Dissecting the molecular basis of foot-and-mouth disease virus evolution

Caroline Frances Wright

BSc MSc

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Institute of Biodiversity, Animal Health and
Comparative Medicine
University of Glasgow

October 2012

Abstract

Foot-and-mouth disease virus (FMDV) causes the most contagious transboundary disease of animals, affecting both wild and domestic cloven-hoofed animals. Similarly to other RNA viruses, FMDV is highly variable as a result of the inherent low fidelity of the viral RNA-dependent RNA polymerase. The accumulation of this variability and relatedness between FMDV sequences was used to provide evidence for modes of transmission (fomite) as well as a constant clock rate across two FMDV topotypes (~8.70 x 10^{-3} substitutions/site/year), during the 1967 UK FMD epidemic, using full genome consensus sequencing. However, during an epidemic, virus replicates within multiple animals, where it is also replicating and evolving within different tissues and cells. Each scale of evolution, from a single cell to multiple animals across the globe, involves evolutionary processes that shape the viral diversity generated below the level of the consensus. During this PhD project, next-generation sequencing (NGS) was used to dissect the fine scale viral population diversity of FMDV. Collaboration with the Institute of Biodiversity, Animal Health and Comparative Medicine at the University of Glasgow provided the specialist bioinformatic and statistical capabilities required for the analysis of NGS datasets. As part of this collaboration, a new systematic approach was developed to process NGS data and distinguish genuine mutations from artefacts. Additionally, evolutionary models were applied to this data to estimate parameters such as the genome-wide mutation rate of FMDV (upper limit of 7.8 x 10^{-4} per nt). Analysis of the mutation spectra generated from a clonal control study established a mutation frequency threshold of 0.5% above which there can be confidence that 95% of mutations are real in the sense that they are present in the sampled virus population. This threshold, together with an optimized protocol, was used for the more extensive investigation of within and between host viral population dynamics during transmission. Analysis of mutation spectra and site-specific mutations revealed that intra-host bottlenecks are typically more pronounced than inter-host bottlenecks. NGS analysis has distinguished between the population structure of multiple samples taken from a single host, which may provide the means to reconstruct both intra- and inter-host transmission routes in the future. A more sophisticated understanding of viral diversity at its finest scales could hold the key to the better understanding of viral pathogenesis and, therefore development of effective and sustainable disease treatment and control strategies.

Declaration

I hereby declare that the research described within this thesis is my own work, unless otherwise stated, and certify that is has never been submitted for any other degree or professional qualification

Caroline F Wright BSc MSc

The Pirbright Institute,

Ash Road,

Surrey

GU24 0NF

Acknowledgements

My sincerest thanks and appreciation must first go to my supervisors, Don, Dan and David, whose patience has known no bounds. This PhD would have been a very different experience without their support and encouragement. Our collaboration with Glasgow has given me the opportunity to work with some amazing people outside of my day-to-day environment. I couldn’t have asked for better people to work with in terms of the specific challenges of this PhD than Marco and Richard. Working alongside them has not only expanded my horizons in terms of field of research, it has reminded me that there’s always more than one way to skin a cat! However, I could not have survived the last four years without the support I’ve received from countless people at the IAH, now The Pirbright Institute. In the lab there has been Ryan, who was always ready to discuss/argue about all things science, Daryl for his help with virus structure prediction, Julian, Terry, Toby, Claudia, Valerie and Begona who put up with my incessant questions and always provided sound advice and Jemma for running the Sanger sequencer. I would also like to thank John Gloster and Antonello especially for their help and advice during Chapter 2. Very special thanks go to Mandy Swan and Sheila Wilsden who were my saving grace during Chapter 6 (sorry again for frying all your cells, Mandy!) and Nick Knowles for his guidance and suggestions. Both inside and outside work, there have been so many people who have seen me through the rough times; Beck, Daryl, Jo and Robyn, my ex housemates, who I could always rely on to make me laugh and have a beer with, Sparks, Emma, Paul, Neil and everyone else in the work bar. I must also thank my family for their support and encouragement over the years (my dad, David; step-mum, Juliet; Gramps; Aunty Clare; sisters Edwina, Jo, Rosie and Amelia; brothers James, Seb, Tom and Ollie; sisters in-law Kristin, Jane, Haley and Emma; and brother in-law Matt). My step-dad, Phil, didn’t get to see me undertake this PhD but I know he would have been very proud and would have encouraged me all the way. I will save the last thanks for two very special people who have pushed, inspired, motivated and calmed me whenever I needed it most. My boyfriend, Joe and my mum, Wendy, I could not have done this without you.

Table of contents

List of figures

List of tables

Abbreviations

°C	degrees Celsius
A/T/G/C	adenine/thymine/guanine/cytosine
BFS	British field sample
BHK	baby hamster kidney
bp	base pairs
BTY	bovine thyroid
cm	centimetre
cDNA	complementary DNA
CI	confidence interval
CO_2	carbon dioxide
cre	*cis*-acting replication element
DEFRA	Department for environment, farming and rural affairs
dN/dS	non-synonymous to synonymous substitution ratio
DNA	deoxy-ribonucleic acid
dNTP	deoxy-nucleotide triphosphate
DTT	dithiothreitol
E.coli	Escherichia coli
eIF4G	eukaryotic initiation factor
ER	endoplasmic reticulum
FMD	foot-and-mouth disease
FMDV	foot-and-mouth disease virus
HCV	hepatitis C virus
HIV	human immunodeficiency virus
HS	heparan sulphate
IAH	Institute for Animal Health
IRES	internal ribosome entry site
kb	kilobase
M	molar
MCMC	Markov Chain Monte Carlo
$MgSO_4$	magnesium sulphate
ml	millilitre
moi	multiplicity of infection
nt	nucleotide
PCR	polymerase chain reaction
PFU	plaque forming units
pH	potential of hydrogen

Amino acids

Amino Acid	**3-letter code**	**1-letter code**	**Side chain polarity**	**Side chain acidity or basicity**
Alanine	Ala	A	Nonpolar	Neutral
Arginine	Arg	R	Polar	Basic (strongly)
Asparagine	Asn	N	Polar	Neutral
Aspartic acid	Asp	D	Polar	Acidic
Cysteine	Cys	C	Polar	Neutral
Glutamic acid	Glu	E	Polar	Acidic
Glutamine	Gln	Q	Polar	Neutral
Glycine	Gly	G	Nonpolar	Neutral
Histidine	His	H	Polar	Basic (weakly)
Isoleucine	Ile	I	Nonpolar	Neutral
Leucine	Leu	L	Nonpolar	Neutral
Lysine	Lys	K	Polar	Basic
Methionine	Met	M	Nonpolar	Neutral
Phenylalanine	Phe	F	Nonpolar	Neutral
Proline	Pro	P	Nonpolar	Neutral
Serine	Ser	S	Polar	Neutral
Threonine	Thr	T	Polar	Neutral
Tryptophan	Trp	W	Nonpolar	Neutral
Tyrosine	Tyr	Y	Polar	Neutral
Valine	Val	V	Nonpolar	Neutral

Chapter 1

Introduction to foot-and-mouth disease and the evolution of its causative agent

1.1 Summary

This introductory chapter is broken down into four primary sections. The first section puts this research into its wider context by providing an introduction to the disease caused by the virus studied here, foot-and-mouth disease (FMD). The second section focuses in more detail on the virus itself, foot-and-mouth disease virus (FMDV), in terms of virion structure, the genome, replication, and mechanism of cell entry. FMDV evolves at a range of spatial and temporal scales, the dynamics of which, in terms of FMDV and other RNA viruses, are discussed within the third section. The fourth section discusses how virus evolution has been studied to date. Finally, the chapter culminates with the overall objectives of this PhD.

1.2 The disease

1.2.1 Foot-and-mouth disease

Foot-and-mouth disease (FMD) is the most contagious transboundary animal disease, affecting both wild and domestic cloven-hoofed animals, including cattle, pigs, sheep and goats. Significant economic loss results from its high morbidity and export trade restrictions imposed on affected countries. Mortality is typically low in adult animals but can be high in young animals due to acute myocarditis. The aetiological agent of this disease is foot-and-mouth disease virus (FMDV; family *Picornaviridae;* genus *Aphthovirus*). Seven serotypes of FMDV have been identified (A, O, C, Asia 1, and South African territories [SAT] 1-3) each of which include multiple subtypes. FMD serotypes differ in their global distribution. Serotypes A and O have the widest distribution, occurring in Africa, Asia and South America, where FMD is endemic. Types SAT 1, 2 and 3 are normally restricted to Africa only and Asia 1 to Asia. The capacity of the disease to invade free areas is common to all types, for example, FMDV SAT-2 serotype is currently causing outbreaks in the Middle East and Asia 1 periodically moves west and east from central Asia. Infection or vaccination against one serotype does not provide protection against other serotypes.

1.2.2 Clinical signs

The disease is characterised by a number of debilitating clinical signs, including fever, lameness and vesicular lesions of the feet, tongue, snout and teats. While clinical disease is often severe and obvious in pigs and cattle, signs, such as fluid filled vesicles in the mouth, can be especially subtle in sheep and goats, due to variations in lingual epithelium thickness. Clinical disease is indistinguishable between FMD serotypes and also other vesicular diseases in pigs, such as swine vesicular disease and vesicular stomatitis (affects cattle, horses and occasionally pigs). An experimental study of FMD transmission in cattle found that, on average, animals were infectious 0.5 days after the onset of clinical signs (Charleston, Bankowski et al. 2011). Therefore, timing of clinical disease and proximity of susceptible naïve animals plays a critical role in terms of disease transmission and spread.

1.3 The virus

1.3.1 Foot-and-mouth disease virus

A substantial step towards understanding FMD was Loeffler and Frosch's 1897 landmark demonstration that the disease was caused by a filterable agent, or virus. This virus has a single stranded positive sense RNA genome and is immediately infectious within a cell. The genome is approximately 8500 nt in length (Forss, Strebel et al. 1984) and contained within a non-enveloped icosahedral capsid approximately 25 nm in diameter (Bachrach 1968). Although the virus is sensitive to acidic conditions (pH < 6.0), high temperatures (> 50ºC) and UV, it has been shown to survive for a number of months outside a susceptible host, under favourable conditions, for example, on wool (McColl, Westbury et al. 1995) and in bovine faeces and slurry (Parker 1971; Haas, Ahl et al. 1995). A review of FMDV survival in animal excretions and on fomites has been provided by Bartley, Donnelly et al. 2002.

1.3.2 Virion structure

FMDV serotypes are determined serologically, based upon differences in the antigenic structure of the viral capsid, which is composed of 60 copies of each of four structural proteins VP1-4 (also termed 1D, 1B, 1C and 1A). The VP4 structural protein is buried on the inside of the capsid and has a distinct extended conformation (Acharya, Fry et al. 1989). Conversely, structural proteins VP1, VP2 and VP3 fold into eight-stranded β barrels, which are connected by loops that form the outer surface of the viral particle. FMDV capsid structure has been determined for a number of strains, including O_1BFS (discussed in this thesis), by use of X-ray crystallography. Five antigenic sites have been described involving all three surface exposed structural proteins for type O1 FMDV (Barnett, Ouldridge et al. 1989; Kitson, McCahon et al. 1990; Crowther, Farias et al. 1993). Of particular interest regarding the cellular binding of FMDV is a prominent surface loop connecting the βG and βH strands (G-H loop) of VP1, which will be discussed further in the following section.

1.3.3 Viral cell attachment and entry

The surface exposed G-H loop of FMDV VP1 contains a highly conserved arginine-glycine-aspartic acid (RGD) motif, which interacts directly with host cell receptors (integrin), resulting in virus binding to cells. Integrin are type I heterodimeric membrane proteins consisting of α and β subunits. The major integrin receptors for FMDV in susceptible hosts are αvβ1, αvβ3, αvβ6 and αvβ8 (Berinstein, Roivainen et al. 1995; Neff, Mason et al. 2000; Neff and Baxt 2001; Jackson, Mould et al. 2002; Jackson, Clark et al. 2004). However, αvβ6 has been shown to be the predominant epithelial cell surface receptor within areas that are commonly targeted by FMDV, in cattle (Monaghan, Gold et al. 2005; O'Donnell, Pacheco et al. 2009). Additionally, αvβ3 has been found in close association with blood vessels in various tissues. Interestingly, there are a variety of tissues that, while expressing αvβ6, do not support FMDV replication, perhaps revealing the presence of alternative cell-specific or tissue-specific host factors as co-determinants of tropism. A comprehensive review of host and virus determinants of Picornavirus pathogenesis and tropism has been provided by Whitton, Cornell et al. 2005.

Multiple passages of type O FMDV in cell culture result in viral utilization of the glycosaminoglycan, heparan sulphate (HS), as an alternative cellular receptor; a change that is correlated to selection of viruses containing an extra positively charged amino acid (Jackson, Ellard et al. 1996a; Fry, Lea et al. 1999a).

Once attached to cells, FMDV enters via receptor-mediated endocytosis (different *in vitro* and *in vivo*), where un-coating of the viral capsid occurs due to acidification within the early endosomal pathway (Johns, Berryman et al. 2009). Where infection by integrin binding viruses occurs via clathrin-mediated endocytosis, HS binding viruses enter cells via a caveola-mediated mechanism (O'Donnell, Larocco et al. 2008).

1.3.4 The foot-and-mouth disease virus genome

The FMDV genome contains a single open reading frame (ORF; approximately 7000 nucleotides (nt) long), flanked either end by untranslated regions (UTR), as depicted in Figure 1.1. The 5’ UTR of the FMDV genome is relatively long at approximately 1300 nt. The 5’ UTR consists of, from the 5’ end, a 350-380 nt ‘Short’ (S) fragment, a 100 to 420 nt long poly ‘C’ tract and the approximately 700 nt 5’ terminus of the genomic ‘Long’ (L) fragment, containing three or four tandemly repeated pseudoknots, a stem-loop *cis*-acting replication element (*cre*) and a type II internal ribosomal entry site (IRES), as reviewed by Carrillo, Tulman et al. 2005. This end of the genome is covalently linked to a small viral protein VPg (or 3B). Two stem-loops have been predicted within the 3’ UTR of the FMDV genome (Carrillo, Tulman et al. 2005), which is polyadenylated (poly ‘A’ tract) and relatively short at around 90 nt long. RNA is highly structured throughout the genome, as reviewed by Carrillo, Tulman et al. 2005.

The function of the S fragment, which has been predicted to fold into a large hairpin structure (Clarke, Brown et al. 1987; Escarmis, Toja et al. 1992), is still unknown but it is thought to be required for replication of the RNA. The ‘cloverleaf’ structure at the 5’ end of poliovirus RNA is much better characterized and has been shown to be involved in the process of RNA replication (Andino, Rieckhof et al. 1990) and to have a major effect on RNA stability (Murray, Roberts et al. 2001). The role of the poly ‘C’ tract is also unclear, but it is interesting to note the slowing of growth in cell culture after the removal of only four C residues from this region of cloned FMDV cDNA (Rieder, Bunch et al. 1993). The role of the multiple pseudoknots predicted to occur at the 3’ end of the poly ‘C’ tract (Clarke, Brown et al. 1987; Escarmis, Dopazo et al. 1995) is also unknown.

The FMDV *cre* is required for replication and has been shown to be able to function in *trans (Tiley, King et al. 2003)* and will be discussed further in section 1.3.4. Serrano et al (2006) demonstrated that the S fragment and the IRES interact specifically with the 3’ UTR, therefore potentially playing a role in translation and/or replication of the genome (Serrano, Pulido et al. 2006).

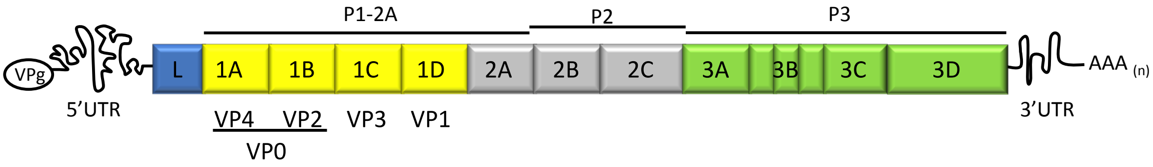


Figure 1.1

A schematic representation of FMDV genome organisation. The FMDV genome is covalently linked to the protein VPg at the 5' end and is polyadenylated at the 3' end. The coding and non-coding regions of FMDV RNA are indicated in the form of a single large open reading frame that encodes a poly-protein flanked either end by untranslated regions (UTR). P1 contains the capsid coding region for viral proteins (VP) 1-4. The remainder of the coding region encodes non-structural proteins. Precursors P1-2A, P2 and P3, including three distinct copies of 3B (VPg), are indicated. Figure adapted from (Belsham 2005).

1.3.5 Foot-and-mouth disease virus translation and replication

The initial role of FMDV RNA upon infection of cells and release into the cytoplasm is as mRNA for the production of viral proteins, which are, in turn, required for RNA replication and packaging of the de novo-synthesised RNA into virions. Different functions of the RNA may occur in discrete compartments within cells. Cap independent translation of the polyprotein is driven by the IRES while a cellular enzyme cleaves the VPg protein. Various structural and non-structural proteins that assist in the replication of the viral genome result from subsequent cleavage of the poly-protein. However, the 5' and 3' UTRs also have significant roles in virus translation and replication (Mason, Grubman et al. 2003). The absence of a cap structure and the presence of an extensive secondary structure with multiple un-used AUG codons are features shared by all picornavirus RNA 5' UTRs.

1.3.5a The proteins

Viral encoded proteases process the viral polyprotein. The FMDV leader (L) protein is also a protease, which cleaves itself and the P1-2A precursor at its N-terminus, which is also cleaved at its C-terminus by the 2A protein. Not all 'cleavage' mechanisms follow a proteolytic reaction, as has been demonstrated during the processing of FMDV 2A/2B by 2A (Donnelly, Luke et al. 2001).

Subsequent processing of the P1-2A precursor by 3C protease yields VP0 (subsequently cleaved into VP4 and VP2), VP3, and VP1 (Figure 1.1). Both P2 and P3 precursors are also cleaved by 3C protease into additional non-structural (NS) proteins, which perform a number of functions that promote virus production (reviewed in Belsham 2005) and block or limit the host response. Replication of the viral genome is driven by the viral RNA-dependent RNA polymerase, 3D, a process that also requires NS protein 3B (VPg) to initiate RNA synthesis and 2C. Additionally, the viral protein 3AB is thought to be required for the initiation of RNA synthesis (reviewed in Grubman and Baxt 2004). The 3D polymerase and 3AB physically associate with each other and with viral RNA replication complexes found on virus induced membranes in infected cells (Hope, Diamond et al. 1997). Cleavage of the host translation initiation factor eIF4G by L results in marked down-regulation of host protein synthesis by shutting off cap dependent translation (Devaney, Vakharia et al. 1988; Medina, Domingo et al. 1993). This cleavage also leads to inhibited trafficking of proteins through the endoplasmic retriculum/Golgi secretory pathway by 2B and 2C and/or 2BC (Moffat, Howell et al. 2005; Moffat, Knox et al. 2007). The NS proteins 2B, 2C and 3A also play a role in the rearrangement of host cell membranes, which become the site of viral RNA replication and capsid assembly (O'Donnell, Pacheco et al. 2001; Pena, Moraes et al. 2008). 2B and 3C have the most conserved amino acid sequences between serotypes (Carrillo, Tulman et al. 2005).

1.3.5b Replication of the genome

Initially, the positive-sense genome of FMDV acts as a template for the synthesis of an anti-sense RNA, which in turn is used for the production of new positive-sense infectious genomes. A large excess of positive compared to negative strands exists within infected cells (Novak and Kirkegaard 1991). The *cis*-acting replication element (*cre*) serves as a template for 3D mediated uridylylation of VPg, forming VPgpUpU in infected cells (Crawford and Baltimore 1983), which acts as a primer for RNA polymerase (3D) driven genome replication. The 3’ poly ‘A’ tract is the initiation site for the synthesis of negative sense RNA that, following elongation, results in the formation of a double-stranded RNA/RNA molecule, the replicative form (RF). Free minus strands are not detectable *in vivo* (Grubman and Baxt 2004). It is interesting to note that, although VPg is linked to both positive and negative strands of FMDV, *cre*-dependent uridylylation of VPg is required for positive but not negative strand synthesis (Murray and Barton 2003).

FMDV RNA polymerase has poor proofreading capability, the mechanism behind which has been analysed in vitro (Arias, Arnold et al. 2008). Therefore, almost every time a genome is replicated, a mutation occurs so that FMDV exists as a population of closely related genomes. This genetic variability within FMDV populations has been extensively demonstrated in cell culture (Sobrino, Davila et al. 1983; Arias, Lazaro et al. 2001; Ruiz-Jarabo, Pariente et al. 2004). The creation and consequences of this genetic variability will be discussed further in the current Chapter, section 1.4.

1.3.5c Virus assembly

Picornavirus virion assembly can be broken down into four general steps, which include: (1) synthesis of the capsid protein precursor, (2) cleavage of precursor into VP0, VP3 and VP1 to form a non-covalent complex (protomer), (3) formation of pentamers from five protomers, (4) formation of icosahedral empty particles from 12 pentamers, linked to packaging of viral RNA (encapsidation) and cleavage of VP0 into VP4 and VP2 (‘maturation’), reviewed by Agol 2002. The specificity of positive strand encapsidation during virus assembly has been confirmed for poliovirus (Novak and Kirkegaard 1991). Although not specifically demonstrated for FMDV, it is assumed that any encapsidation of negative strands would be accidental as this would not be expected to be of advantage to the virus. Therefore, viral genome encapsidation forms an integral, potentially *rate* influencing*,* part of the intra-cellular replication process.

