease domain of RNase L. *RNA*, 14(6):1026–1036, Jun 2008.

[296] J. B. Rohll, D. H. Moon, D. J. Evans, and J. W. Almond. The 3' untranslated region of picornavirus RNA: features required for efficient genome replication. *J. Virol.*, 69(12):7835–7844, Dec 1995.

[297] S. Todd, J. S. Towner, D. M. Brown, and B. L. Semler. Replication-competent picornaviruses with complete genomic RNA 3' noncoding region deletions. *J. Virol.*, 71(11):8868–8874, Nov 1997.

[298] C. P. Burrill, O. Westesson, M. B. Schulte, V. R. Strings, M. Segal, and R. Andino. Global RNA structure analysis of poliovirus identifies a conserved RNA structure

involved in viral replication and infectivity. *J. Virol.*, 87(21):11670–11683, Nov 2013.

[299] M. M. Falk, F. Sobrino, and E. Beck. VPg gene amplification correlates with infective particle formation in foot-and-mouth disease virus. *J. Virol.*, 66(4):2251–2260, Apr 1992.

[300] S. Forss and H. Schaller. A tandem repeat gene in a picornavirus. *Nucleic Acids Res.*, 10(20):6441–6450, Oct 1982.

[301] D. V. Sangar, D. N. Black, D. J. Rowlands, T. J. Harris, and F. Brown. Location of the initiation site for protein synthesis on foot-and-mouth disease virus RNA by in vitro translation of defined fragments of the RNA. *J. Virol.*, 33(1):59–68, Jan 1980.

[302] A. M. Yoffe, P. Prinsen, A. Gopal, C. M. Knobler, W. M. Gelbart, and A. Ben-Shaul. Predicting the sizes of large RNA molecules. *Proc. Natl. Acad. Sci. U.S.A.*, 105(42):16153–16158, Oct 2008.

[303] L. Tubiana, A. L. Božič, C. Micheletti, and R. Podgornik. Synonymous mutations reduce genome compactness in icosahedral ssRNA viruses. *Biophys. J.*, 108(1): 194–202, Jan 2015.

[304] L. K. Johansen and C. D. Morrow. The RNA encompassing the internal ribosome entry site in the poliovirus 5' nontranslated region enhances the encapsidation of genomic RNA. *Virology*, 273(2):391–399, Aug 2000.

[305] H. H. Lu and E. Wimmer. Poliovirus chimeras replicating under the translational control of genetic elements of hepatitis C virus reveal unusual properties of the internal ribosomal entry site of hepatitis C virus. *Proc. Natl. Acad. Sci. U.S.A.*, 93(4):1412–1417, Feb 1996.

[306] B. L. Semler, V. H. Johnson, and S. Tracy. A chimeric plasmid from cDNA clones of poliovirus and coxsackievirus produces a recombinant virus that is temperature-sensitive. *Proc. Natl. Acad. Sci. U.S.A.*, 83(6):1777–1781, Mar 1986.

[307] M. Saiz, S. Gomez, E. Martinez-Salas, and F. Sobrino. Deletion or substitution of the aphthovirus 3' NCR abrogates infectivity and virus replication. *J. Gen. Virol.*, 82(Pt 1):93–101, Jan 2001.

[308] J. K. Pfeiffer and K. Kirkegaard. Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. *PLoS Pathog.*, 1(2):e11, Oct 2005.

[309] E. Domingo, C. Escarmis, M. A. Martinez, E. Martinez-Salas, and M. G. Mateu. Foot-and-mouth disease virus populations are quasispecies. *Curr. Top. Microbiol. Immunol.*, 176:33–47, 1992.

[310] G. M. Jenkins, M. Worobey, C. H. Woelk, and E. C. Holmes. Evidence for the non-quasispecies evolution of RNA viruses [corrected]. *Mol. Biol. Evol.*, 18(6): 987–994, Jun 2001.

[311] E. Domingo, D. Sabo, T. Taniguchi, and C. Weissmann. Nucleotide sequence heterogeneity of an RNA phage population. *Cell*, 13(4):735–744, Apr 1978.

[312] D. A. Steinhauer, J. C. de la Torre, E. Meier, and J. J. Holland. Extreme heterogeneity in populations of vesicular stomatitis virus. *J. Virol.*, 63(5):2072–2080, May 1989.

[313] T. K. Sikombe, A. S. Mweene, J. Muma, C. Kasanga, Y. Sinkala, F. Banda, M. Mulumba, E. M. Fana, C. Mundia, and M. Simuunza. Serological Survey of Foot-and-Mouth Disease Virus in Buffaloes (Syncerus caffer) in Zambia. *Vet Med Int*, 2015:264528, 2015.