1.4 RNA virus evolution

There have been many cases of host cell tropism and host range modification associated with the genetic variability of RNA viruses, as reviewed by Baranowski, Ruiz-Jarabo et al. 2003. Domingo et al suggested over three decades ago that extensive genetic variability, across genome sites, was the basis for FMDV antigenic diversity (Domingo, Davila et al. 1980). High mutation rates, rapid replication kinetics and large population sizes all contribute to the heterogeneity of RNA virus populations. However, over evolutionary time, viral populations have been subjected to positive selection, negative selection and random drift. When the effective population size of a virus is small, it is predicted that genetic drift is the critical determinant of mutation frequency (Rouzine, Rodrigo et al. 2001). This is typically the case during multiple host-to-host transmissions and has been demonstrated for FMDV (Cottam, Haydon et al. 2006). Evolutionary transformations are less predictable during such stochastic processes.

1.4.1 The impact of replicative mode on mutation distribution

As discussed in previous sections, when a virus replicates inside a host cell, it uses its own genome as a template. However, once produced, the progeny can, in turn, become template for further replication. The distribution of mutants produced during an infection can subsequently vary greatly, depending on whether, and how many of, the progeny genomes become templates. Therefore, replication strategy provides important information about the generation of mutation. A recent study looked at the relationship between mutation frequency and replication strategy in positive-sense single-stranded RNA viruses (Thebaud, Chadoeuf et al. 2010), discussed below.

Numerous studies have been conducted to address the question of *optimal* replicative mode in positive sense RNA viruses (Chao, Rang et al. 2002; Krakauer and Komarova 2003; Regoes, Crotty et al. 2005; Sardanyes, Sole et al. 2009; Thebaud, Chadoeuf et al. 2010). Such studies are based on the assumption of either one of two basic modes of viral replication. One results in a Poisson distribution of mutations, whereby the parental virus is the only template used for production of progeny, the so called, ‘Stamping machine’ replication (SMR) model first proposed by Luria (Luria 1951). Conversely, if all genomic strand progeny are used as template for additional progeny the replication mode is effectively geometric (GR). An extension of this rationale is that, if a fraction of progeny acts as template for further replication, the replication mode will be a mixture of both SMR and GR. Thebaud et al. (2010) proposed that at high mutation rates, or when a high proportion of mutations are deleterious, the optimal replication strategy shifts towards the synthesis of more negative strands per positive strand, and “*in extremis”* towards the SMR mode. An equivalent model proposed by Sardanyes, Sole et al. (2009) also predicted that by employing the SMR mode, RNA viruses may increase their robustness against the accumulation of deleterious mutations. The same study also predicted that this increase in robustness would depend on assumptions made about the fitness landscape topology, such as strength of antagonistic and synergistic epistasis (Sardanyes, Sole et al. 2009).

1.4.2 Mutation rates and quasispecies

Mutation drives the heterogeneity upon which selection, recombination and genetic drift operate. Therefore, in order to understand the course of evolution through the progression of viral population structure over time, a clear understanding of both mutation and substitution rate is critical. Duffy et al. 2008 provides a review of current understanding of virus evolutionary rates, their determinants and how they are measured. *Mutation rate* is defined as the number of genetic mutations (point mutations, insertions and deletions) that accumulate per unit time or, for obligately lytic viruses per burst, per generation, or, per round of genomic replication. The low fidelity of the RNA-dependent RNA polymerase means RNA viruses often mutate at a higher rate than DNA viruses, which utilize higher fidelity DNA polymerases. However, similar mutation rates between some DNA and RNA viruses (an example of which is provided in Figure 1.2) suggest that polymerase fidelity may not be the only contributing factor and aspects of viral biology as genomic structure, size and replication speed may also play a role.

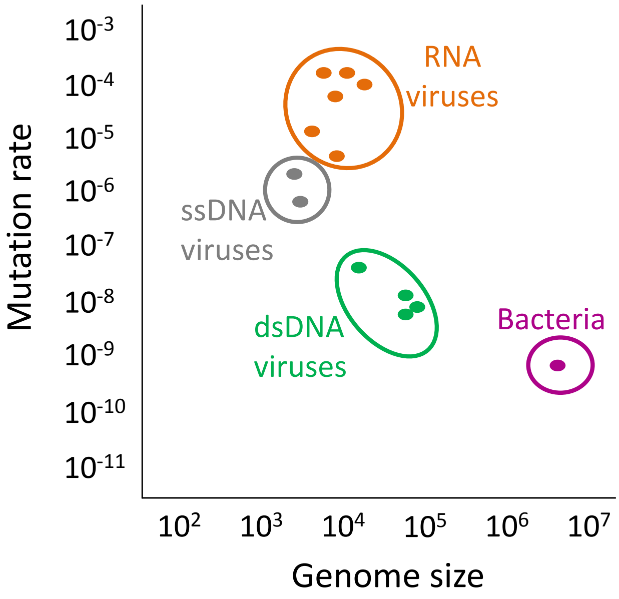


Figure 1.2

Per-site mutation rate against genome size (adapted from Gago, Elena et al. 2009). RNA viruses (left to right) are tobacco mosaic virus, human rhinovirus, poliovirus, vesicular stomatitis virus, bacteriophage φ6, and measles virus, Single-stranded DNA viruses are bacteriophage φX174 and bacteriophage m13. Double-stranded DNA viruses are bacteriophage λ, herpes simplex virus, bacteriophage T2, and bacteriophage T4. Bacteria is *Escherichia coli.*

The *mutation rate* of RNA viruses has been extensively reviewed (Holland, Spindler et al. 1982; Domingo and Holland 1997; Drake and Holland 1999; Duffy, Shackelton et al. 2008), and is commonly quoted to range between 1 x 10^{-3} to 1 x 10^{-5} misincorporations per nt per genome replication, with transitions occurring much more frequently than transversions (Kuge, Kawamura et al. 1989). Mutation rate is commonly measured either by the Luria-Delbruck fluctuation tests or mutation accumulation studies, as reviewed by Duffy, Shackelton et al. 2008. Estimates have been made, as depicted in Figure 1.2; however, it becomes substantially more difficult to drill down within a mutation rate *range* to ascertain a consistent rate for any one specific virus, including FMDV*.* A number of *in vitro* studies have looked at the *frequency* of mutations within populations of FMDV (Sobrino, Davila et al. 1983; Sierra, Davila et al. 2000; Arias, Lazaro et al. 2001; Pariente, Sierra et al. 2001; Airaksinen, Pariente et al. 2003; Gu, Zheng et al. 2006) but this frequency cannot be used to directly estimate mutation rate as it results from the combined action of mutation and selection. Mutation generated by poliovirus RNA polymerase have been more extensively studied (Parvin, Moscona et al. 1986; Sedivy, Capone et al. 1987; Ward, Stokes et al. 1988; Ward and Flanegan 1992; Wells, Plotch et al. 2001; Freistadt, Vaccaro et al. 2007). However, due to questions raised about the validity of experimental determination and possible over estimations, poliovirus mutation rate also remains to be accurately measured.

A possible reason for the uncertainty of mutation rate estimations lies with the difficulty in quantifying the number of generations of replication over which mutations are generated. Moreover, as discussed previously, the definition of genome replication requires care. The *in silico* study by Thebaud et al. 2010, which considers the replication strategy of RNA viruses, suggests combining such a model with measures of mutation frequency to achieve more accurate measures of mutation rate (Thebaud, Chadoeuf et al. 2010).

Substitution rate, defined as the number of fixed mutations (by natural selection or genetic drift), per nt site, per unit time, has been described as a complex product of four component factors including, underlying mutation rate, generation time, effective population size and fitness (Duffy, Shackelton et al. 2008). The substitution rate observed in the field for different serotypes of FMDV has been measured and found to lie in the range of 0.0004 – 0.045 substitutions per nt per year (Sobrino, Palma et al. 1986; Haydon, Samuel et al. 2001; Bastos, Haydon et al. 2003). This leads to a substantial level of genetic diversity, seen particularly within the nt sequence of the capsid proteins (Carrillo, Tulman et al. 2005), which accumulates, predominantly as synonymous changes, in a continuous, linear fashion over broad temporal and spatial scales (Villaverde, Martinez et al. 1991a; Elena, Gonzalez-Candelas et al. 1992; Cottam, Haydon et al. 2006; Valdazo-Gonzalez, Knowles et al. 2011). However, this rate is significantly lower than might have been expected, given the potential degree of genetic diversity generated within a single animal according to the mutation rate for RNA viruses quoted above. Potential explanations for this discrepancy are discussed in (Haydon, Samuel et al. 2001). It could be theorized, however, that the requirement to maintain an optimal balance between viral population stability and variability, for reasons discussed in the preceding section, occurs throughout the different spatial and temporal scales of infection.

Under the Darwinian model of evolution, ‘Survival of the fittest’ stipulates the best-adapted replicator is favoured by natural selection. However, under the quasispecies model of molecular evolution, first proposed by M. Eigen and his colleagues (Eigen 1971a; Eigen 1978), selection acts on ‘clouds’, of mutants, the quasispecies*,* not on individual sequences provided that mutation rate is high enough. At high mutation rates, the fittest organisms may not be the fastest replicators but rather those able to tolerate deleterious mutational effects, even at the cost of a low replication rate, dubbed ‘Survival of the flattest’. As discussed previously, RNA viruses have characteristically high mutation rates, consequently, the quasispecies model is often used to describe the evolutionary dynamics of RNA virus populations, including FMDV (Martinez, Carrillo et al. 1991; Villaverde, Martinez et al. 1991a; Domingo, Escarmis et al. 1992; Ibanez, Clotet et al. 2000; Ruiz-Jarabo, Arias et al. 2000; Mullan, Kenny-Walsh et al. 2001).

The original model assumed infinite population sizes and predicted deterministic dynamics, whereas, although large, viral populations are finite and subject to stochastic dynamics and neutral drift. Consequently, use of the quasispecies model for RNA virus evolution has been criticised (Jenkins, Worobey et al. 2001; Holmes and Moya 2002a). However, although not infinite in size, the co-operative population structure of RNA viruses (see below), in the form of mutational robustness, induced by mutational coupling, does not disappear when populations are finite (Bornberg-Bauer and Chan 1999; van Nimwegen, Crutchfield et al. 1999; Wilke 2001; Wilke and Adami 2003). RNA secondary structure has also been used as a fitness determinant in order to demonstrate quasispecies dynamics in finite populations of self-replicating RNA sequences (Forster, Adami et al. 2006).

The role played by selection in establishing the quasispecies dynamic has been elegantly investigated in poliovirus (Pfeiffer and Kirkegaard 2005; Vignuzzi, Stone et al. 2006). Vignuzzi et al, found that a poliovirus generating less genomic diversity, as a result of using a high fidelity polymerase, led to a loss of neurotropism and attenuated pathogenesis in mice. By observing *co-operative* interactions between different variants in the 'cloud', the authors propose a rationale for the role of quasispecies diversity in infectivity. This rationale hypothesized that different *sub*-populations may serve different roles within the overall 'cloud' and facilitate colonization of different tissues and therefore, maintaining quasispecies complexity enables the systematic spread of viral populations at the intra-host scale (Vignuzzi, Stone et al. 2006). A similar study by (Pfeiffer and Kirkegaard 2005) had also found links between decreased viral diversity and reduced ability of the viral population as a whole to adapt within the host environment.

Properties consistent with the quasispecies theory, such as an error threshold and differences in mutation spectra affecting fitness, have been demonstrated for FMDV populations *in vitro.* These so called, 'quasispecies' dynamics are reviewed in terms of antiviral strategies based on virus entry into error catastrophe in (Domingo, Escarmis et al. 2005). However, whether selection does indeed act on FMDV populations, as a single unit, in a natural setting is still be demonstrated. Therefore, I will use the term viral swarm in place of *quasispecies* to describe the heterogeneity of FMDV populations within this thesis.

1.4.3 Approaches used to study viral sequence variability

Prior to the advent of next-generation sequencing, which will be discussed in the current Chapter, section 1.5, there were a number of alternative methods used for the study of viral populations. Biological cloning has been extensively used to produce multiple plaque purified viral clones to investigate viral population heterogeneity and implications for fitness (Escarmis, Davila et al. 1999; Arias, Ruiz-Jarabo et al. 2004; Domingo, Pariente et al. 2005; Escarmis, Lazaro et al. 2008). This technique only provides a partial picture of the original viral population as it is restricted to *fit* viral particles that initially create the plaques. Alternatively, molecular cloning techniques have also been used to investigate viral population diversity at the within host scale (Cottam, King et al. 2009b; Murcia, Baillie et al. 2010; Bull, Luciani et al. 2011; Bull, Eden et al. 2012). Full length cDNA sequences can be cloned into plasmid vectors, as has been demonstrated for viruses such as hepatitis C virus (HCV) (Date, Kato et al. 2012) and FMDV (Ellard, Drew et al. 1999). It should be noted, however, that these techniques still incur artefactual mutations (artefacts) during RT-PCR. These artefacts can be avoided by directly sequencing PCR products derived by the amplification of a single target molecule (Simmonds, Balfe et al. 1990). Even if an artefact occurred in the first PCR cycle, half of the templates will still have the original nucleotide at that position (Smith, McAllister et al. 1997). However, all these techniques are very time consuming and offer only relatively limited resolution of the viral population.

1.4.4 Foot-and-mouth disease evolution

It has been hypothesized that the population heterogeneity of RNA viruses affords them greater adaptability during infection of a host (Coffin 1995; Domingo, Escarmis et al. 1996; Garcia-Arriaza, Manrubia et al. 2004). FMDV viral populations have been shown to exhibit what is known as genomic ‘memory’ (Ruiz-Jarabo, Arias et al. 2000; Arias, Ruiz-Jarabo et al. 2004), which refers to the maintenance at low frequency of a previously dominant virus variant, potentially aiding such adaptability. Additionally, populations of FMDV have been shown to adapt to different multiplicities of infection (MOI) *in vitro* (Sevilla, Ruiz-Jarabo et al. 1998) and co-evolve alongside different cell lines (Ruiz-Jarabo, Pariente et al. 2004) as well as maintain defective RNAs (Charpentier, Davila et al. 1996).

Processes and events that shape RNA virus populations during virus/ host interaction include those within diverse cellular environments as well as the host immune response itself, which exerts selective pressure upon specific regions of the FMDV genome. Such processes also occur during transmission across ‘host-to-host’ and ‘tissue-to-tissue’ barriers. Processes that shape viral populations can either be driven by selection or random genetic drift, such as bottlenecking. Bottleneck events occur both within and between hosts, so will be discussed separately. The question remains whether conventional Sanger sequencing of the consensus can provide sufficient resolution to distinguish between within host FMDV populations and, subsequently, to characterise the transmission of diversity between hosts.

1.4.4a Inter-host viral transmission

Transmission of FMDV is achieved either through direct or indirect contact with an infected animal, product or object (fomite). It is believed that fomite virus outside a susceptible host does not replicate and therefore remains evolutionary ‘dormant’;

which can impact on the predicted error rate at the epidemic scale. Genetic sequence data has been used to successfully trace the movement of FMDV at the global scale and within a single epidemic. However, both the heterogeneous nature of within host viral populations and the number of transmitted viruses between hosts may influence the rate of mutation fixation (Kinnunen, Poyry et al. 1991; Villaverde, Martinez et al. 1991b).

Although virus may enter a susceptible host through damaged integument, transfer of airborne droplets (from the breath of infected animals or after atmospheric re-suspension from contaminated materials) to the respiratory tract of recipient animals is the most probable form of transmission between animals in close proximity. However, the species of both donor and recipient animal influences the means of transmission; in turn, the means of transmission will influence the route and dose of infection, which in turn may influence the site/s of primary replication and pattern of viral dissemination.

1.4.4b Intra-host viral transmission

The intra-host scale studies discussed in this thesis were conducted on bovine samples; therefore viral dissemination in cattle will be discussed in some detail. As discussed previously for poliovirus, population heterogeneity may influence site of viral replication, in the form of tissue tropism. Therefore, it is important to put observed viral population characteristics into context with both dissemination route and location of viral replication if intra-host scale microevolutionary dynamics are to be investigated.

The most accepted route of FMDV intra-host dissemination, after inhalation of airborne virus, involves initial replication within the pharynx followed by virus spread through regional lymph nodes (Henderson 1948) and into the circulation (McVicar and Sutmoller 1976; Burrows, Mann et al. 1981; Alexandersen, Zhang et al. 2002a; Alexandersen, Zhang et al. 2003). Serum-associated viraemia, commonly lasting 4-5 days (Cottral and Bachrach 1968; Alexandersen, Zhang et al. 2002b; Alexandersen, Quan et al. 2003), seeds secondary sites and multiple cycles of viral replication mainly in the cornified epithelia of the skin, tongue and mouth (Burrows, Mann et al. 1981; Alexandersen, Oleksiewicz et al. 2001; Oleksiewicz, Donaldson et al. 2001). However, tissue specific sampling can offer improved resolution and further dissection of the acute phase of disease, in addition to the analysis of oesophageal-pharyngeal fluid (OP, ‘probang’) or serum samples.

Arzt et al. (2010) took up to 40 tissue samples per animal and demonstrated that the nasopharyngeal region was the probable site of primary replication in previraemic cattle aerosol inoculated with FMDV O$_1$Manisa. The same study also demonstrated that progression towards viraemia coincided with a marked increase of viral load in pulmonary tissues and a substantial decrease in nasopharyngeal tissues (Arzt, Pacheco et al. 2010). A number of factors have been indicated or identified as potential causes for localized infection of the nasopharyngeal region. These include the concentration of virus in the air or initial infective dose (McVicar and Sutmoller 1976), other resident viruses (Graves, McVicar et al. 1971), the proportion of defective interfering particles or variants in the viral population (Sutmoller and McVicar 1972) and local antibody (McVicar and Sutmoller 1974). The size distribution of aerosol particles carrying FMDV will also affect whether viral challenge will be primarily at the upper or lower respiratory tract. Interestingly, as well as infection via direct exposure to aerosolized virus, the pharyngeal area can also be infected via the blood stream (Sutmoller and McVicar 1976), as was also demonstrated within hours after introduction of virus into the udder through the teat canal (Burrows, Mann et al. 1971). By isolating the upper and lower respiratory tract, the study by Sutmoller et al. (1976) was able to show that either were able to serve as potential portals of FMDV entry into the systemic circulation in cattle.

It has been hypothesized that sustained FMDV viraemia is maintained by viral replication in lesional and/or nonlesional skin (Brown, Olander et al. 1995; Alexandersen and Mowat 2005), however, this has never been unambiguously established. Conversely, there are several factors that support the concept that the lungs may be important amplifiers of FMDV, maintaining high titer viraemia within an infected host. Arzt, Pacheco et al. (2010) demonstrated the relatively high titre of FMDV RNA per mg and large quantities of FMDV structural and non-structural antigens in the lungs of viraemic cattle. This study also highlighted the overall mass and extensive vascularity of the lungs as additional supporting factors (Arzt, Pacheco et al. 2010).

1.4.4c Bottlenecks

Bottleneck events can occur during transmission both within and between hosts. The impact of intra-host scale bottleneck events has been described for a range of pathogens, including poliovirus (Pfeiffer and Kirkegaard 2006) and FMDV (Carrillo, Lu et al. 2007a). It has been theorized that the fine balance between population stability and flexibility has been optimized during the evolution of the virus. Where too many mutations per genome can bring a viral population to extinction (Sierra, Davila et al. 2000), too few can also cause extinction by reducing the ability of the virus to adapt to different environments. However, it is interesting to note the advantageous impact of bottleneck events on RNA virus populations that are evolving at an increased error rate (Cases-Gonzalez, Arribas et al. 2008).

1.4.4d The immune response

Much investigation has been conducted to characterize the host response to FMDV, and a review of this work has been provided by Golde, de Los Santos et al. 2011. *In vitro* studies have provided invaluable insights that would not have been possible to conduct *in vivo*, some of which have been confirmed to be relevant in live animals.

To ensure the effective spread of infection within and between susceptible hosts, FMDV has evolved multiple mechanisms to subvert the early immune response. Infection is initiated, disseminated throughout the body and infectious progeny produced in less than 7 days. Interactions between FMDV, antigen-presenting cells and their precursors results in suboptimal immune function, favouring viral replication and delayed specific adaptive T-cell responses (Golde, de Los Santos et al. 2011). Numerous studies have identified antigenic variants within FMDV populations in vivo and in vitro (Diez, Mateu et al. 1989; Borrego, Novella et al. 1993; Holguin, Hernandez et al. 1997).

A thorough understanding of virus/host interaction and response is required for effective control of disease in livestock populations through targeted vaccination strategies.

1.4.5 Constraints on genetic variability

While the advantage of population heterogeneity through mutation is the promotion of viral adaptability, there are obvious limitations to these in terms of virus viability. A number of studies have investigated the concept of an error threshold, above which there may be a decrease in infectivity and population extinction (Holmes 2003; Pariente, Airaksinen et al. 2003; Domingo, Pariente et al. 2005; Cases-Gonzalez, Arribas et al. 2008). Epistasis offers another constraint on mutation, whereby in order for a particular mutation to be viable, it is necessary for a compensatory mutation to occur elsewhere in the genome to counteract any deleterious effect. The canalization effect of epistasis has been investigated in the RNA bacteriophage φ6 (Burch and Chao 2004). Even synonymous changes may impact virus viability through RNA secondary structure alteration, which has been found to influence virus pathogenesis (Witwer, Rauscher et al. 2001; Hofacker, Stadler et al. 2004; Simmonds, Tuplin et al. 2004).

1.4.6 Genome rearrangements

As discussed above, the potential to generate mutation is part of the replicative mechanism of RNA viruses, including Picornaviruses. These mutations come in different forms, nt substitutions and rearrangements (deletions, duplications, insertions and recombinations). However, long-term evolution results in the consecutive fixation of only a minute proportion of these mutations, so that rearrangements may occur rarely. The biological significance of these rearrangements is not clear-cut; they may contribute to genome flexibility and/or stability. For example, recombination between altered genomes containing deleterious mutations may result in the regeneration of the parent-type genome. Such a ‘rescue’ mechanism may be particularly important during low dose inter-host transmissions that can be accompanied by virus fitness decline (Agol, Belov et al. 2001).

Naturally occurring recombination between virus genomes in an infected cell may occur via replicative mechanisms (synthesis of a nascent strand is started on one parental RNA molecule and is completed on another due to template switching) or non-replicative mechanisms (fragments of different parental RNA molecules may be covalently joined in a non-replicative reaction). Recombination has been estimated to play an important role in the evolution of an increasing number of RNA viruses, including Coronaviruses (Liao and Lai 1992), a range of plant viruses (reviewed in Sztuba-Solinska, Urbanowicz et al. 2011), Human immunodeficiency virus (HIV) (Lole, Bollinger et al. 1999; Rhodes, Wargo et al. 2003), and FMDV (Haydon, Bastos et al. 2004; Heath, van der Walt et al. 2006; Jackson, O'Neill et al. 2007; Li, Shang et al. 2007; Lewis-Rogers, McClellan et al. 2008; Lee, Oem et al. 2009), to name but a few. However, a large proportion of these studies rely on bioinformatic methods and sequence data to reconstruct recombination events. Exceptions include HIV (Rhodes, Wargo et al. 2003), plant viruses (Froissart, Roze et al. 2005; Sztuba-Solinska, Dzianott et al. 2011) and animal virus (Liao and Lai 1992) studies, which have used the insertion of neutral genome markers to experimentally assess recombination rate within the viral populations. Such techniques are yet to be applied to FMDV.

The impact of mechanistically induced recombination during RT-PCR will be discussed in the following section.

1.5 Sequencing approaches and technologies

1.5.1 First-generation to second-generation sequencing

Fundamental to any genetics investigation is the determination of genotypes through DNA sequencing. Dideoxynucleotide sequencing of DNA, first described by Sanger *et al*. in 1977 (Sanger, Nicklen et al. 1977), has undergone steady and substantial upgrades over the years. The scale at which sequence data is produced now requires a specialized and devoted infrastructure of bioinformatics, computer databases and instrumentation. These advances have been especially evident over the last ten years, largely due to the efforts necessary to sequence the human genome (the Human Genome Project [HGP], coordinated by the U.S. Department of Energy and National Institute of Health, was completed in 2003). Two pioneering papers reporting new sequencing developments in 2005 provided the first glimpse of things to come (Margulies, Egholm et al. 2005; Shendure, Porreca et al. 2005). These new DNA sequencing technologies are collectively referred to as 'next-generation' sequencing (NGS), 'high-throughput' sequencing, 'ultra-deep' sequencing (UDS), or 'massively parallel' sequencing. The term 'next-generation' sequencing (NGS) will be used throughout this thesis. Figure 1.3a and b provides a comparison of conventional Sanger and next-generation sequencing (a generalized view of NGS is given, as although some aspects of different platforms vary, the principles of DNA fragmentation, cluster generation and cyclic sequencing to image-based data collection, are common to these technologies). A synopsis of both Sanger and early NGS platforms is given by Shendure et al. (2008), a summary of which is given below, using RNA virus genetic analysis as an example.

Using the Sanger method, once target RNA has been isolated, reverse transcribed and PCR amplified, DNA template, or 'PCR product' is subjected to a 'cycle sequencing' reaction. Within this reaction, cycles of template denaturation, primer annealing and primer extension are performed. Each round of primer extension is stochastically determined by the incorporation of deoxynucleotides (dNTPs) and fluorescently labelled dideoxynucleotides (ddNTPs). In the resulting mixture of end-labelled extension products, the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. The sequence is determined by high-resolution electrophoretic separation of the single-stranded, end-labelled extension products in a capillary-based polymer gel. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, coupled to four-colour detection of emission spectra, provides the readout that is represented in a Sanger sequencing pherogram. Software translate these traces into DNA sequence, while also generating error probabilities for each base-call (Ewing and Green 1998; Ewing, Hillier et al. 1998).

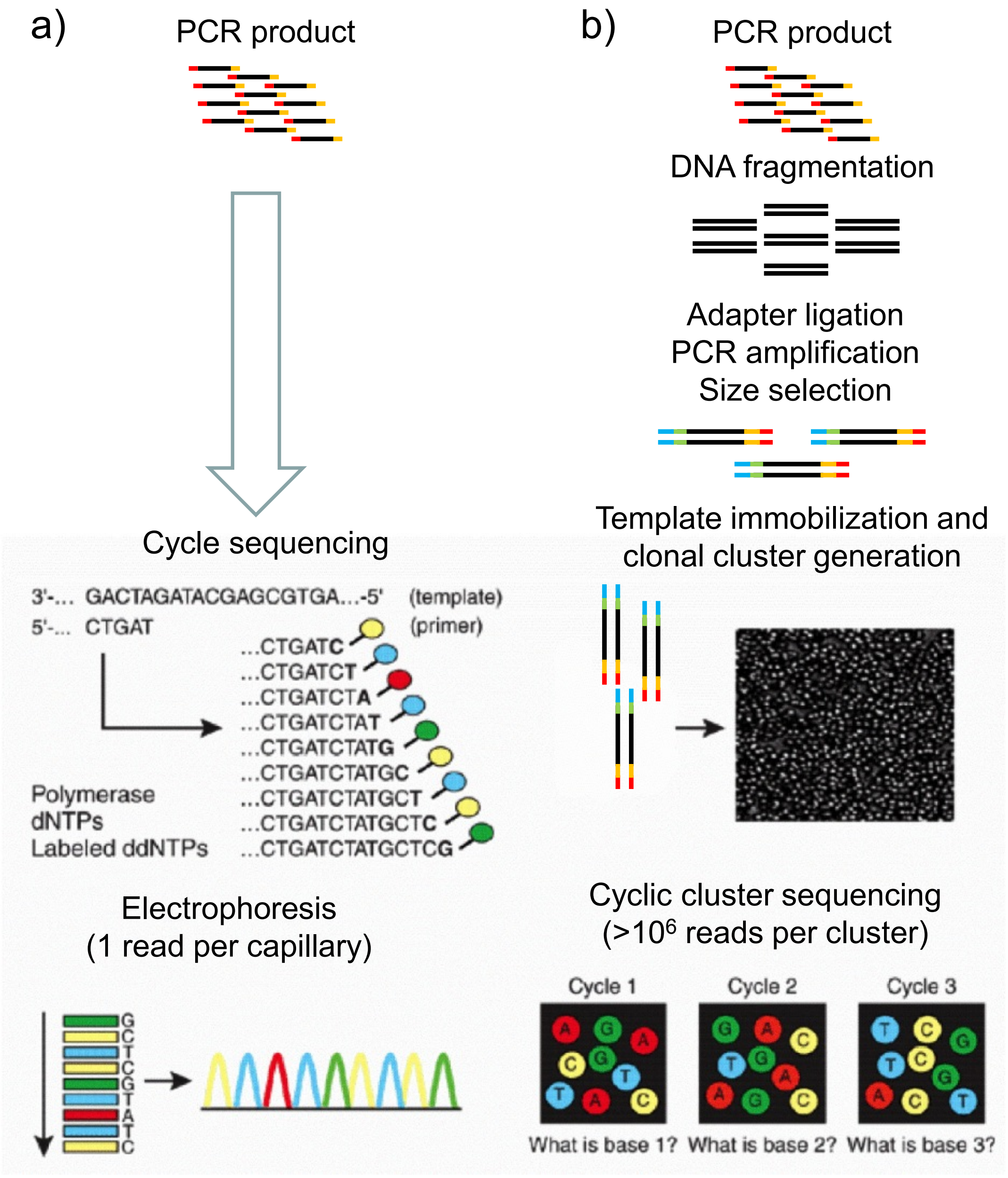


Figure 1.3

Schematic of a generalized workflow for **A** conventional Sanger versus **B** next-generation sequencing. Adapted from (Shendure and Ji 2008)

Library preparation for NGS typically starts with the fragmentation of viral template DNA ('PCR product'), followed by the ligation of adapter sequences. Alternative protocols can be used with mate-paired or paired-end sequencing tags for additional distance information, which will be discussed in more detail for the Illumina platform in section 1.5.2a. The generation of clonally clustered amplicons to serve as sequencing templates can be achieved by several approaches, including emulsion PCR (Dressman, Yan et al. 2003) and bridge PCR (Adessi, Matton et al. 2000). Common to the library preparation of many NGS platforms is that PCR amplicons, derived from any given molecule, are spatially clustered, either to a single location on a planar substrate (bridge PCR), or to the surface of micro-scale beads, which can be recovered and arrayed (emulsion PCR). The sequencing process itself is based on alternating cycles of enzyme-driven biochemistry and image-based data acquisition. Metzker (2010) provides a comprehensive review of the template preparation, sequencing and imaging methodologies used by the main NGS platforms (Metzker 2010).

The majority of NGS platforms, excluding the most advanced single molecule platforms (Table 1.0), are still to achieve read lengths equivalent to those possible with Sanger sequencing (up to ~ 1,000 bp). However, the micro-scale, template immobilization (or 'cluster') based sequencing of the NGS platforms enables a much higher degree of parallelism compared to conventional capillary-based sequencing. Additionally, because clusters of sequencing template are immobilized on a planar surface at the micro-scale, they can be enzymatically manipulated by a single, picrolitre reagent volume, drastically reducing costs. Shendure et al. (2008) valued the cost of 'high-throughput' shotgun genomic *Sanger* sequencing at $0.05 per kilobase. The same study valued an average cost of five major NGS platforms at $13.00 per megabase (Shendure and Ji 2008). [Using current exchange rates, this equates to a cost differential of just over £300 per kilobase].

The *consensus* base given at each nucleotide position by the Sanger sequencing method only identifies the predominant or major viral sequence present in a sample. If background fluorescence is low and sequence peaks are relatively high, ambiguities present within 20% of the viral population may be distinguished using Sanger sequencing. Therefore, consensus sequencing remains uninformative about minority variants that are present. Conversely, the ultra-deep coverage provided by NGS may potentially reveal information about the viral swarm missed by consensus sequencing by identifying mutations present in only a small fraction of the population. This additional information provided by NGS may allow differentiation between closely related viral populations at the inter and intra-host scale.

1.5.2 Next-generation sequencing

The following section provides an overview of NGS as a method. A more detailed account of where this technology has been applied to further our knowledge of viral evolution is provided in Chapter 3 and 5. Numerous reviews of the different NGS platforms available, their respective strengths and weaknesses in terms of application and cost, have been compiled (Mardis 2008; Marguerat, Wilhelm et al. 2008; Shendure and Ji 2008; Ansorge 2009; Metzker 2010; Glenn 2011); therefore, only a brief summary will be given here.

The main commercially available NGS platforms are from Roche (454), Illumina (Solexa) and Life Technologies (Ion Torrent), however, new commercial providers and platforms continue to be established, each platform applying distinct principles, which result in differences in number and length of sequence reads, which may provide particular advantages and disadvantages for individual applications. A comparison of some of these platforms is provided in Table 1.0; however, as a rapidly developing field of technology, the statistics given here are likely to change and are therefore meant as demonstrative only. It should be noted, for a variety of reasons, a comparison of error rate between platforms is particularly problematic (Glenn 2011) and therefore these values should be taken as approximations. All the commercial providers discussed here maintain informative websites that detail the most up-to-date specifications for each of the individual platforms.

Table 1.1 Comparison of next-generation sequencing platforms

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	GB per run	Primary errors	Error rate (%)	Pros	Cons	Biological applications
Roche/454's GS FLX Titanium	Sr, Pr/ emPCR	PS	330[1]	0.35	0.45	Indel	1[5]	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homopolymer repeats	Bacterial and insect genome de novo assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics
Solexa/Illumina GA IIx	Sr, Pr/ solid-phase bridge	RT	75 or 100	4[2], 9[3]	18[2], 35[3]	Substitution	≥0.1[5]	Currently the most widely used platform in the field	Relatively low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
Life's/APG's SOLiD 3	Sr, Pr/ emPCR	Cleavable probe SBL	50	7[2], 14[3]	30[2], 50[3]	A-T bias	>0.06[5]	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
Polonator G.007	Pr only/ emPCR	Non-cleavable probe SBL	26	5[3]	12[3]	N/A	N/A	Second cheapest platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery
Ion torrent by Life technologies. Ion PGM-318 chip	Sr, Pr/ emPCR	Semicon	200	0.19[2]	1[2]	Indel	~1[4]	Shortest run time and cheapest machine	High error rates in homopolymer repeats	Bacterial whole-genome sequencing; whole-genome resequencing; de novo RNA expression studies
Helicos BioSciences Heliscope	Sr, Pr/ single molecule	RT	32[1]	8[3]	37[3]	Substitution Insertion Deletion	0.2[4] 1.5 3.0	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods
Pacific Biosciences RS	Sr only/single molecule	Real-time	~3000[1]	<0.42	0.12	CG deletions	16[5]	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome seq; complements other resequencing efforts in discovering large structural variants&haplotype blocks

[1] Average read-length [2] Single read run [3] Paired read run (either mate-paired or paired-end) [4] Information based on company sources alone [5] Information taken from (Glenn 2011). Sr, Single read; Pr, Paired read; GA, Genome Analyzer; GS, Genome Sequencer; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; Semicon, semiconductor sequencing; SOLid, support oligonucleotide ligation detection. Adapted from (Metzker 2010).

The machines in Table 1.0 range in price from approximately £623,000 (Helicos Biosciences HeliScope) to £30,000 (Life Technologies, Ion Torrent, Ion PGM). Similarly to the Ion Torrent PGM sequencer (Life Technologies), other 'desktop' versions of the previously mentioned platforms are being made available, including the MiSeq (Illumina) and GS Junior (Roche), which predominantly boast improved run times. Currently under development at Oxford Nanopore Technologies, is an alternative single molecule sequencing technology based on nanopore sensing (reviewed by Branton, Deamer et al. 2008). The technology is based on the principle of molecule induced changes in ionic current across a pore containing membrane bilayer. When a target molecule, for example, a DNA or RNA base, passes through the pore or near its aperture, this event creates a characteristic disruption in current. It is subsequently possible to identify the molecule by measuring this change in current. This technology is very scalable and, similarly to the Pacific Biosciences (PacBio) RS platform, enables real-time analyses. Multiple desktop nanopore instruments (GridION nodes) can be networked into larger co-operating units compared to the miniaturised, disposable version of the technology (MinION) which is the size of a USB stick that can be plugged directly into a laptop or desktop computer. Currently, the Oxford Nanopore and Helicos systems are the only systems capable of direct RNA sequencing; however, this is likely to change in the near future.

The sheer volume of data that can be produced relatively cheaply has resulted in such genetic analysis tools as microarrays being replaced by sequence-based methods for certain applications, including gene expression studies. Here rare transcripts can be identified and quantified by NGS without prior knowledge of a particular gene where it can also provide information about alternative splicing and sequence variation in identified genes (Wang, Gerstein et al. 2009). However, microarrays used for enrichment in conjunction with NGS can also be powerful tools in terms of high-throughput targeting strategies (Chou, Liu et al. 2010; Milan, Coppe et al. 2011; Hong, Doddapaneni et al. 2012).

The application of NGS to the estimation of within-host virus population diversity was recognised early on, where it was initially used for the detection of low-frequency drug resistance mutations in HIV (Hoffmann, Minkah et al. 2007; Wang, Mitsuya et al. 2007). The use of NGS to investigate human, animal and plant viral population dynamics is increasing rapidly (as reviewed by Beerenwinkel and Zagordi 2011). As discussed previously, the choice of platform will depend on its application and target organism. The 454 platform (Roche) is often utilized for viral population investigations, as the comparatively longer reads it produces may be advantageous for *de novo* sequencing and assembly, for example, novel Orthobunyavirus discovery (Schmallenberg virus) (Hoffmann, Scheuch et al. 2012). Alternatively, this platform can be useful for re-sequencing of viral genomes with a high proportion of homopolymeric repeats. Conversely, NGS platforms, such as the Genome Analyzer (Illumina) and SOLiD (Life Technologies), produce a higher volume of short reads. The lower costs and increased number of reads associated with shorter read-lengths are better suited for re-sequencing and frequency (or counting) based applications. The low error rates of these platforms also increase their application for variant discovery. However, of the two ‘short read length to high volume’ platforms, the Illumina Genome Analyzer provided the longest read length and therefore, for a combination of these reasons, was selected as the NGS platform for this project.

1.5.2a The Illumina Genome Analyzer platform

The process of sequencing on the Illumina GA system can be separated into three phases, ‘library preparation’, ‘cluster generation’ and ‘sequencing’. Developments within the platform from the GA II to GA IIx systems involved improvements in chemistry and read length (from 50 nt to 75 nt), however the workflow essentially remained the same. As mentioned previously, template preparation, or DNA ‘library preparation’, for NGS typically starts with DNA fragmentation, which can be achieved via a range of techniques, including sonication, nebulization and enzymatic methods. The majority of imaging systems are not able to detect single fluorescent events, therefore template amplification is required. Although subsequent immobilization and amplification of template fragments to a solid surface is common amongst NGS platforms, the precise method by which this is achieved can vary. For the Illumina platform, randomly distributed, clonally amplified clusters are produced by solid-phase, or ‘Bridge’ amplification (Figure 1.4a), whereby multiple samples can be combined, or ‘Multiplexed’ (Figure 1.4b), and amplified simultaneously on a glass slide, or ‘Flow cell’, containing eight channels, during ‘cluster generation’. During the multiplexed sequencing method, DNA libraries are “tagged” with a unique six nt long identifier, or index. An automated two or three-read sequencing strategy (Figure 1.3b) identifies each uniquely tagged sample for individual downstream analysis. This approach is highly accurate and, due to the inherent redundancy in the index design, allows for indexes that differ by one base still to be used as sample identifiers. The option for either a two or three-read sequencing strategy, occurs according to whether the option to perform single (used during this project) or ‘paired’ read sequencing is taken.

Using the Illumina platform, ‘paired’ read sequencing can be either ‘paired-end’ or ‘mate-paired’, both of which provide information about physical distance between the two synthesised reads, in addition to the sequence information. This distance information is particularly useful to resolve larger structural rearrangements (insertions, deletions, inversions), or for *de novo* assembly and assemblies across repetitive regions (Van Nieuwerburgh, Thompson et al. 2012), none of which was necessary for this project. A summary of the Illumina GA library preparation workflow is included in Appendix 1.

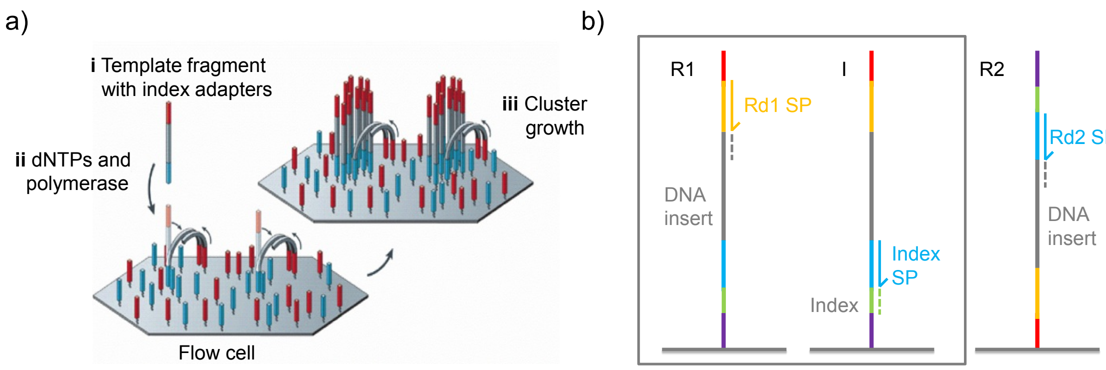


Figure 1.4

Schematic of Illumina GA cluster generation **a)** Bridge amplification: i template fragments ligated to index adapters randomly attach to a dense lawn of primers covalently bound to the inside surface of the flow cell channels, ii Unlabelled nucleotides and enzyme are added to initiate sold-phase bridge amplification, iii The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate. Adapted from (Metzker 2010). **b)** Multiplexed sequencing process: R1 (Read 1 dotted line), is generated using the R1 sequencing primer (Rd1 SP in orange); I (6 bp index read dotted line), is generated after the R1 product is removed and the Index sequencing primer (Index SP in blue) is annealed to the same strand. A single-read sequencing strategy consists of both R1 and I, as indicated by the grey box (used during this project). If a paired-end read is required, the original template strand is used to regenerate the complementary strand after which the original strand is removed and the complementary strand acts as a template for application of R2 (Read 2 dotted line) primed by the R2 sequencing primer (Rd2 SP in blue). Adapted from information provided at www.illumina.com.

The final ‘sequencing’ phase of the Illumina GA platform is performed by the *cyclic reversible termination* (CRT) method, using CRT chemistry (Bentley, Balasubramanian et al. 2008). A brief summary of this process is provided in Figure 1.5, however, a more comprehensive review is given by Metzket (2010), which includes details of the modified nucleotides (reversible terminators) used (Metzker 2010).

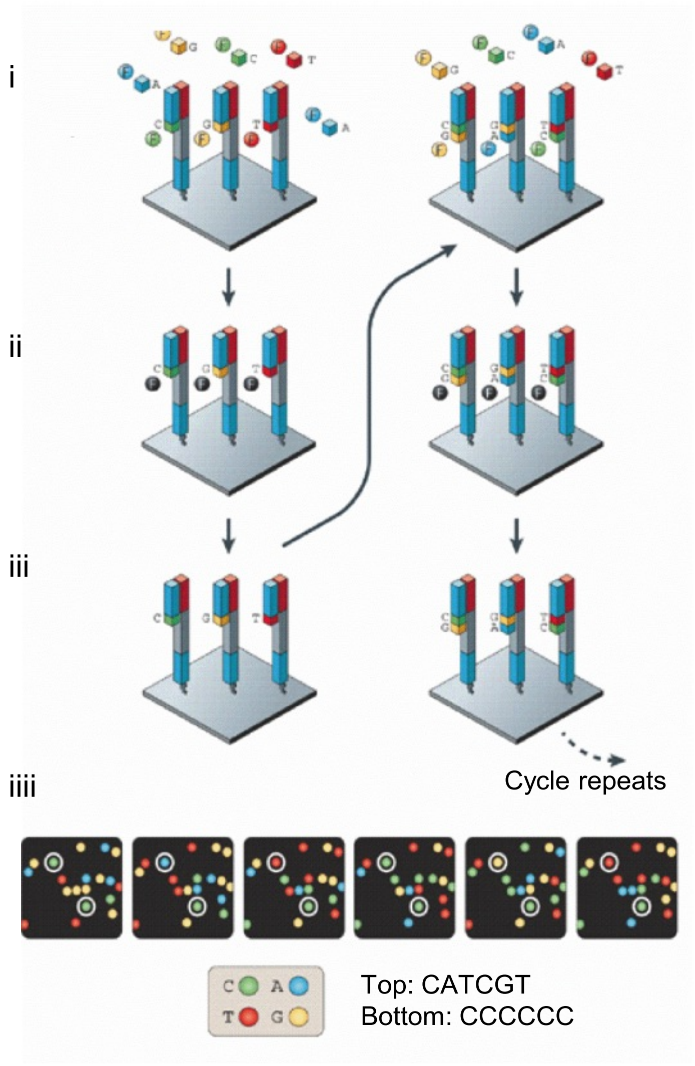


Figure 1.5

A schematic of the four-colour cyclic reversible termination (CRT) sequencing method used within the Illumina GA platform. i All four nucleotides labelled with a different dye (hence 'four-colour' CRT) are added simultaneously, ii A wash followed by four colour imaging is performed, iii Addition of the reducing agent tris(2-carboxyethyl)phosphate (TCEP), simultaneously removes the fluorescent dyes and regenerates the 3'-blocked reversible terminator, which is followed by another wash before the cycle is repeated iiii The four-colour images here highlight the sequencing data from two clonally amplified templates. Adapted from (Metzker 2010).

1.5.2b Challenges for viral population analysis

The experimental, or *in vitro*, processing of viral RNA, including the production of cDNA via reverse transcription (RT), followed by PCR amplification, produce errors within viral genomes. True viral mutations incurred *in vivo* are problematic to separate from these methodologically introduced errors if they occur at a frequency below or at the same level. This therefore becomes an increasing problem with the depth of detection possible using NGS.

Recombination, or rearrangement, is an additional genetic phenomenon that occurs within viral populations both *in vivo* and during *in vitro* processing by RT-PCR. As discussed for FMDV in the current chapter, section 1.4.4, many notable viruses that cause human disease, for example, HIV, Hepatitis B and C virus (HBV and HCV respectively) all recombine within their hosts. However, viral RNA sequence data will also include chimeric cDNA artefacts that are generated by template switching during RT and PCR amplification (Mathieu-Daude, Welsh et al. 1996). The impact of such artefactual events on the estimation of *true* viral population diversity is a problem that will become clearer as improved NGS technologies appear, with longer reads and, as such, will be discussed further in the final chapter of this thesis.

Potential artefacts and biases arise during and because of the actual process of sequencing itself. For example, upon imaging of a clonally amplified cluster (Illumina bridge amplification), the observed signal is a consensus of the nucleotides or probes added to the identical templates for a given cycle. A greater demand is therefore placed on the efficiency of the addition process, whereby incomplete extension of the template ensemble results in lagging and leading-strand dephasing (Metzker 2010). Signal dephasing, during step-wise addition methods, increases fluorescence noise, which can cause base miss-calling and shorter reads (Erlich, Mitra et al. 2008) as referenced in (Metzker 2010). A study by Nakamura et al. (2011), speculated that certain sequence-specific errors (SSE) incurred within reads from the Illumina GA favour dephasing by inhibiting single-base elongation, by i) folding single-stranded DNA and ii) alternating enzyme preference. The authors of this study highlight the substantial contribution this phenomenon has to variations in sequence coverage and its potential cause of false single-nucleotide polymorphism (SNP) calls (Nakamura, Oshima et al. 2011). It has been noted that an overrepresentation of amplicon ends, in particular the last 50 bp, can account for more than 50% of the sequenced bases (Harismendy and Frazer 2009). It has also been speculated that this overrepresentation and the per-base sequencing coverage variability, known to be an important issue for

NGS, regardless of input material type (Hillier, Marth et al. 2008; Ossowski, Schneeberger et al. 2008), could result from the sample preparation and fragmentation method (Harismendy and Frazer 2009).

Substitutions are the most common error type incurred during Illumina sequencing itself, with a higher proportion of errors occurring when the previous nucleotide incorporated is a ‘G’ base (Dohm, Lottaz et al. 2008). Additionally, an underrepresentation of AT-rich (Dohm, Lottaz et al. 2008; Hillier, Marth et al. 2008; Harismendy, Ng et al. 2009) and GC-rich regions (Hillier, Marth et al. 2008; Harismendy, Ng et al. 2009) has been revealed on genome analysis of Illumina sequence data, most likely caused by amplification bias during template preparation (Hillier, Marth et al. 2008).

Once NGS sequence reads have been generated, they can either be aligned to a known reference sequence or assembled *de novo.* Limitations of the alignment method become especially apparent when attempting to place short NGS reads within repetitive regions of the reference genome (Metzker 2010). Longer reads or reads generated by paired-end or mate-paired strategies can help to resolve such alignment issues. Consequently, the length of reads themselves forms a substantial limitation of current NGS platforms by providing incomplete information on the viral population structure. These constraints are compounded by the lack of linkage between mutations observed on different reads leading to difficulties in haplotype reconstruction and haplotype frequency estimations. While the lack of linkage between mutations does not directly impact upon the *frequency* based analysis in this thesis, it limits the extrapolation of functional impacts to mutations within single reads.

1.5.2c Data Analysis

The common file format for NGS sequence data is FASTQ, as reviewed by Cock, Fields et al. 2010. Briefly, the FASTQ file format provides a simple extension to the FASTA format by storing a numeric quality score (PHRED qualities) associated with each nucleotide in a sequence. FASTQ was first widely used by the Sanger Institute, which is why the Sanger specification of the *standard* FASTQ format is commonly used. Although the Illumina output file looks almost identical to the standard, Sanger FASTQ format (see Figure 1.6 for example FASTQ-Illumina file format), quality is scaled differently. The FASTQ variant used by the current Illumina pipeline encodes PHRED scores with an ASCII offset of 64, and so can hold PHRED scores from 0 to 62 (ASCII 64-126), although scores were only expected to range between 0-40 within raw Illumina data at the time of review by Cock et al. 2010.

```
@HWI-B5-690_0092_FC:2:1:2420:1185#CTTGTA/1
CAGTTTCCCGATTATGATTTTTATTGCCGTGGTAGTGTTCGGCTTTAAGGCTTTT
GTGATTGTGCCGCAGCAG
+HWI-B5-690_0092_FC:2:1:2420:1185#CTTGTA/1
g]gggfffffcggeggegggggfgecfdfefceeededeg_afcabeb]dggddff`ba`ad]ab\aaaaWTT
```

Figure 1.6

Example of an Illumina FASTQ file (based on the Sanger FASTQ file format). @ indicates the title and optional description, followed by the sequence line(s), + indicates optional repeat of the title line, followed by the quality line(s).

The quality scores that accompany NGS sequences within the FASTQ file format allow a certain level of data *filtering*. However, due to the degree of errors introduced throughout the experimental process, and the incomplete nature of the sequence information, several steps of filtering, alignment and error correction are required. Beerenwinkle *et al.* (2001) provide a good review of these three processes, which are also discussed further in Chapters 3 and 5.

The production of millions of NGS reads, with the accompanying need for more substantial and complex quality control, alignment strategies and computational analysis, has challenged the infrastructure of existing information technology systems. Advances in bioinformatics continue to be made. In the last three years alone, the number of commercially and publically available software packages, available for the analysis of NGS data, has increased from those detailed within Table 1.2 to over a hundred programmes and web based services. Skew remains towards alignment, assembly and mapping software compared to single nucleotide polymorphism (SNP) calling/discovery, error correction and filtering. However, Oxford Journals, ‘Bioinformatics’, maintains a virtual issue, ‘Bioinformatics for Next Generation Sequencing’, which is continually updated with the latest papers published on the tools and algorithms relevant to next-generation sequencing applications. The increasing number of publications relevant to general data-handling ‘Pipeline’ and ‘Variant detection’ is testament to the sustained development in these areas. Nevertheless, further investment into the improvement of these analysis systems is necessary if they are to keep pace with the continuing expansion of NGS technologies.

Table 1.2 Next-generation sequencing software available in the commercial and public domain in 2009

Name	Bioinformatics method	Operating system	Language
CLCbio Genomics Workbench[1]	Integrated solutions	Windows, MacOSX, Linux	
Galaxy		Job webportal	
Genomatix		N/A	
JMP Genomics		N/A	
NextGENe[1]		Windows, MacOSX	
Seqman Genome Analyser		Windows, MacOSX	
Shore		POSIX	
SlimSearch		N/A	
BFAST	Alignment	N/A	
Bowtie		N/A	
BWA			C++
ELAND		N/A	
Exonerate		POSIX	
GenomeMapper		POSIX	
GMAP			C/Pearl
gnumap			C
MAQ			C++
MOSAIK		Windows, Linux, MacOSX	
MrFAST and MrsFAST			C
MUMmer		POSIX	
Novocraft		Linux, MacOSX	
PASS		Windows, Linux	
RMAP		POSIX	
SeqMap		Most OS’s	
SHRiMP		POSIX	
Slider		N/A	
SOAP			C++
SSAHA			C++
SOCS		N/A	
SWIFT		N/A	
SXOligoSearch		OS independent	
Vmatch		POSIX	
Zoom		N/A	
ssahaSNP	SNP/Indel Discovery	Linux, Solaris, MacOSX	
PolyBayesShort		Linux	
PyroBayes		N/A	

[1] Includes SNP detection; SNP, single nucleotide polymorphism; N/A, not available; OS, operating system

The NCBI Sequence Read Archive (SRA) provides an online repository for raw data from NGS platforms including Roche 454, Illumina GA, Life’s SOLid, Helicos Heliscope, Complete Genomics, and PacBio’s SMRT.

1.6 Objectives of PhD thesis

The evolution of FMDV population diversity, a result of the inherent low fidelity of viral RNA-dependent RNA polymerase, can be observed at multiple spatial and temporal scales. Inherent to each of these scales are key processes that shape the dynamics of virus evolution. These processes are driven by selection or more random events such as transmission and bottlenecking. Understanding the impacts of such processes on the characteristics of viral population diversity, at the finer-scales (*within* host), may ultimately inform our knowledge of the acquisition and fixation of nucleotide changes within consensus sequences at broader scales (*between* hosts and above). Therefore, the key questions posed in this thesis are,

i) Can we use ultra-deep sequence data provided by NGS to characterise viral diversity below the level of the consensus?

ii) How are nucleotide changes fixed in the consensus sequence?

iii) How related are viral populations at the intra and inter-host scale?

The overall aim of this PhD project is to further dissect the genetic evolution of FMDV at different spatial and temporal scales, both *in vivo* and *in vitro,* using a novel sequencing technology. Figure 1.7 provides a basic schematic of these scales. The ultimate goal is to use such data towards constructing more representative and unified models of RNA virus mutation, evolution and transmission.

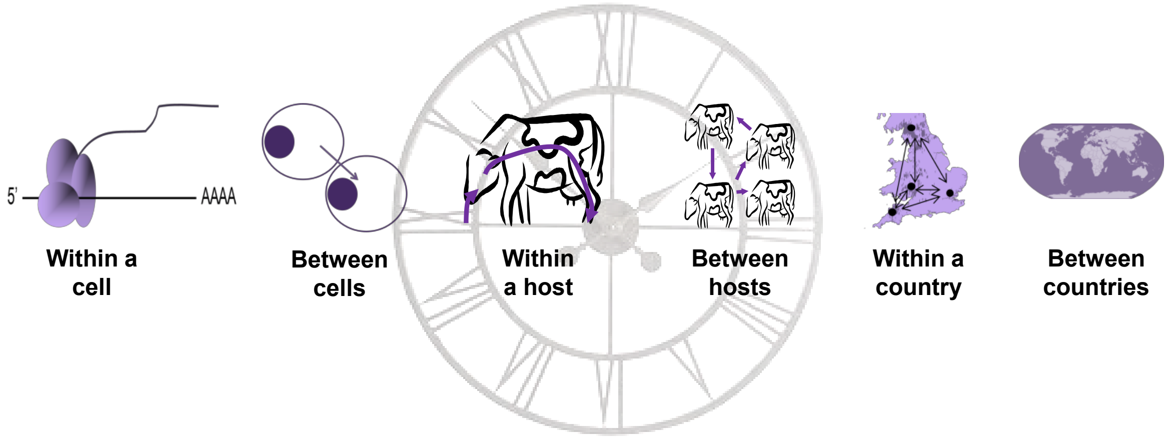


Figure 1.7

Schematic of the multiple scales of FMDV evolution. The clock face is indicative of time.

Chapter 2 describes the genetic evolution of FMDV at the epidemic scale, through the sequence analysis of samples taken during an outbreak of FMD in the United Kingdom (UK) in 1967. The objective was to use full genome consensus sequencing to clarify the relatedness of intra-epidemic samples and samples from additional contemporary FMD outbreaks. Chapter 3 describes a pilot study designed to test the appropriateness of NGS technology for the resequencing of FMDV and the analysis of viral population diversity at the within-host scale. This chapter also includes the development of a novel pipeline for the filtering, alignment and error correction of NGS data produced by the Illunima GA II platform. An assessment was made as to whether such a technique could be used to quantify genuine viral mutations taking into account the impact of artefactual mutations. Following the positive assessment of the Illumina GA II platform (Chapter 3), further optimization of the experimental protocol for producing FMDV samples for NGS analysis was performed and is described in Chapter 4. This chapter also includes details of a clonal control experiment designed to further quantify the contribution made by artefactual mutation to FMDV sequence variability. This study aimed to produce a more accurate mutation frequency threshold above which there can be relative confidence in the identification of genuine mutations.

Using the aforementioned optimized protocol, Chapter 5 describes an investigation of FMDV population dynamics at the within-host scale during serial transmission in bovine hosts. This study hypothesized that both inter and intra-host bottleneck size may be inferred by variations in viral population diversity, characterised at the ultra-deep level using NGS. Chapter 6 describes a pilot *in vitro* study designed to further investigate the impact of bottleneck size on the acquisition and fixation of mutations within FMDV populations, characterised at the ultra-deep level using NGS. The findings of this thesis are concluded and discussed in Chapter 7.

Chapter 2

Reconstructing the origin of the UK 1967-68 foot-and-mouth disease outbreak

The work in this chapter is to be submitted to Infection, Genetics and Evolution:

Caroline F. Wright[1], Nick J. Knowles[1], Antonello DiNardo[1], David J. Paton[1], Daniel T. Haydon[2], Donald P. King[1].

[1] Institute for Animal Health, Ash Road, Pirbright, GU24 0NF, United Kingdom

[2] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom.

Complete genome sequences generated in this study have been submitted to GenBank

2.1 Summary

A large epidemic of foot-and-mouth disease (FMD) occurred in the United Kingdom (UK) over a seven-month period in Northwest England from late 1967 to the summer of 1968, following smaller outbreaks in 1967 in Hampshire and Warwickshire. The causative agent of all three events was identified as foot-and-mouth disease virus (FMDV) serotype O and the source of the largest one was attributed to infected bone in lamb products imported from Argentina. However, available diagnostic tools were unable to entirely rule out connections with the earlier Warwickshire UK FMD outbreak and questions remained about other potential sources from Europe. The aim of this study was to apply molecular sequencing to answer these questions about a historic UK FMD outbreak and, by doing so, dissect this event at a previously unobtainable depth of resolution. VP1 region and full genome (FG) sequences were recovered directly from clinical epithelium samples (n=13) or cell culture isolates (n=6), from contemporary UK, European and South American outbreaks. Analysis of the VP1 sequences provided evidence for at least three separate incursions of FMDV into the UK, one of which caused the main 1967/68 epidemic. Analysis of FG sequences from the main 1967/68 outbreak (n=10) revealed nucleotide substitutions at 94 genomic sites. Viral FG sequences have provided further evidence for a linear accumulation of nucleotide substitutions (rate = 8.7×10^{-3} substitutions per site per year whilst continually replicating inside a host). However, where this linear relationship was absent, evidence is provided for the virus having spent periods of time outside a host and therefore not replicating or incurring genomic mutations. Genetic scale clarification of past disease outbreak dynamics will further add to the knowledge and understanding from which to base future outbreak control strategies.

2.2 Introduction

The severe productivity losses, associated with foot-and-mouth disease (FMD), are a result of its debilitating effects on cloven-hoofed animals, characterized by vesicular lesions of the feet, tongue, snout and teats as well as fever and lameness, rather than high mortality rates (Arzt, Juleff et al. 2011). The etiological agent, FMDV, is a member of the genus *Aphthovirus* in the family *Picornaviridae*. Of the seven serotypes of FMD (A, O, C, Asia 1, SAT 1-3) serotype O is the most prevalent and was classically divided into eleven antigenic subtypes (O1 – O11) (Davie, 1964). High mutation rates of FMDV RNA polymerase (10^{-3} – 10^{-5} per nt per transcription cycle), coupled with large population sizes and a rapid rate of replication, results in the fast evolution of this virus within infected hosts (Domingo, Escarmis et al. 2003). Although predominantly spread by direct or indirect contact with infected animals, their secretions or associated products, FMDV can travel over extensive distances by air- and windborne routes or via fomites, causing incursions in areas that were previously free from the disease (Arzt, Juleff et al. 2011).

FMD has not been endemic in the UK since 1884, but outbreaks occurred sporadically until the 1960's, with virus introductions attributed to spread from Europe and South America. A major epidemic, starting in October 1967 in Shropshire, caused outbreaks on 2,346 farms, 18 of which were infected twice. The epidemic was controlled after seven months by a stamping out policy, combined with movement restrictions and the last outbreak was reported on the 4th of June 1968. Although outbreaks occurred over the North-West Midlands and North Wales, Lancashire & Westmorland, Derbyshire, South-West Midlands and South Wales and the East Midlands, the vast majority of affected farms (2,228) were located in the North-West Midlands, which had the highest density of dairy cattle in the country at the time. Apart from two years (1963 and 1964), FMD had sporadically been present in the UK during the thirteen years preceding this epidemic, including three sets of outbreaks affecting Northumberland (32 outbreaks over approximately three months, 1966), Hampshire (29 outbreaks over approximately one month, January, 1967) (Sellers and Forman 1973) and Warwickshire (5 outbreaks over 3 days, September, 1967).

The UK Government commissioned the Northumberland report (Anon April 1969) that made four principal conclusions regarding the primary source and initial spread of the epidemic. Firstly, that it was not possible to categorically establish the origin of the 1967/68 epidemic. Secondly, that there was a basis for inference that the most probable source of the epidemic was infected meat from South America. Thirdly, that, as the Ministry of Agriculture had also attributed an earlier outbreak in 1967 in Warwickshire to South American meat, this event remained as a potential link to the main 1967/68 epidemic. Finally, that it was difficult to explain the epidemic's rapid development and extension other than by accepting that a number of foci were established almost simultaneously.

Since 1968, FMDV diagnostic tools have advanced from serologically based tests, such as the complement fixation and virus neutralization test, towards molecular sequencing. Determination of the genetic sequence for one of three surface exposed capsid proteins of FMDV (VP1) and the FG, has not only enabled the global tracing of FMD transmission but provided the resolution at which disease spread can be monitored at the single outbreak scale (Samuel and Knowles 2001; Knowles and Samuel 2003; Cottam, Wadsworth et al. 2008a; Abdul-Hamid, Firat-Sarac et al. 2011; Kasambula, Belsham et al. 2011; Valdazo-Gonzalez, Knowles et al. 2011). As well as providing the genetic profile of an outbreak, FG sequencing of FMDV can yield insights into the processes shaping this profile, for example, by testing for the presence of a molecular clock in terms of nucleotide (nt) substitution rate. Previous studies, at the outbreak scale, have demonstrated that nt changes, from the earliest FMDV sample, accrue in a linear 'clock-like' way with time (Villaverde, Martinez et al. 1991b; Elena, Gonzalez-Candelas et al. 1992; Haydon, Samuel et al. 2001; Cottam, Haydon et al. 2006; Valdazo-Gonzalez, Knowles et al. 2011), resulting from continuous viral replication within susceptible hosts. The aim of this study was to apply current molecular sequencing tools to further characterize the 1967/68 FMD outbreak, along with other contemporary events, and, by doing so, clarify some of the issues highlighted by the Northumberland Report, expanding the knowledge fed into future disease control programmes. It was not the aim of this study to reconstruct the viral transmission pathways of this historic outbreak.

2.3 Methods

2.3.1 Samples

This study accessed archived vesicular epithelium samples (n=13) from the World Reference Laboratory for FMD (WRLFMD), Institute for Animal Health, Pirbright, collection which had been stored at -20^{o}C in 0.04 phosphate buffer (M25; disodium hydrogen phosphate, potassium dihydrogen phosphate, pH 7.5) and 50% (vol/vol) glycerol. The 13 clinical samples were collected from early in the Northumberland outbreak (OB/North), the beginning of the Hampshire outbreak (OB/Hants), the beginning of the Warwickshire outbreak (OB/Warks), and a total of ten collected approximately every month from the beginning (isolate from the index case on Bryn Farm not available), to the end of the 1967/68 outbreak. As the index case for the 1967/68 outbreak occurred in Shropshire, for the sake of this study, this outbreak will be geographically known as the Shropshire outbreak (OB/Salop). All 13 samples had previously been found FMDV positive by the complement fixation test on original submission to the WRLFMD at the time of these outbreaks. All clinical samples from which FG sequences were determined (n=12) are detailed in Table 2.1

Table 2.1 Details of epithelium samples from which FMDV full genome sequences were derived

Within outbreak sample ID	Outbreak (OB/)	Outbreak duration	Epi. sample type	Sample collection date (day/mo/yr)	County	World Reference Laboratory no.(OBFS)	Approximate lesion age (hrs)	Sequence length (nt)	Total no. of nts sequenced	Average times coverage of each base	GenBank accession No.
-	Hants	6.1.67 – 3.2.67	Bovine tongue	6.1.67	Hants.	1810A	N/A	8177	26393	3.23	JX869177
-	Warks	8.9.67 – 11.9.67	Bovine DP	8.9.67	Warks.	1836	6	8177	45434	5.48	JX869178
A	Salop	21.10.67 – 6.6.68	Bovine foot	31.10.67	Salop.	1848	6	8176	35045	4.29	JX869179
B			Bovine foot	14.11.67	Salop.	1889	12	8176	30647	3.75	JX869180
C			Bovine tongue	8.12.67	Salop.	1950	6-8	8176	25659	3.14	JX869181
D			Bovine tongue	6.1.68	Heref.	11/68	12	8176	30485	3.73	JX869182
E			Bovine tongue	10.2.68	Staffs.	41/68	8	8173	50125	6.13	JX869183
F			Bovine toot	13.2.68	Ches.	45/68	288	8176	87275	10.68	JX869184
G			Bovine tongue	22.3.68	Ches.	63/68	8	8176	51187	6.26	JX869185
H			Bovine tongue	7.4.68	Ches.	69/68	18	8176	48464	5.84	JX869186
I			Bovine tongue	5.5.68	Salop.	86/68	24	8176	54360	6.61	JX869187
J			Ovine	4.6.68	Salop.	89/68	N/A	8176	44809	5.46	JX869188

N/A not available

DP Dental Pad

Additionally, FMDV VP1 sequences from contemporary outbreaks in the UK, Europe and South America were analysed (n=12), and comprised of those determined from the single epithelium sample from OB/North and cell culture isolates (n=6), as well as those available from GenBank (n=5), as detailed in Table 2.2. Finally, VP1 sequences determined from isolates from the UK FMD outbreak in 1981 (n=3) were also included in VP1 sequence analysis for comparative purposes (Table 2.2).

Table 2.2 Details of cell culture isolates and one epithelium sample from which FMDV VP1 sequences were derived

Isolate name	Original material collection date (dd/mm/yyyy)	Location, Country	GenBank accession No.
O_1/Lombardy/ITL/46 [2]	1946	Lombardy, Italy	M58601
O_2/Flanders/BEL/47 [1]	1947	Flanders, Belgium	JX869189
O_2/Brescia/ITL/47 [2]	1947	Brescia, Italy	AY593826
O/M11/MEX/52 [1]	1952	Mexico	JX869190
O_1/Campos/BRA/58 [2]	1958	Campos, Brazil	AY593819
O/GRE/1/63 [1]	1963	Greece	JX869191
O_1/Brugge/BEL/63 [2]	1963	Bruges, Belgium	EU553836
O_1/Lausanne/SWI/65 [2]	1965	Lausanne, Switzerland	M15974
O_1/Argentina/c.65 [2]	c. 1965	Argentina	AY593814
O/UKG/66 (1782) [1,3]	22.7.66	Northumberland, UK	Submitted
O_1/Kaufbeuren/FRG/66 [2]	1966	Kaufbeuren, Germany	X00871
O_1/BFS 1860/UK/67 (OBFS18) [2]	01/11/1967	Wrexham, Cheshire, UK	AY593815
O/FRA/1/81 [1]	Mar-1981	Côtes-du-Nord, France	JX869192
O/UKG/15/81 [1]	19/03/1981	Jersey, Channel Islands, UK	JX869193
O/UKG/16/81 [1]	21/03/1981	Isle of Wight, UK.	JX869194

[1] VP1 sequence determined for the current analysis
[2] VP1 sequence obtained from GenBank
[3] VP1 sequence derived from a clinical epithelium sample

2.3.2 Sample preparation and RNA extraction

The suspension preparation protocol for epithelium samples was as described by (Cottam, Haydon et al. 2006). Briefly, a 10% tissue suspension was prepared with a pestle and mortar in a class II safety cabinet using 0.04 M phosphate buffer and approximately 1.5 g of each the 13 vesicular epithelium samples. The suspension was then centrifuged for 10 min at 3500 x g at room temperature and the supernatant removed to be stored at – 80^{o}C until tested. Total RNA was extracted (TRIzol, Invitrogen, Paisley, UK) from all epithelium samples before reverse transcription and amplification by PCR. Total RNA was extracted from 460 µL cell culture supernatant by using RNeasy kits (Qiagen Ltd., Crawley, West Sussex, UK), according to the manufacturer's instructions, resuspended in 50 µL nuclease-free water and stored at -20°C.

2.3.3 RT-PCR

The following reverse transcription method was modified from that used in (Cottam, Haydon et al. 2006). Briefly, extracted RNA (15 µl) was added to 3 µl 10 mM oligo-dT primer UKFMD/Rev6, 3 µl 10 mM deoxynucleoside triphosphate mix and then incubated at 70^{o}C for 3 min followed by 4^{o}C for 3 min. Nineteen microlitres of freshly prepared RT mix (8 µl 5x RT buffer [Invitrogen], 2 µl 0.1 mM dithiothreitol, 2 µl RNase OUT [Invitrogen], 5 µl nuclease-free water) was added to the sample followed by 2 µl of an enzyme with high specificity (Superscript III reverse transcriptase, Invitrogen). The sample was then incubated at 45^{o}C for 60 min, after which the cDNA synthesis reaction was terminated by incubation at 85^{o}C for 5 min. The cDNA was then cleaned using QIAquick PCR purification kits (QIAGEN), eluted in 40 µl of nuclease-free water before storage at -20^{o}C.
VP1 region amplification was achieved as per the method previously described using two different primer sets (Knowles, Samuel et al. 2005; Abdul-Hamid, Hussein et al. 2011). The protocol used for FG PCR amplification was modified from that previously described (Cottam, Wadsworth et al. 2008a). Briefly, twenty-three overlapping PCR fragments covering the FMDV genome were amplified by adding 3 µl of each cDNA to 47 µl of master mix (5 µl 10x buffer, 2 µl $MgSO_4$, 1 µl 10 mM deoxynucleoside triphosphate mix, 1 µl 10 mM forward primer, 1 µl 10 mM reverse primer, 0,25 µl Platinum Taq DNA Polymerase Hi-Fidelity [Invitrogen], 37 µl nuclease-free water). Details of the RT and PCR primers used are as previously

published (Cottam, Wadsworth et al. 2008a). Samples were run on a PCR program cycle of initial denaturation at 94^{o}C for 5 min and then 39 cycles of 94^{o}C for 30 s, 55^{o}C for 30 s, and 72^{o}C for 4 min, ending with incubation at 72^{o}C for 7 min. PCR products were cleaned up using QIAquick PCR purification kits (QIAGEN), eluting in 50 μl of nuclease-free water. In order to visualize amplified DNA to check quality and specificity of the product, 3 μl was run on a 1% agarose gel at 100 V for 35 min alongside a quantitative ladder (GeneRuler 100bp ladderPlus, MBI Fermentas).

2.3.4 VP1 and full genome sequencing

The protocols used for VP1 (Abdul-Hamid, Hussein et al. 2011) and FG sequencing (Cottam, Wadsworth et al. 2008a) have been described. However, sequencing reactions were performed using the Applied Biosystems BigDye Terminator V3.1 Cycle Sequencing Kit and an ABI 3730 genetic analyser.

2.3.5 Sequence analysis

The raw sequence data from all epithelium samples and cell culture isolates (n=7 as detailed in Table 2) were assembled using SeqMan Pro™ 10.0.1 (DNASTAR, Madison, WI) followed by BioEdit v7.1.3.0 (Hall 1999) for all subsequent sequence manipulations and nt difference counts between sequences. The evolutionary history of all VP1 sequences was inferred using the Neighbor-joining method implemented in MEGA5 (Tamura, Peterson et al. 2011), where a bootstrap consensus tree was inferred from 1000 replicates. Evolutionary distances (branch length) were computed using the Tamura-Nei method and were in units of the number of base substitutions per site. Rate variations among sites were modelled with a gamma distribution (shape parameter = 4). All positions containing gaps and missing data were eliminated.

The genealogical network underlying the relationship between all 12 FG sequences examined was computed based on statistical parsimony implemented in the software package TCS 1.21 (Clement, Posada et al. 2000). In order to include a candidate most likely common ancestor in the TCS analysis, a FASTA search of all publically available FG sequences was completed using the FG sequences for OB/Salop and the top six hits included in the TCS analysis. In addition, by correlating the position of each node tip in the TCS analysis to a timeline according to when each sample was collected, branch length could be used to provide an indication of nt substitution rate for each sample.

In order to compare the rate of nt substitution observed during OB/Salop and a more recent UK FMD outbreak of equivalent size and duration, a molecular clock was fitted to the first five sequences from OB/Salop (A – E) and 42 previously analysed sequences collected during the UK 2001 FMD outbreak (Cottam, Haydon et al. 2006; Cottam, Thebaud et al. 2008; Konig, Cottam et al. 2009). Before performing the phylogenetic reconstruction, jModelTest 0.1.1 analysis was employed for determining the best-fitting nucleotide substitution model by the Bayesian information criterion (BIC). Markov chain Monte Carlo techniques were implemented in the software package BEAST 1.7.2 (Bayesian evolutionary analysis sampling trees) (Drummond and Rambaut 2007) where a strict molecular clock with constant rate, no prior assumption of population size (Bayesian Skyline plot), and the TN93+Γ4 (Tamura and Nei 1993) model of base substitution with empirical base frequencies were assumed. Once extracted from BEAST, the difference in clock rate observed during OB/Salop and the 2001 UK outbreak was tested using the Student's t-test with Welch's approximation (Welch 1947) in R 2.15.1. A full timed phylogeny using BEAST was not appropriate for the number of sequences available.

Recombination analysis was performed on all 12 FG sequences using Simplot 3.5.1 software (Lole, Bollinger et al. 1999). Pairwise genetic similarity plots were generated using the Kimura 2-parameter model to calculate evolutionary distance. Bootstrap replicates were also performed to infer statistical significance. The genetic similarity plots between each query and reference sequence were plotted in a moving window along the alignment.

2.4 Results

2.4.1 VP1 sequence analysis

VP1 sequence analysis was used to define the genetic relationships between OB/Salop, previous FMD outbreaks in the UK, as well as contemporary outbreaks in Europe and South America. Figure 2.1 clearly shows that sequences determined from the four UK FMD outbreaks (OB/Salop, OB/Warks, OB/Hants and OB/North) were on three separate phylogenetic lineages (lineages were supported by bootstrap values >60%). OB/Hants and /Warks samples are shown to be more closely related to each other then to any other sample analysed here.

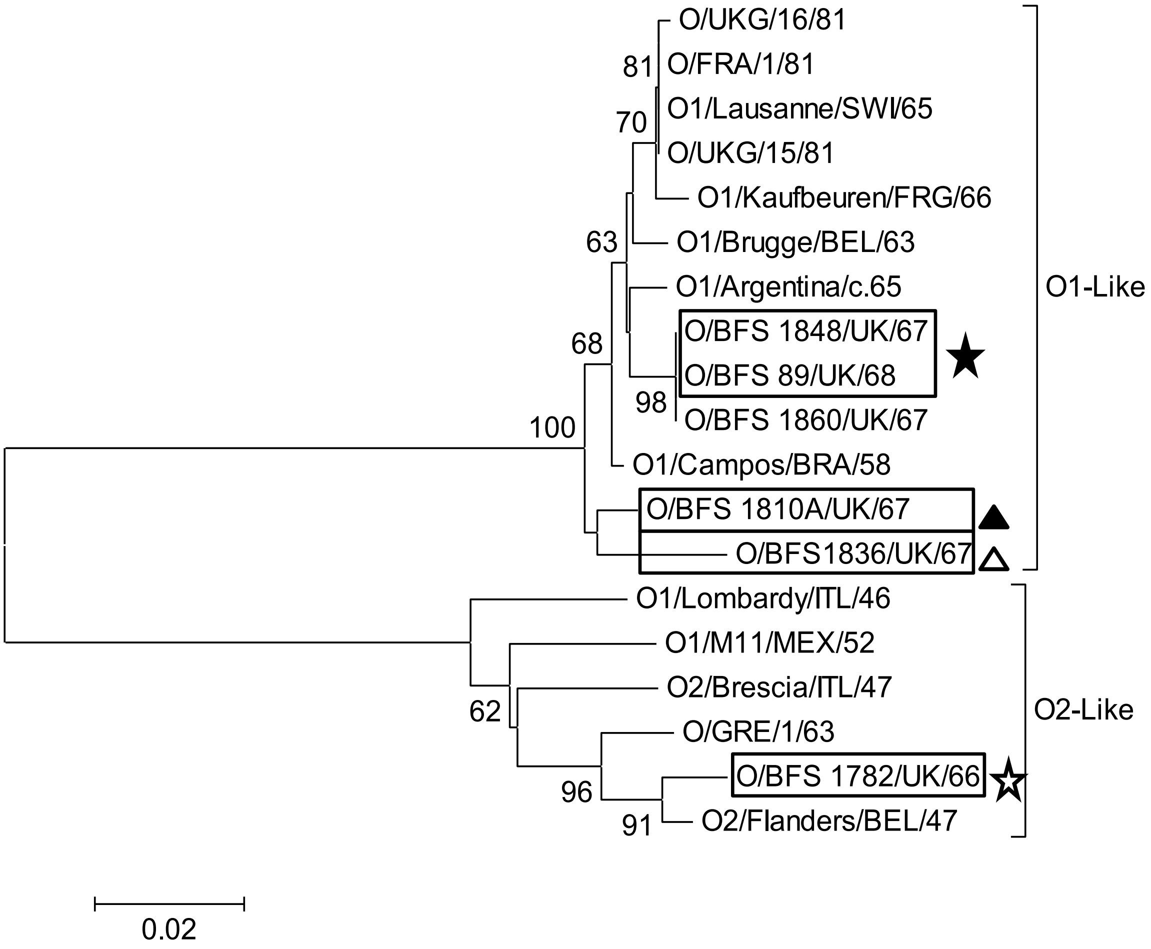


Figure 2.1

Un-rooted Neighbor-joining tree showing the relationships between 20 selected complete FMDV VP1 sequences. The percentage of replicate trees in which associated sequences clustered together in the bootstrap test (1000 replicates) is shown next to the branches (>60%). Sequences from the four UK FMD outbreaks analysed are highlighted within black boxes. White star, OB/North; black star, OB/Salop; white triangle OB/Warks; black triangle, OB/Hants.

Most distant was the OB/North sample, which was more like the O2 than O1 antigenic subtype and demonstrated relatedness to earlier outbreaks in Europe (Belgium and Greece). Furthermore, the selected samples from OB/Salop were shown to be more closely related to South American isolates compared to those collected from previous UK FMD outbreaks.

Interestingly, no nt differences were observed within the VP1 sequence between samples taken at the beginning and the end of OB/Salop, although the samples were taken 217 days apart (a total of 7 nt differences were observed when VP1 sequences were compared from all ten OB/Salop samples [sequences not included in phylogentic analysis]). For example, taking the rate of fixation of nt substitutions within the VP1 region of 6×10^{-3} per nt per year, during a defined period of acute disease (Villaverde, Martinez et al. 1991a), this would equate to approximately 2 substitutions during a 217 day long period, assuming a VP1 sequence length of 639 nt. The branch containing VP1 sequences from early and late samples of OB/Salop also contained the VP1 sequence of a previously determined isolate from this outbreak, GenBank sequence AY593815 (see section 3.3 for further details and FG sequence analysis of this isolate).

2.4.2 Full genome sequence analysis between outbreaks

The assembled FG sequences of the 12 epithelium samples analysed were unique and ranged in length between 8183 nucleotides (nts) (isolate E, OB/Salop), 8187 nts (single isolate for both OB/Hants and /Warks), and 8186 for the remaining nine samples. Primer derived sequences (<0.4% of the total genome length), were omitted at the 3' and 5'ends of the genome (15 and 8-nts respectively), as well as at the 3' and 5' ends of the 10-nt long artificial poly(C) tract (5 and 4-nts respectively). A 10-nt long poly(A) tract was also included at the 3' end of the genome. No ambiguities were found.

Although having only occurred one and nine months previously, the FG sequence from OB/Warks and /Hants demonstrated large genetic differences when compared to the FG sequences from OB/Salop (232-255 nts for OB/Warks and 102-125 nts for OB/Hants). Of the samples analysed, OB/Warks and /Hants showed the closest genetic relationship (difference of 214 nts). In order to better understand the genetic relationship between these three UK FMD outbreaks, recombination analysis was performed on all FG sequences. Figure 2.2 (A) shows a similarity plot analysis where regions of similarity between the query (O1/Campos/BRA/58[AY593819]) and reference sequences (OB/Warks [black trace], OB/Hants and all ten from OB/Salop [grey trace]) increased or decreased, approximately identifying recombination breakpoints (clustering of the query to reference sequences was supported by 1000 bootstrap replicates). In this instance, within the second half of the genome, from the 5' end of P2 (2C) to the P3 region (3D), there are three main areas where the OB/Warks curve drops (nt positions 4600-4900; 5800-6200 and 6600-7000), indicating diverse regions of genome compared to all other reference sequences. Consequently, between a 300 and 400 nt long section from the OB/Warks sequence, spanning each of these areas, was submitted to a FASTA similarity search returning the result of A26/Argentina/ARG/66 [AY593770] (97-98% similarity). Figure 2.2 (B) shows the subsequent similarity plot (specifications as described above) of the OB/Warks sequence queried against all other determined FG sequences (grey traces) plus A26/Argentina (black trace). Regions of highest similarity between OB/Warks and A26/Argentina are clearly correlated with regions of lowest similarity between all other reference sequences and the OB/Warks sequence. Clustering of the OB/Warks FG sequence and A26/Argentina was supported by bootscanning, with the parental threshold set to 70% (plot not shown).

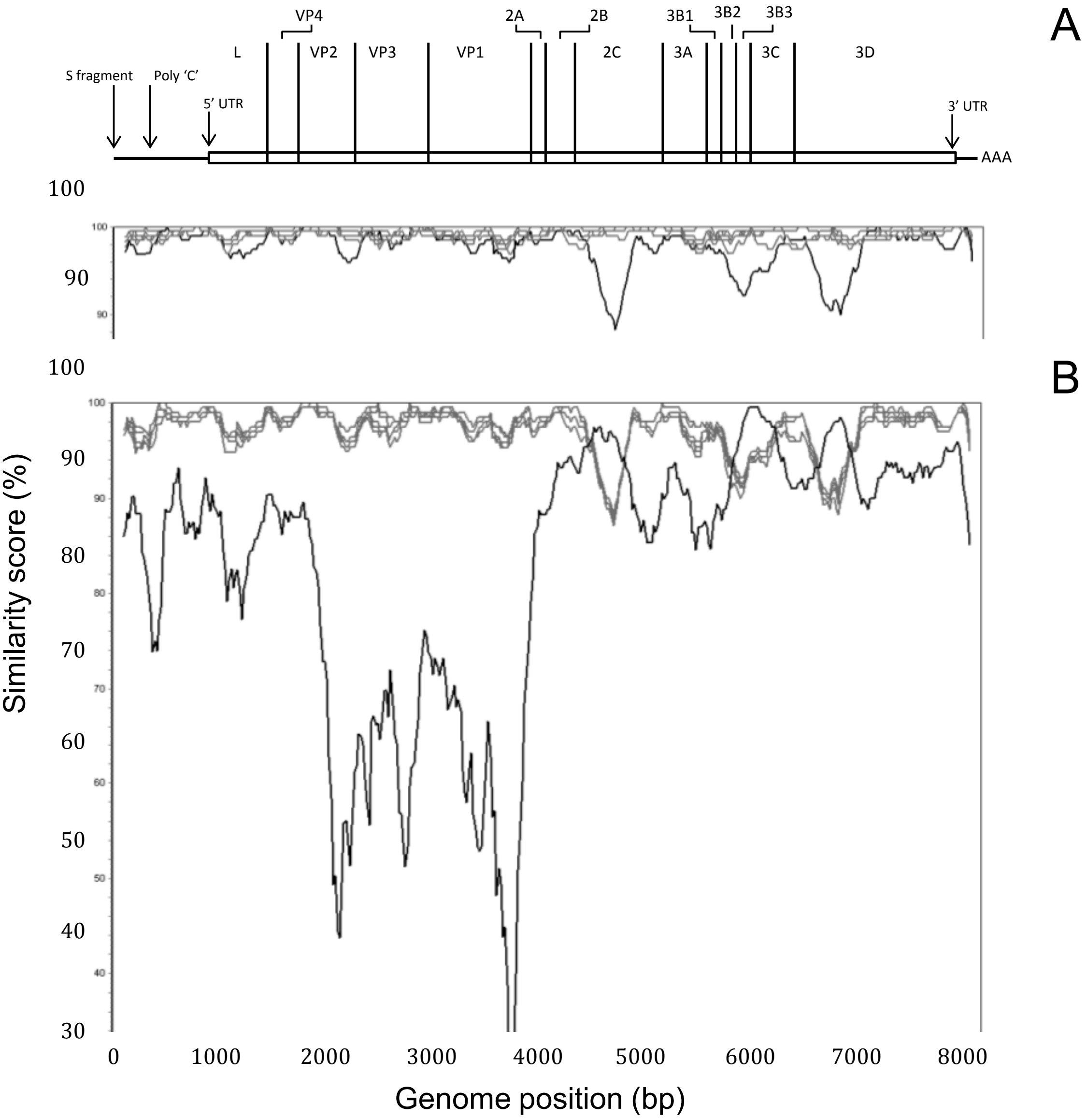


Figure 2.2

Recombination analysis using Simplot 3.5.1. **A** all FG sequences from OB/Salop (gray trace), OB/Hants (gray trace) and /Warks (black trace) queried against O1/Campos/BRA/58/(AY593819). **B** all previous FG sequences analysed (gray traces) plus that of A26/Argentina/(AY593770) (black trace) queried against the FG sequence of the OB/Warks sample. Analysis performed, Kimura (2-parameter) in a sliding window size of 200 bp moving in steps of 20 bps along the alignment. The pairwise similarity values were plotted at the midpoint of the 200 bp window. At the top of the figure, a fully annotated FMDV FG sequence is represented.

2.4.3 Full genome sequence analysis within an outbreak

Using the earliest sequence determined from OB/Salop (sample A) as a reference, the remaining nine FG sequences of this outbreak were found to contain 94 point mutations across the genome. Within the coding region, these nt changes were mainly synonymous (n=65), the majority of which were transitions (n=62), compared to non-synonymous mutations (n= 17). Eleven mutations were found within the 5’ UTR and one was found within the 3’ UTR. In addition, a single deletion of three nts (ACC) was found within the 5’ UTR (sample E). The single stem-loop secondary structure previously predicted for the FMDV S-fragment RNA sequence (Newton, Carroll et al. 1985; Clarke, Brown et al. 1987), was tested using RNAStructure v5.3 (Mathews 2006), and found to be maintained, with small conformational adjustments, despite this deletion. While the loop apex of this stem-loop structure maintained an A-C-C-T-C conformation within the sequence for sample E, the adjacent stem was elongated by two base pairs and the following two loops were smaller by one and six nts respectively (data not shown). However, the nt composition of the loop apex was not conserved on examination of approximately 100 FMDV S-fragment sequences (representing all seven FMDV serotypes).

The maximum genetic difference seen between samples studied from OB/Salop was of 42 nts (E – G), with a minimum of two nts (I – J). Figure 2.3 provides a map of the geographical collection points of all OB/Salop samples.

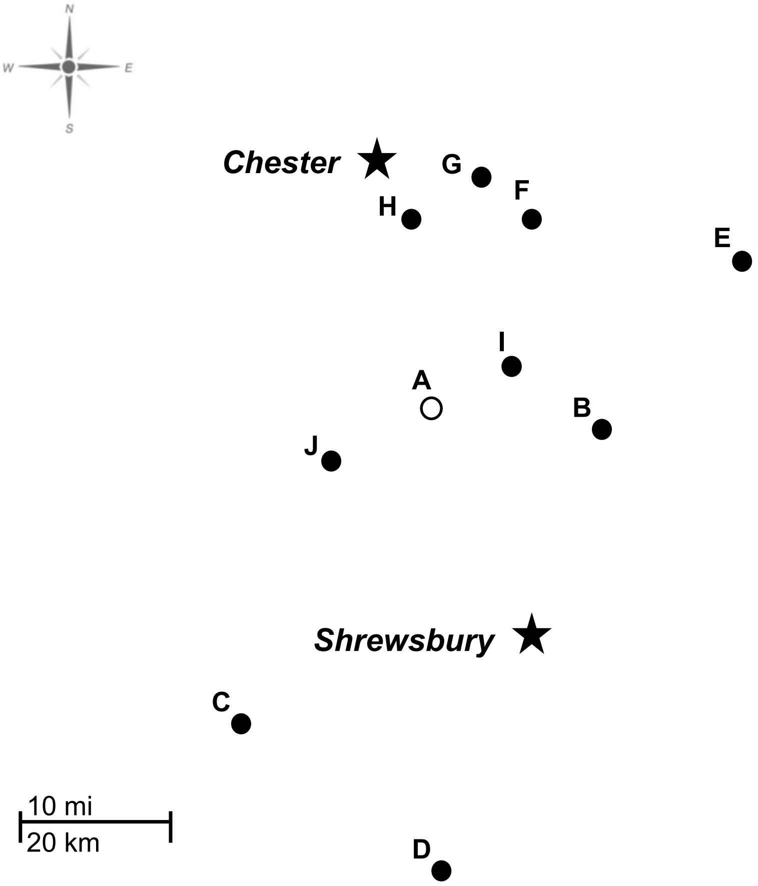


Figure 2.3

Map of the geographical collection point for all ten samples from OB/Salop. Open circle represents the outbreak with the index case.

The genealogy network of all ten samples from this outbreak, in relation to OB/Hants and /Warks samples plus the most likely common ancestor using O1/Campos/BRA/58[AY593819], implemented by the software package TCS (Clement, Posada et al. 2000) is shown in Figure 2.4. The most likely common ancestor, as estimated by TCS analysis, was maintained after addition of all publically available FG sequences for the 1967/68 outbreak (EU448370, EU448369, EU448368, AY593816, AY593815 [data not including in Figure 2.4]). All publically available FG sequences were from cell culture passaged viruses (passage numbers unknown), originally derived from a single epithelium sample (collected on the 1st of November 1967 in Wrexham), which was designated serotype O_1 British Field Strain 1860 (O_1BFS1860) and used as the type virus for this outbreak. Although additional nt substitutions may have been introduced into the viral genome of cell culture passaged isolates, these would not change the genetic relationship of the *in vivo* samples analysed here, having been subject to very different selective pressures *in vitro.*

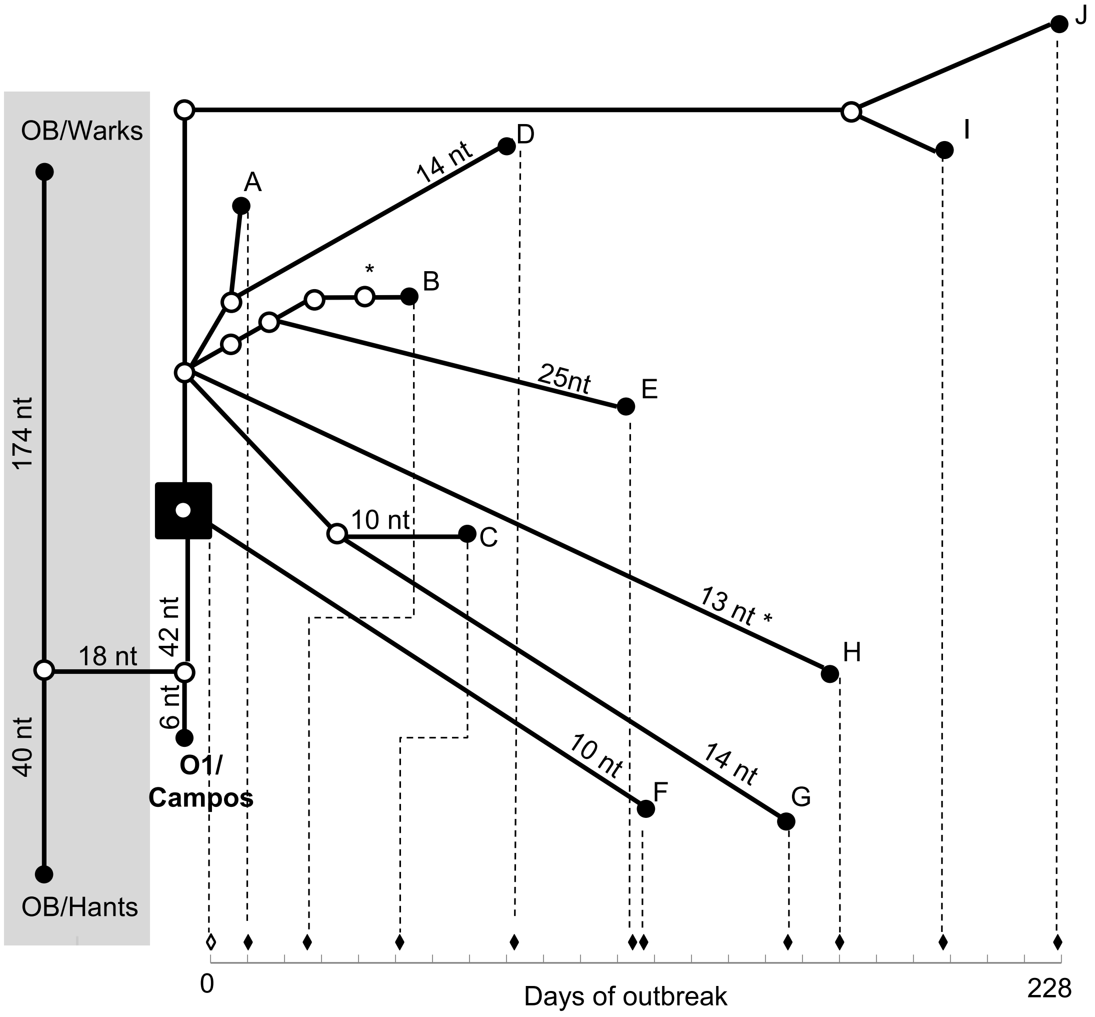


Figure 2.4

Statistical parsimony analysis by TCS. Ten FG sequences (A – J) derived from clinical epithelium samples collected during OB/Salop are shown in relation to those from OB/Hants and /Warks (past UK outbreaks highlighted in gray box) as well as the most similar South American sequence (O1/Campos/BRA/58[AY593819]). The estimated most likely common ancestor is highlighted within a black box. Unless otherwise stated, each connecting branch line represents a single nt substitution, with each dot representing a putative ancestor virus. Node tips for all OB/Salop sequences are correlated to an outbreak timeline so that branch length is proportional to time. (*) nt substitution occurred twice on independent branches

Longer branch lengths in Figure 2.4 indicate a relatively slow accumulation of nt substitutions (observed in samples F-J), whereas shorter branch lengths indicate a relatively fast accumulation of nt substitutions (observed in samples A-E). Figure 2.5 shows the accumulation of nt substitutions from the most likely common ancestor (estimated by TCS analysis), for the ten OB/Salop samples against time. Note, for the sake of this analysis, the estimated most likely common ancestor was dated the 21st of October 1967, according to the date on which FMD symptoms were first reported on Bryn Farm (Anon April 1969). Plotting the linear regression analysis for samples A-E with corresponding 95% confidence intervals, clearly demonstrated samples F-J do not follow the same linear accumulation of nt substitutions (Figure 2.5). A subsequent plot of nt substitution number for samples A-E from OB/Salop against equivalent numbers from the UK 2001 samples, against time, demonstrated that the two regression line slopes were almost identical (0.24 and 0.21 respectively) (data not shown). No statistical difference was observed between the BEAST estimated molecular clock rate for samples A-E from OB/Salop (8.74×10^{-3}, 95% CI: 8.73×10^{-3} to 8.75×10^{-3}) and that estimated for the 42 samples from the UK 2001 outbreak (8.89×10^{-3}, 95% CI: 8.88×10^{-3} to 8.91×10^{-3}), when computed for the Student t-test (t=15.142, p=0.000). It should be noted that although sample (F) from OB/Salop came from a lesion estimated at 12 days old at collection, this does not significantly impact the calculated nt substitution rate (9×10^{-2} according to date of collection and 1×10^{-1} if calculated with a date 12 days earlier).

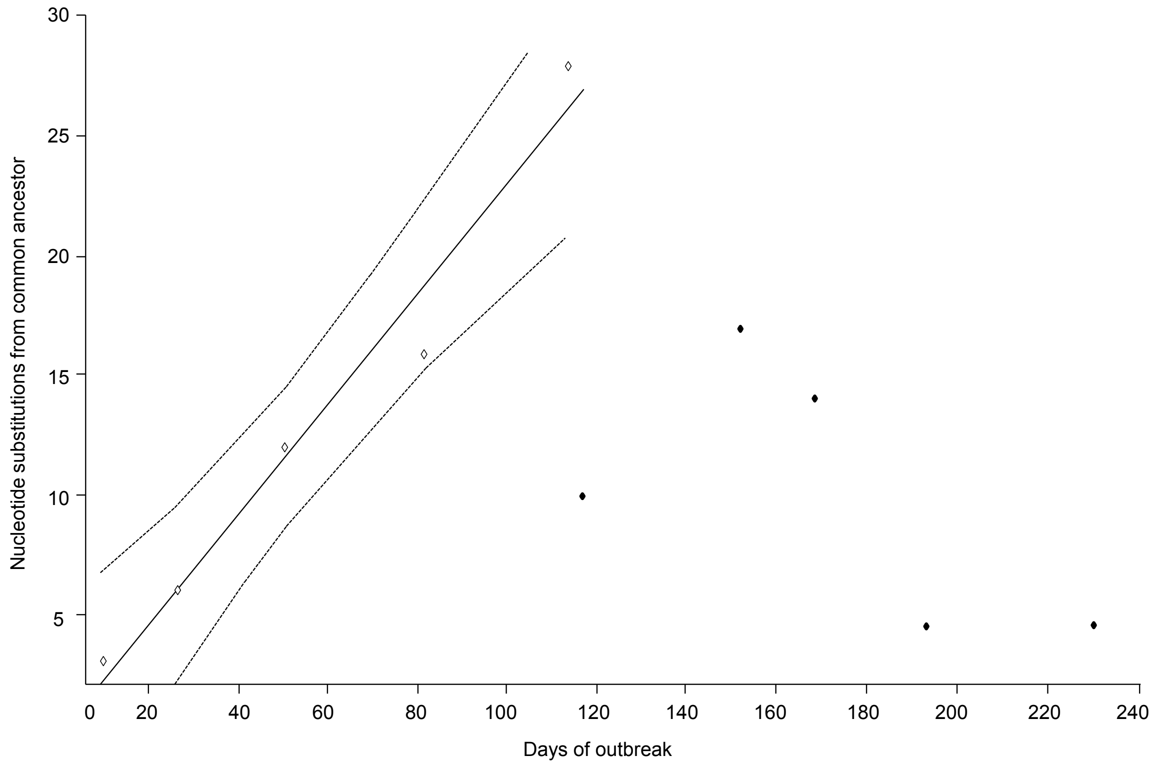


Figure 2.5

Accumulation of nt substitutions over time during OB/Salop. FG sequences derived from samples A-E (open diamonds) and those derived from samples F-J (closed diamonds) are included. A 95% confidence interval either side of the regression line for samples A-E, calculated using the R statistical package, is indicated.

2.5 Discussion

Contemporary molecular based examination of FMDV VP1 and FG sequences has allowed the first fine-scale inter and intra-event analysis of FMD outbreaks that occurred from 1966 to 1968 in the UK. Unlike previous *within outbreak* FMD studies (Cottam, Haydon et al. 2006; Cottam, Wadsworth et al. 2008a), this analysis did not attempt to reconstruct viral transmission pathways. As a consequence of the temporally broad but outbreak shallow sample set (10 samples over 8 months), a high degree of missing farms negated the possibility of transmission pathway reconstruction. However, this retrospective study has granted the means to further solidify viral evolutionary characteristics at the overall outbreak scale.

Phylogenetic analysis of the VP1 sequence of two samples from OB/Salop, one from OB/North, OB/Hants and OB/Warks, along with contemporary isolates from Europe and South America, has shown that there were potentially at least three separate incursions of FMD into the UK. These findings finally confirm that earlier UK outbreaks, including those in Northumberland, Hampshire and Warwickshire were not responsible for the large 1967/68 epidemic. In accordance with the findings of the Northumberland report, samples from OB/Salop showed more genetic relatedness to contemporary South American isolates compared to all other field samples or cell culture isolates analysed. However, identification of both the source and route of virus introduction into the UK is limited by the size and temporal distribution of the samples and isolates tested. As speculated in the Northumberland report, it remains possible that the route of virus introduction that caused the 1967/68 outbreak was via an as yet uncharacterized source in Europe. Although the genetic relationship between the sequences analysed indicates that these UK outbreaks were not linked to each other, it would be interesting to establish whether this holds true for other outbreaks that occurred during the 10 years preceding 1967. Determining whether past outbreaks where due to multiple individual incursions of virus, or potentially undetected persistent infections, could provide valuable insights with regards to future disease prevention measures.

Beyond the resolution provided by VP1 sequences, the FG sequence can yield further insights into the genetic relatedness of outbreaks. For example, the common ancestor estimated for OB/Hants and /Warks is called into question by the large genetic difference observed between the FG sequence of these two outbreaks. However, in order to clarify the relationship between OB/Hants and /Warks, additional FG sequences would need to be determined and analysed (additional outbreak samples and contemporary international samples). The large genetic difference (214-255 nts) observed between the FG sequence from OB/Warks and all other sequences was partially clarified by recombination analysis. Where samples collected from OB/Salop and OB/Hants showed greatest sequence similarity (between 96 and 100% across the genome) with the South American strain O1/Campos/BRA/58[AY593819], the single FG sequence from OB/Warks showed three distinct regions of divergence from this serotype O strain. Subsequent re-analysis indicated that these regions of divergence shared greatest sequence similarity to a serotype A strain (A26/Argentina/[AY593770]), providing evidence of multi-region recombination events between this A serotype strain and the OB/Warks sequence. This finding is supported by those made by Jackson et al. (2007) who conducted a pairwise scanning analysis on genome sequence data from 156 interserotypic FMDV isolates. This previous study concluded that recombination is most likely widespread throughout the non-structural genes of FMDV genomes (Jackson, O’Neil et al. 2007). Evidence has also been found of interserotypic recombination within non-structural regions of FMDV field samples (Li, Shang et al. 2007). Unless there is an awareness of such potential events when reconstructing the dynamics of any epidemic caused by a virus capable of recombination, there is a risk of misleading phylogenetic relationships being presented, especially when considering timed phylogenies. However, further samples from OB/Warks would need to be analysed to confirm the high genetic diversity, and hence low relatedness, observed.

The improved resolution of FG sequences has enabled the investigation of evolutionary dynamics, across the viral genome, over the timescale of a single outbreak, including nt substitution rate. Although it would have been useful to have included the FG sequence for the *actual* index case from Bryn Farm (isolate not available), extrapolations from the estimated most likely common ancestor have provided a number of interesting observations. For example, late Salop samples (I&J) were more closely related (5 nt difference each) to the earliest sample sequenced from this epidemic (A), although they were collected 187 and 217 days later respectively. Genomic sequences from the remaining three late Salop samples (F-H), also demonstrated characteristics of having been outside a susceptible host for a period of time (all lying significantly outside the linear accumulation of nt substitution demonstrated by the early Salop samples). Virus outside a host would not be replicating or incurring nt substitutions, and would therefore appear more like the ancestral virus than would be expected given the time that had elapsed. The suggestion was made by the Ministry of Agriculture (Anon April 1969) that 12/18 farms where FMD occurred twice were due to a recrudescence of the disease, possibly as a result of infected hay remaining on the farm after the original outbreak. However, none of the 5 late samples from Salop occurred on farms from which infected samples had previously been collected. Fomite transmission of infected material to previously FMD-free premises, including hay, could also result in this apparent ‘slowing’ of virus evolution during an epidemic. Evidence for potential fomite spread was therefore provided by the observed reduction in the rate at which nt substitutions were accumulated within the 5 late Salop samples. The equivalent clock rates observed for sequences determined from early Salop samples (8.73×10^{-3} substitutions per site per year) and the 2001 UK FMD epidemic (8.66×10^{-3} substitutions per site per year), provide evidence for a constant clock rate across two FMDV topotypes, Europe-South America (EURO-SA) and Middle East- South Asia (ME-SA), respectively.

Although FG sequencing enabled a more detailed analysis of a single epidemic compared to VP1 sequencing, the consensus sequence did not provide evidence of mixed viral populations within the samples studied here. Mixed viral populations, in the form of ambiguous sites within the consensus Sanger sequence, were found to be rare during a study by Cottam et al. (2006), at an equivalent epidemiological scale. In this study, only three out 21 isolates contained a total of six ambiguous sites. Identification of the major or predominant nucleotide at each genomic position during consensus sequencing leads to a ‘masking’ effect of minority variants present within individual viral samples and will be discussed further in the following chapters. Future work may include the characterisation of FMDV within sample population structures, during a national scale epidemic, using next-generation sequencing (NGS).

This study confirms the rate of evolution in FMDV determined from 2001 data for serial transmission within and between herds comprised of fully susceptible cattle. In piecing together events, this study shows the utility of this method for establishing links between outbreaks separated by time and distance but shows an important caveat when utilizing FG sequencing to infer outbreak timescales - namely the potential for the virus to lie dormant in the environment. Although these data do not provide direct evidence of *recrudescence* in these cases, they do highlight the need for vigilance regarding site/material disinfection before re-introduction of susceptible hosts or transport from a previously infected site to a site containing susceptible hosts.

Chapter 3

Beyond the consensus: Dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation sequencing

The work in this chapter has been published:
Wright, C. F.[1], M. J. Morelli[1], et al. (2011). Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. J Virol **85**(5): 2266-2275.

[1] These authors contributed equally to this work

This work was supported by the Biotechnology and Biological Sciences Research Council, United Kingdom via a DTA PhD studentship, project BB/E018505/1, a SYSBIO postdoctoral grant, project BB/F005733/1, and the IAH's Institute Strategic Programme Grant on FMD. I would like to thank Dr N. Juleff for providing the samples used in the study and J. Galbraith at Glasgow Polyomics Facility, University of Glasgow, for sequencing the samples and for providing advice. I would also like to thank C. Randall, L. Fitzpatrick and M. Jenkins for their care of the animals used in the experimental infection study and Derek Gatherer and Gareth Elvidge for useful discussions.
The analytical and statistical pipeline described in this chapter was constructed and performed by Dr Marco Morelli.

3.1 Summary

The sequence diversity of viral populations within individual hosts is the starting material for selection and subsequent evolution of RNA viruses such as foot-and-mouth disease virus (FMDV). Using next-generation sequencing (NGS) performed on a Genome Analyzer platform (Illumina), this study compared the viral populations within two bovine epithelial samples (foot lesions) from a single animal with the Inoculum used to initiate experimental infection. Genomic sequences were determined in duplicate sequencing runs, and the consensus sequence determined by NGS, for the Inoculum, was identical to that previously determined using the Sanger method. However, NGS reveals the fine polymorphic sub-structure of the viral population, from nucleotide variants present at just below 50% frequency to those present at fractions of 1%. Some of the higher frequency polymorphisms identified encoded changes within codons associated with heparan sulphate binding and were present in both feet lesions revealing intermediate stages in the evolution of a tissue-culture adapted virus replicating within a mammalian host. We identified 2,622, 1,434 and 1,703 polymorphisms in the Inoculum, and in the two foot lesions respectively: most of the substitutions occurred only in a small fraction of the population and represent the progeny from recent cellular replication prior to onset of any selective pressures. We estimated an upper limit for the genome-wide mutation rate of the virus within a cell to be 7.8 x 10^{-4} per nt. The greater depth of detection, achieved by NGS, demonstrates that this method is a powerful and valuable tool for the dissection of FMDV populations within-hosts.

3.2 Introduction

RNA viruses evolve rapidly due to their large population size, high replication rate and poor proof-reading ability of their RNA-dependent RNA polymerase. These viruses exist as heterogeneous and complex populations comprising similar but non-identical genomes, but the evolutionary importance of this phenomenon remains unclear (Eigen 1971b; Eigen and Schuster 1978; Holmes and Moya 2002b). Consensus sequencing identifies the predominant or major viral sequence present in a sample, but is uninformative about minority variants that are present. Evidence for population heterogeneity, where individual sequences differ from the consensus sequence, has been routinely obtained using cloning approaches (Airaksinen, Pariente et al. 2003; Cottam, King et al. 2009a), providing insights into the evolutionary processes that shape viral populations. Unfortunately, these cloning processes are laborious and usually provide only a limited resolution of the mutant spectrum within a sample.

Next-Generation Sequencing (NGS) techniques offer an unprecedented ‘step-change’ increase in the amount of sequence data that can be generated from a sample. Albeit mostly used for de-novo sequencing of large genomes, NGS can be applied to re-sequence short viral genomes to obtain an ultra-deep coverage. Therefore, NGS has the potential to provide information beyond the consensus for a viral sample by revealing nucleotide substitutions present in only a small fraction of the population. Several studies have previously used the 454 pyrosequencing platform (Roche Applied Science) to detect minority sequence variants for human viruses such as HIV-1 (Hoffmann, Minkah et al. 2007; Wang, Mitsuya et al. 2007; Eriksson, Pachter et al. ; Le, Chiarella et al. 2009; Rozera, Abbate et al. 2009; Simen, Simons et al. 2009; Tsibris, Korber et al.), hepatitis B (Margeridon-Thermet, Shulman et al. 2009; Solmone, Vincenti et al. 2009), hepatitis C (Wang, Sherrill-Mix et al. 2010a) and attenuated virus (Victoria, Wang et al. 2010). A promising alternative to 454, is reversible terminator-based sequencing chemistry utilized by the Illumina sequencing platform (Genome Analyzer II). The lower costs of the runs and the higher throughput of this NGS approach are likely to make it widely used for deep-sequencing genomic investigations in the future (Shendure and Ji 2008). Illumina sequencing was recently used to obtain sequences of West Nile Virus, through virus-derived siRNA (Brackney, Beane et al. 2009), mutant viruses of severe acute respiratory syndrome (Eckerle, Becker et al. 2010), and human rhinovirus (Cordey, Junier et al. 2010).

The aim of this study was to explore the extent to which the Illumina sequencing platform can be used to characterize and monitor changes in viral sequence diversity that occurs during replication of a positive-stranded RNA virus within a host. This study uses NGS to dissect foot-and-mouth disease virus (FMDV) within-host population structure, at a depth unobtainable by previous cloning techniques. Belonging to the *Picornaviridae* family, FMDV is highly infectious causing vesicular lesions in the mouth and on the feet of cloven-hoofed animals. The samples analysed here were collected during an infection experiment, in which a bovine host was inoculated with FMDV. We developed a protocol that enabled identification of artefacts introduced during amplification and sequencing which was used to validate and quantify the minority sequence variants that were detected. In particular, we expected to see evidence for the reversion of capsid amino acid residues responsible for heparan sulphate (HS) binding associated with replication of a cell culture adapted strain of FMDV in a mammalian host (Sa-Carvalho, Rieder et al. 1997b; Fry, Lea et al. 1999). Although this study was conducted using FMDV, we anticipate that the features we observe may be broadly representative of populations found in samples obtained from other positive-stranded RNA viruses.

3.3 Methods

3.3.1 Sample preparation and genome amplification

The samples analysed were collected during an infection experiment, in which a single bovine host was inoculated intradermolingually with a dose of $10^{5.7}$ 50% tissue culture infective doses ($TCID_{50}$) of FMDV (O_1BFS 1860). The full-length FMDV genome sequence of this sample had been previously determined using Sanger sequencing (EU448369) and was used as a reference genome in this study. The Inoculum was derived from a bovine tongue vesicle specimen that had been passaged extensively in cell culture (Cottam, Wadsworth et al. 2008).

Total RNA (TRIzol, Invitrogen, Paisley, UK) was extracted from a sample of the Inoculum as well as two 10% tissue suspensions prepared from epithelial lesions (front left foot [FLF] and back right foot [BRF]) collected from the animal at 2 days post inoculation. Reverse transcription was performed using an enzyme with high specificity (Superscript III reverse transcriptase, Invitrogen), and an oligo-dT primer (see Table 3.1).

Table 3.1 Oligonucleotide primers used for the amplification of the two large, overlapping FMDV genome fragments studied (omitting the S fragment up to and including the poly(c) tract), for both the first and second run

PCR Set	Primer [1]	Primer Sequence (5' to 3')	Location on Genome [2]	Amplicon Size (bp)	Overlap (bp)
1	BFS-370F	CCCCCCCCCCCCTAAG	351-366	4557	1051
	BFS-4926R	AAGTCCTTGCCGTCAGGGT	4891-4909		
2	BFS-3876F	AAATTGTGGCACCGGTGA	3859-3876	4317	
	BFS-8193R	TTTTTTTTTTTTTTGATTAAGG	8155-8176		
-	UKFMD/Rev6	GGCGGCCGCTTTTTTTTTTTTTTT	poly(A)		

[1] Last letter indicates a forward or reverse primer

[2] Numbering according to GenBank sequence EU448369

For each sample, two PCR reactions generating long overlapping fragments (4557 bp and 4317 bp respectively) were carried out using a proof-reading enzyme mixture (Platinum Taq Hi-Fidelity, Invitrogen). For biosecurity reasons these individual fragments together comprised <80% of the complete FMDV genome, and corresponded to nts 351-4909 and 3859-8176 of EU448369. This enabled the amplified DNA to be transported outside of the high containment FMD laboratory for sequencing at The Sir Henry Wellcome Functional Genomics Facility (University of Glasgow). The samples were amplified using the following cycling programme: 94 ˚C for 5 min, followed by 94 ˚C for 30 s, 55 ˚C for 30 s and 70 ˚C for 4 min, with a final step of 72 ˚C for 7 min. For each RNA sample, the number of PCR cycles used was optimized (using parallel reactions undertaken using Picogreen) such that products were collected from the exponential part of the amplification curve prior to the plateau phase. Once established for each sample, the same optimized cycle number was used for both runs. Individual PCR products were visualized using agarose-gel electrophoresis and quantified (Nanodrop, Labtech), after which the concentrations of each PCR fragment were adjusted to equimolar ratios for each of the three samples prior to sequence analysis. We repeated the PCR of the original reverse-transcribed sample in order to obtain an independent replica of the amplified sample. The number of viral RNA copies put into the initial PCR reaction was established by quantitative PCR for each of the samples (Callahan, Brown et al. 2002).

3.3.2 Next-generation sequencing

Sequencing was carried out on the Genome Analyzer II platform (Illumina). Briefly, DNA was fragmented using sonication and the resultant fragment distribution assessed by an Agilent BioAnalyzer 2100. After size selection of fragments between 300 and 400 bp, a library of purified genomic DNA was prepared by ligating adapters onto the fragment ends to generate flow-cell suitable templates. A unique 6-nt sequence index, or ‘tag’ for identification during analysis, was added to each sample by PCR. Once the adapter/index modified fragments were pooled and attached to the flow-cell by complimentary surface-bound primers, isothermal ‘bridging’ amplification formed multiple DNA ‘clusters’ for reversible-terminator sequencing, yielding reads of 50 nucleotides. We conducted two sequencing runs: in the first, we sequenced on a single lane the three amplified viral populations (Inoculum, FLF and BRF) after tagging. The second run was performed on a different flow cell: again, we sequenced the same populations on a single lane, using a second, independent amplification of the three original cDNAs. Ideally this second, independent sequencing run would have been conducted on template RNA that had additionally been through two, independent RT reactions in order to account for potential errors introduced during this process and not just PCR amplification (see following section 3.3.4). The second run was performed after the Illumina Genome Analyzer went through an upgrade and was able to deliver longer reads of 70 nucleotides.

3.3.3 Data filtering

In order to make direct comparisons between the two runs, we trimmed reads from the second run to 50nt. Typically, quality scores decreased along a read, as the reliability of the sequencing process decreased with the number of cycles of the

Sequencing Platform. The second run yielded much better qualities thanks to an upgrade of the Illumina platform. For both runs, reads with average error per nt below a fixed threshold ($\theta = 0.2\%$) were discarded to generate a flatter error profile along the read (see Appendix 2, Figure 1). The first and last 5 nts of each aligned read were removed from the analysis as they showed a higher number of mismatches to the reference sequence due to insertions or deletions close to the edges of the reads. More details can be found in Appendix 2, Figure 2.

3.3.4 Validation and analysis of sequence diversity in the samples

The frequency of site-specific polymorphisms was estimated from the frequency of mismatches of the aligned reads to the reference genome. A proportion of these mismatches were expected to be artefactual, arising from a base mis-calling in the sequencing process, or from a PCR error in the amplification of the sample. In order to identify polymorphisms arising from possible base mis-calls in the sequencing reaction, we used the quality score of each nucleotide read to compute the average probability of a sequencing error, p_i, at each site i. Typical values of p_i are around 0.1%. Assuming sequencing errors to be independent, we computed the expected number of such errors as the mean of the binomial distribution B(x; p_i, n_i), where n_i is the coverage of site i. If the observed number of mismatches exceeded this expected number of errors in both runs then we excluded the possibility of a sequencing error. On the other hand, we hypothesize that the probability that PCR errors in both runs independently generated identical base changes at the same site is very low. Based on values quoted for the enzymes used, we estimate that the error rate for the combined RT-PCR amplification process to be 7.7×10^{-6} per base pair copied. We therefore defined polymorphic sites that could not be attributed to sequencing errors and at which both the most common and second most common nucleotides were the same between the two runs to be ‘qualitatively validated sites’. For each site in the set of qualitatively validated polymorphisms, we computed the 95% confidence intervals for the polymorphism frequency using the binomial distribution above. If the 95% confidence intervals from each run overlapped, we defined the polymorphism frequency estimates from the two runs to be in quantitative agreement.

We assessed the quantitative repeatability of site-specific polymorphism frequency estimates by calculating Spearman Rank correlation coefficients between polymorphism frequencies in the samples within each run and between polymorphism frequencies from runs 1 and 2.

We counted the number of transitions (Ts) and transversions (Tv) observed at qualitatively validated sites across the genome, we computed κ=2Ts/Tv, and the relative distribution of mutations across the 1st, 2nd, and 3rd codon positions across the open reading frame (ORF). We obtained an estimate for dN/dS as follows: for each codon of the reference ORF, we computed the expected number of synonymous (s_i) and non-synonymous site (n_i) and, for each read j spanning that codon, the number of observed synonymous (s^r_{ij}) and non-synonymous substitutions (n^r_{ij}). Using all the codons where $s_i>0$, we the obtained the proportion of synonymous (p_S) and non-synonymous (p_N) observations:

$$p_S = \frac{1}{n_{cod}} \sum_{i=1}^{n_{cod}} \frac{1}{m_i} \sum_{j=1}^{n_i} \frac{s_{ij}^r}{s_i}$$

(and analogously for p_N) ,where m_i is the number of reads covering codon i and n_{cod} is the total number of codons in the ORF. From p_N and p_S we have obtained dN/dS according to (Nei and Gojobori 1986).

We calculated the number of validated sites at which STOP-codons are observed within the reading frame, and used these counts to estimate an upper limit on the mutation rate. Let n_i be the coverage at the ith nucleotide position, and let $x_{i,obs}$ be the number of reads indicating a STOP codon at the ith position. Assuming independence, the probability density function describing the number of mutations, x_i, that might be observed at site i is the binomial $B(x_i;\lambda,n_i)$ where λ is the mutation frequency, corresponding to the number of mutations accumulated by a site during a cellular passage. The maximum likelihood estimate of λ is $\sum_i x_{i,obs} / \sum_i n_i$ (Evans, Hastings et al. 2000). Using a flat conjugate prior distribution (beta function with shape parameters set to 1), we obtained confidence intervals for λ from the corresponding posterior distribution (beta function with parameters $1+\sum_i x_{i,obs}$ and $1+\sum_i (n_i - x_{i,obs})$ (Gelman 2004)). Assuming an equal probability for each mutation, λ is related to the mutation rate μ (per nucleotide, per single copying event) via the relation $\lambda = 2ga\mu$ (Thebaud, Chadoeuf et al. 2010), where g is the number of transcription generations (positive -> negative -> positive) the virus underwent in the cell. Here, we assume g=1, which corresponds to a stamping machine replication strategy and therefore to the minimum number of copying events in a cell. *a* is a factor weighting the fraction of mutations generating a STOP codon among all the possible changes that could arise at a single nucleotide position: we only consider sites whose mutations can lead to a STOP codon. Among the 18 codons that are one mutation away from a STOP, 5 of them (UCA, UUA, UAC, UAU, UGG) can reach a STOP codon through either two different mutations to the same position, or a single mutation to one of two different positions. Assuming the same probability for each of the 3 nucleotide mutations, we obtain then a = (4*2+15*1)/(3*19) = 0.4035.

Randomizations were conducted whereby we assembled putative 'clones' from the read data by sampling nucleotides randomly from (qualitatively validated) nucleotide frequencies observed at each site along the genome. We computed the median number of observed nucleotide substitutions (those differing from the consensus of the resampled clones) in sets of 26 independently such assembled clones and these numbers were compared with equivalent numbers from real clones obtained from an individual cow naturally infected with FMDV (Cottam, King et al. 2009a).

The complexity of the viral populations was characterized by computing the entropy of the viral populations:

$$S = -\frac{1}{N}\sum_{1=1}^{N}\sum_{j\in\{A,C,G,T\}} p_{ij}\ln p_{ij}$$

where *N* is the number of sites and p_{iX} is the fraction of reads bearing nt *X* at site *i*. The entropy measures the amount of "disorder" in the population, and it is maximum at a site when all four bases are equally represented.

3.4 Results

In this section, we discuss the results of the Illumina sequencing of three FMDV populations: the Inoculum (a field sample used to artificially infect a bovine host), and two lesions developed on two different feet of the host, 2 days after inoculation. Sequence read data from this study have been deposited in the NVBI Sequence Read Archive (SRA) under accession numbers ERA015837 and ERA015838.

3.4.1 Description and filtering of Illumina data

Sequences from the Illumina Genome Analyzer platform consist of a collection of several million short reads. Sequencing was repeated following independent amplification of cDNA generated through PCR. In the first run ~8% of the reads were discarded because of unresolved nucleotides or corrupted tags. In the second run, ~3% of the reads were discarded. Each nucleotide (nt) of each read is characterized by a quality score, which quantifies the reliability of the base-calling process during the sequencing. Only reads whose average error per nt was below 0.2% (66% for the first run and 95% for the second run) were considered for this analysis. Further details about the reads and the filtering process can be found in the Appendix 2.

3.4.2 Coverage and consensus genomes

Reads that passed the quality test were aligned to the consensus genome sequence of the starting material from which the Inoculum was prepared (see Appendix 2). The mean coverage of the reference genome in the first run was 4863x for the Inoculum, 8665x for the Front Left Foot (FLF) sample and 6594x for the Back Right Foot (BRF) respectively, while for the second run it was 16827x for the Inoculum, 11924x for FLF and 15945x for BRF (Figure 3.1A and B). For some samples (Inoculum and BRF, first run and FLF, second run), the coverage for the two PCR fragments composing the viral genome was not equal. More details on the statistics of the Illumina yield can be found in Appendix 2.

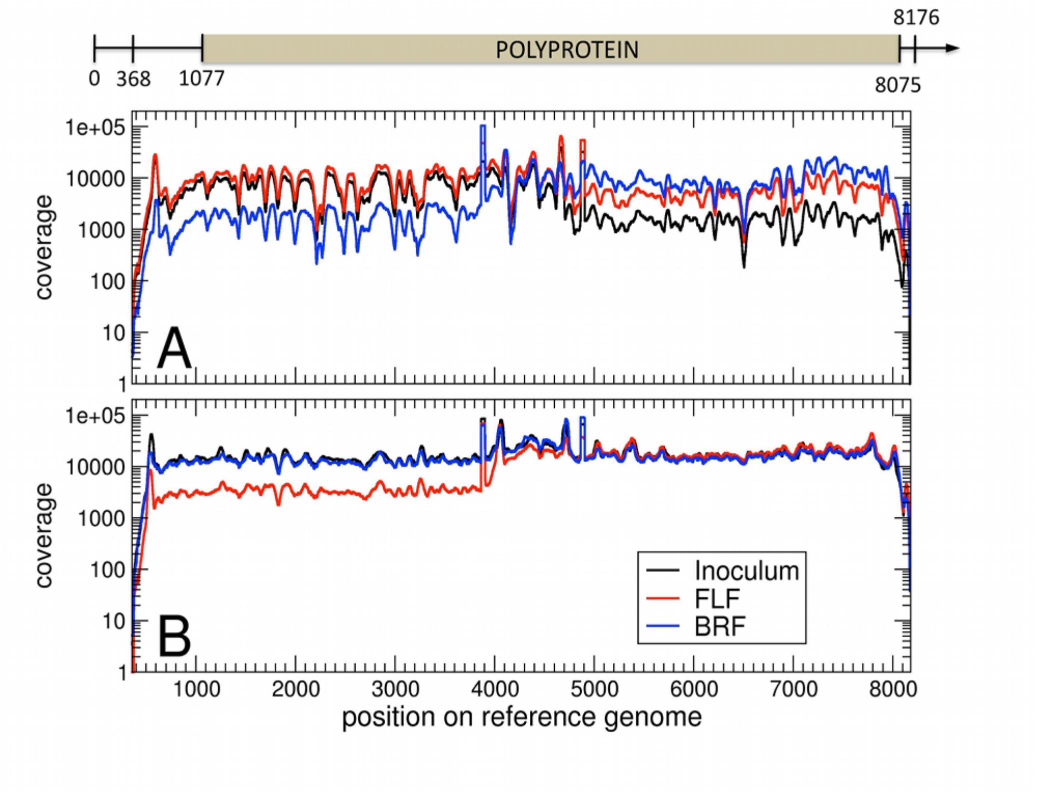


Figure 3.1

Coverage of the reference genome. Obtained with the filtered, trimmed reads. Panel A: first dataset, panel B: second dataset. The three samples (Inoculum, Front Left Foot and Back Right Foot) receive a generous coverage from both runs, while fluctuations are higher on the first dataset. Average coverage is 4873x (Inoculum), 8665x (FLF) and 6594x (BRF) for the first dataset, and 16827x (Inoculum), 11924x (FLF) and 15945x (BRF) for the second dataset. On top of the figure, the sequenced fraction of the genome (nt 368-8176) is represented, together with the position of the polyprotein.

We obtained consensus genomes for each sample, by identifying, site by site, the most abundant nucleotide in the aligned reads. As expected, the consensus for the Inoculum exactly matched the reference genome at all sites. For FLF, both runs indicated two substitutions (nt 2767, G–>A, and nt 8140, G–>T). For BRF sample, the two runs suggested slightly different consensus sequences: the first run revealed five substitutions (nt 2767, G–>A, nt 3138, G–>A, nt 5138, T–>C, nt 7354, C–>T, nt 8134, C–>T), whereas the second run had none. However, at position 8134 about 30% of the reads in the second run showed a T in place of a C, and at position 2767 5% of the reads had an A in place of a T. At the remaining 3 sites, the second run had a small number of reads confirming the polymorphism found in the first run. This result indicates that the same pattern of variation is present in both runs, although the frequency of the mutations is not in quantitative agreement across the two runs for BRF. Finally, the second run showed an almost-consensus substitution in 49.9% of the reads (nt 2754 C->T), which was present at a 10% frequency in the first run.

3.4.3 Validation of polymorphic sites

Mismatch frequencies, obtained by showing site by site the fraction of reads differing from the consensus genome, are shown in Figure 3.2 (first run) and Figure 3.3 (second run). An evident correlation is present between the regions of the sample genomes receiving low coverage and those with the largest fraction of sites showing no variation (Figure 3.2A, second half, 3.2C, first half and 3.3B, first half). Using these raw data, and considering only sites receiving coverage of 100x or more, we found polymorphisms at 7,755 sites in the Inoculum, 7,730 in FLF and 7,710 in BRF, out of the 7,825 nt sequenced. While a few sites exhibited higher levels of polymorphism, the vast majority of sites displayed a mismatch frequency around 0.1%.

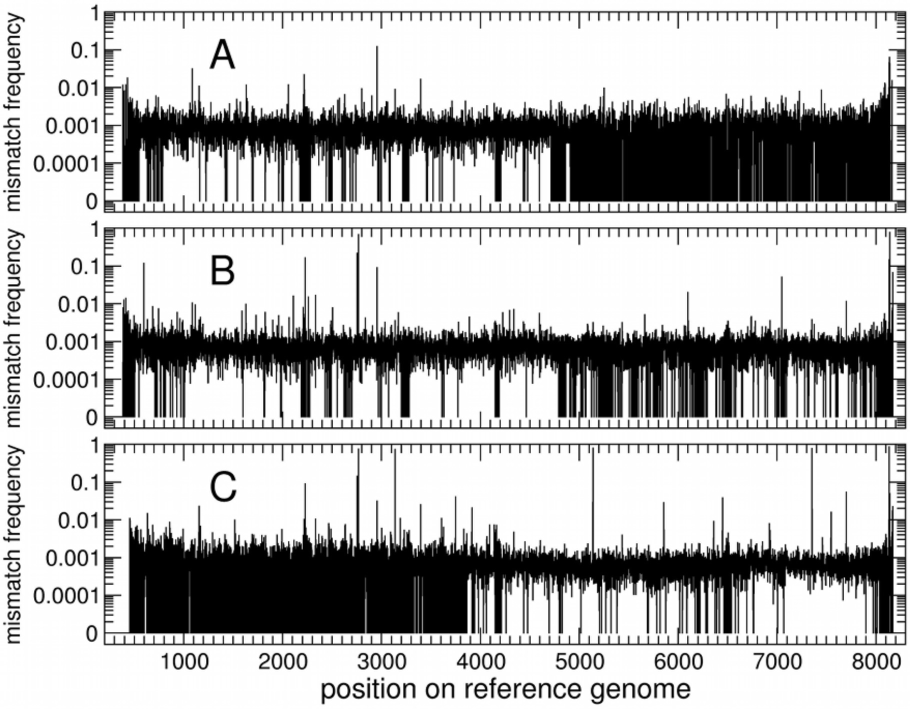


Figure 3.2

Frequency of mismatches (first dataset). Obtained by aligning the reads to the reference genome. Panel A: Inoculum, panel B: FLF, panel C: BRF. The average mismatch frequency lies around 0.1% for all the three samples. At few sites, the mismatch frequency is higher; as expected, the number of these peaks is larger in the FLF and BRF than in the Inoculum. A small fraction of sites show perfect agreement of all the reads with the reference genome (mismatch frequency = 0).

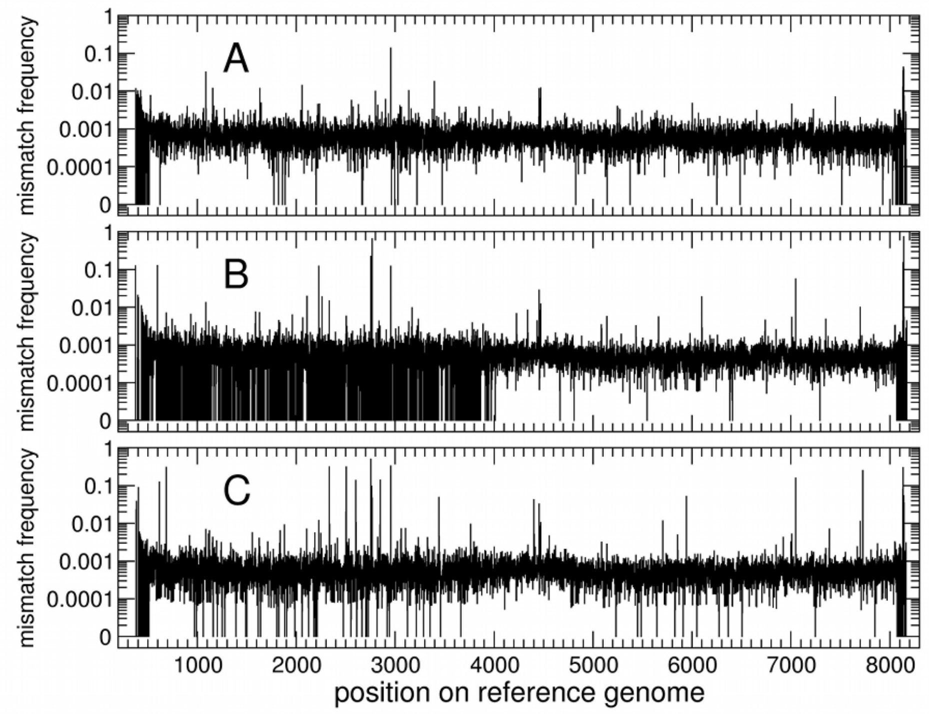


Figure 3.3

Frequency of mismatches (second dataset). Obtained by aligning the reads to the reference genome. Panel A: Inoculum, panel B: FLF, panel C: BRF. This second dataset has higher coverage than the first one, and a lower fraction of sites with no mismatches. The average mismatch frequency is very similar to that of the first dataset.

After screening for possible PCR and sequencing artefacts, we found that qualitatively validated polymorphisms were present at 2,622 sites for the Inoculum, 1,434 in FLF and 1,703 for BRF. The different consensus genomes obtained for BRF in the two runs can be in part reconciled by noting that all six substitutions observed (nt 2754, 2767, 3138, 5138, 7354, 8134) are qualitatively validated in each run. We observed 2,469 quantitatively validated sites in the Inoculum (94% of qualitatively validated sites), 1,303 sites from the FLF (91% of qualitatively validated sites) and 1,528 sites (90% of qualitatively validated sites) from the BRF

Site-specific polymorphism (SSP) frequency at qualitatively validated sites was correlated between the two runs for each of the three samples (Figure 3.4).

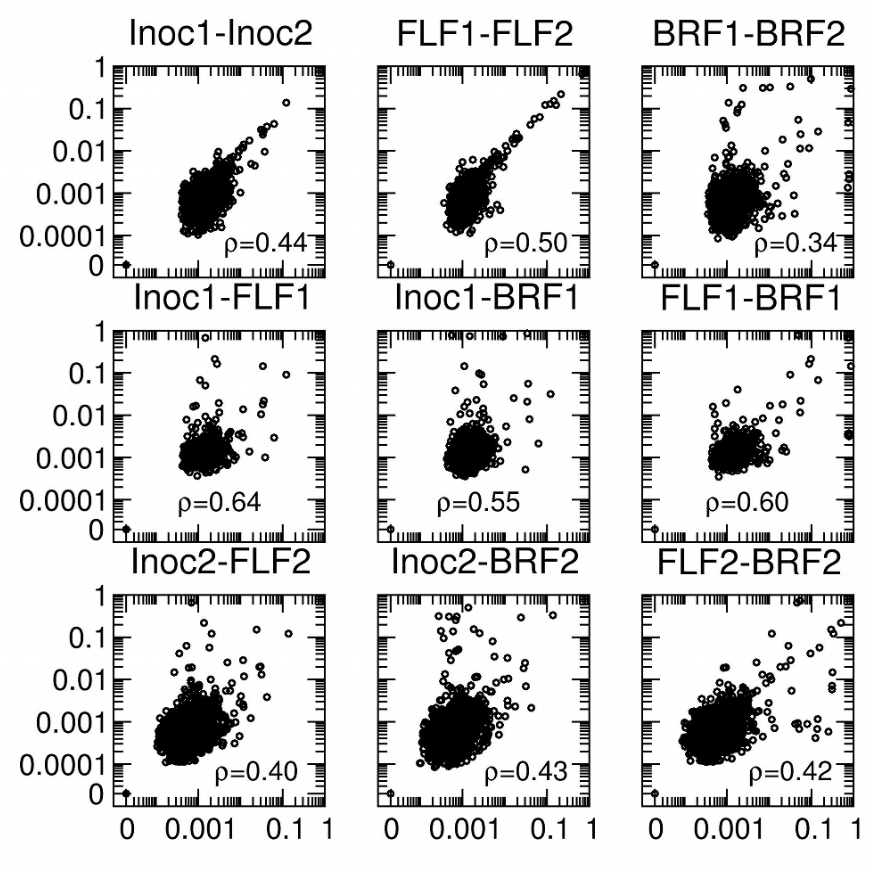


Figure 3.4

Correlations of polymorphism frequencies in the viral populations. Correlations were computed between the two runs (first row) and within each run (second and third row). The Spearman rank correlation ρ is indicated for each pair of datasets. Only qualitatively validated SSPs receiving coverage above 100x in both runs are shown. The correlation coefficients between the two runs in the Inoculum and FLF are similar, while they are lower for BRF. The remaining panels show that the first run is more correlated than the second.

The intra-run correlation for run 1 (Spearman Rank correlation: 0.64 [Inoc-FLF], 0.55 [Inoc-BRF] and 0.60 [FLF-BRF]) was higher than run 2 (Spearman Rank correlation: 0.40 [Inoc-FLF], 0.43 [Inoc-BRF] and 0.42 [FLF-BRF]). The reason for the poor intra-run correlation for run 2 is unclear. The number of viral RNA copies put into the initial PCR reactions was found to be large ($3.2x10^{9}$ for the Inoculum, $6.4x10^{8}$ for FLF and $2.4x10^{8}$ for BRF): assuming that the PCR process amplifies all genomes with the same probability, the probability of resequencing the same genome is exceedingly low ($<10^{-5}$), thus excluding the possibility of biases due to low viral load in the RNA. However, the second run in comparison to run 1 yielded lower amounts DNA library concentrations per sample prior to sequencing (3.4 vs 4.9 ng/µl, 3.7 vs 10.6 and 3.4 vs 9.5 ng/µl ng/µl for the inoculum, FLF and BRF respectively): factors that may have introduced bias into the representative nature of the reads. The intra-run correlation, together with the high fraction of quantitative validation among the qualitatively validates SSPs provides sound evidence that nt changes are linked between the different samples. Inter-run correlation between the samples (Spearman Rank correlation: 0.34 vs 0.44 and 0.50) indicates that validated polymorphisms are unlikely to be artefacts.

3.4.4 Distribution of polymorphisms across the genome

There were 12 SSPs, whose average frequency between the two runs is above 1% in the Inoculum, 19 in FLF, and 25 in BRF (see Supplementary Table). Some of these were clustered in the capsid protein region (beginning of protein VP3) (1 in the Inoculum, 4 in FLF and 5 in BRF) and in the 3’ untranslated region (UTR) (6 in the Inoculum, 5 in FLF and 6 in BRF). Where single reads spanning these sites within the VP3 or 3’UTR were available, there was no evidence that that these mutations were linked together on individual FMDV genomes. In particular, the first cluster was shared between the two foot samples and corresponded to changes encoding amino acid residues associated with heparan sulphate (HS) binding. The Inoculum used in this experiment had undergone extensive cell culture passage and, in common with other in-vitro adapted viruses, utilizes HS as a cellular receptor (Jackson, Ellard et al. 1996b). Subsequent replication in mammalian hosts drives the reversion of positively charged amino acid residues at specific sites in the viral capsid (Sa-Carvalho, Rieder et al. 1997b; Fry, Lea et al. 1999b). A consensus level substitution (>50%) exists within both feet samples of run 1 compared to the reference sequence (see above and the Supplementary Table). This polymorphism corresponded to a change within the 60th codon of protein VP3 ($VP3^{60}$). Although below the level of the consensus sequence, additional qualitatively validated SSPs that were present in both feet samples were detected at four further sites (one codon position in $VP2^{134}$, two codon positions within $VP3^{56}$ and one codon position in $VP3^{59}$) that impact on the ability of FMDV to bind HS. All but one of the mutations that clustered within the 3' UTR of the three samples were located within the first four RNA-RNA pairings either side of the apex of a conserved stem-loop. This structure, one of two stem-loops previously predicted for FMDV and other picornaviruses (Carrillo, Tulman et al. 2005) (Melchers, Bakkers et al. 2000) is thought to generate long-distance RNA-RNA interactions that may impact upon viral replication (Serrano, Pulido et al. 2006). The presence of shared mutations between the two foot samples suggests a common history for the viruses arising as a result of the shared route of intra-host transmission from initial replication sites in the tongue to epithelial sites in the feet via the blood. However an alternative explanation – that the virus is subject to a common selection pressure in both sites cannot be ruled out.

3.4.5 Frequency of site-specific polymorphisms

Some variability was present almost everywhere on the genome. Above minimum coverage of 100x, only 61 sites exhibited no polymorphism (0.79%) in the Inoculum, 59 (0.76%) in FLF and 49 (0.64%) in BRF. These sites received relatively low coverage, suggesting that the absence of observed genetic variability may be due to lack of power to detect it. By grouping the site-specific polymorphism frequencies into discrete bins, we can examine the proportion of sites experiencing different polymorphic frequencies and thereby obtain a comprehensive picture of the heterogeneity in the viral populations (Figure 3.5). Across the three samples, most sites exhibit a range of low-frequency SSPs between 0.01% - 1%.

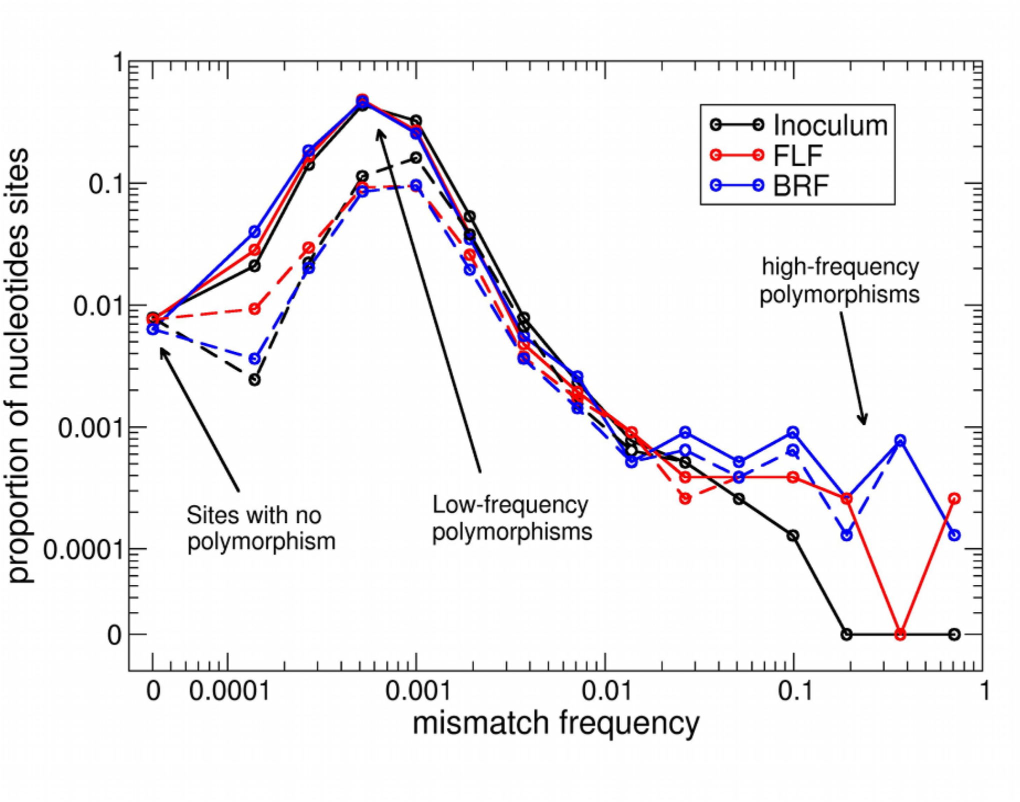


Figure 3.5

Variability in the viral populations. Frequency distribution of the weighted averaged mismatch frequencies between the two runs, for the three samples (the ordinate represents the frequencies of sites showing that fraction of mismatches). Solid lines: all sites receiving minimum coverage of 100 in both runs (7,755 sites for Inoculum, 7,730 for FLF and 7,710 sites for BRF). Dashed lines: sites receiving coverage of 100 or more in both runs, and classified as validated site specific polymorphisms (SSPs) (2,622 sites for Inoculum, 1,434 for FLF and 1,703 for BRF). All lines show a similar trend: a small fraction of the sites (<1%) display no variability in both runs, most of the sites show a very mild amount polymorphism in the viral population (between 0.01% and 1%), while a very small fraction of the sites (0.14% for Inoculum, 0.22% for FLF and 0.39% for BRF) present variation at a level above 1%.

Only a few sites showed higher frequency polymorphism, and these sites were more numerous for the samples from the feet than from the Inoculum, indicating the generation of new high-frequency substitutions during the host passage. The dashed lines (Figure 3.5) correspond to the same analysis restricted to qualitatively validated sites and reveal a similar pattern.

3.4.6 Statistics of polymorphic sites

NGS provided sufficient resolution to detect polymorphisms where two alternative substitutions are simultaneously present. The secondary substitutions (the third most abundant nucleotides in the reads at any particular site) that would have been qualitatively valid even in the absence of the second most abundant nucleotide substitution were present in 67 sites in the Inoculum, 15 in FLF and 41 in BRF. Secondary substitutions typically appear at frequencies below 0.5%, confirming the large amount of low-frequency variability in the samples.

Table 3.2 Statistics of polymorphic sites. General statistics of qualitatively and quantitatively validated SSPs receiving coverage larger than 100x. Ts: transitions in SSPs, Tv: transversions in SSPs, K=2Ts/Tv, dN: non-synonymous mutations in the ORF, dS: synonymous mutations in the ORF, 1st, 2nd and 3rd: mutations in codon positions in the ORF

	Sites	SSPs	Ts	Tv	κ	dN/dS	1st	2nd	3rd
Inoc	7755	2622	2562	60	85.40	0.651	0.288	0.286	0.427
FLF	7730	1434	1400	34	82.36	1.065	0.326	0.333	0.341
BRF	7710	1703	1649	54	61.08	0.680	0.334	0.307	0.359

Table 3.2 shows that transversions are rare among the validated mutations, and thus κ (defined as 2Ts/Tv) is high (however, similar values were reported in (Cottam, King et al. 2009a)). The ratio of non-synonymous to synonymous substitutions in the open reading frame, dN/dS, is higher for FLF than for the other two samples because of the presence of the non-synonymous mutations in a large number of reads at positions 2754 and 2767, associated with heparan sulphate binding amino acid reversions within $VP3^{56}$ and $VP3^{60}$ respectively . The mutation frequency at the third codon position is only marginally higher than in the first and second positions. Taken together, these observations suggest that the observed polymorphisms are dominated by mutations arising during the last round of intra-cellular replication and that have not been subject to extensive purifying selection. Further evidence of this lack of selective pressure is provided by the presence of validated polymorphisms generating STOP codons within the ORF. If it were assumed these mutations are lethal for the virus and therefore subject to purifying

selection during infection of another cell, it follows they would have arisen during the most recent rounds of viral replication. STOP codons were found at 24 sites in the Inoculum, 9 sites in FLF and 21 sites in BRF, mostly at frequencies around 0.1% (with a single exception in BRF where a mutation generating a STOP codon is present in 0.7% of the reads).

The presence of STOP codons can be used to obtain an upper limit on the mutation rate (per nucleotide per transcription event) of this virus. We make the hypothesis that these mutations are lethal and are therefore generated in the last round of cellular replication. Moreover, we assume the replication strategy involving the minimum number of copying events in the cell (the “stamping machine” strategy, see Thebaud, Chadoeuf et al. 2010), and obtained an upper bound for the mutation rate (μ) of 7.8 x 10^{-4} per nucleotide per transcription event (95% CI: 7.4 x 10^{-4} – 8.3 x 10^{-4}), in line with previous estimates (Drake 1993; Drake and Holland 1999; Schrag, Rota et al. 1999).

Finally, we can ask whether these results are broadly consistent with those acquired from cloning studies. In ref. (Cottam, King et al. 2009a), Cottam et al. generated 26 viral capsid clones from an FMDV sample taken from a single lesion of a bovine host. We simulated 10,000 sets of 26 viral capsid ‘clones’, essentially bootstrapping from the nucleotide frequencies revealed by the NGS alignments to be present at each site within the capsid genes. Of these 26 clones, the median number of sequences in each of the 10,000 simulated data sets that were identical to the consensus was 12 (95% CI: 5-17), compared to 15 observed in ref. (Cottam, King et al. 2009a). The median number of simulated clones containing 1, 2, 3, and 4 differences compared to the consensus were 9 (95% CI: 4-14), 3 (95% CI 1-7), 1, (95% CI: 0-3), and 0 (95% CI: 0-1) respectively. These numbers correspond well with those obtained by Cottam et al., (Cottam, King et al. 2009a) which were 6, 3, 2, and 0 respectively.

3.4.7 Complexity of the viral populaiton

In the host, the viral population evolves via extensive replication, mutation, and, at the same time, selection. The result of these combined processes can be quantified by computing how much ‘diversity’ is present within the three samples, using an entropy-like measure *S* that, site by site, takes a maximum value when all nts are present in the same proportion. The entropy of the three populations, computed over the qualitatively validated sites, shows higher values for the feet than for the Inoculum (*S*=0.01138 for FLF, *S*=0.01198 for BRF and *S*=0.00841 for Inoculum), suggesting that repeated cycles of cellular replication during passage in the host does result in greater viral population diversity relative to the Inoculum.

3.5 Discussion

This study describes a novel use of Illumina NGS to investigate the population genetic structure of a positive-stranded RNA virus causing an acute-acting disease in hosts. These experiments generated an unprecedented amount of sequence data and required a new systematic approach to confidently distinguish between sequences that were actually present in the samples from artefacts introduced during the amplification and sequencing steps of the sample processing. Results obtained here were consistent with the findings of previous investigations, providing validation on the use of NGS in the study of FMDV evolution within a host: Carrillo et al. (Carrillo, Lu et al. 2007b) reports an average of 1-5 substitutions per animal passage during an infection experiment in pigs, in line with the 2 substitutions we found in FLF. However, the case of the BRF points out a more complex scenario that could not have been observed with consensus sequences only: the drift of mutations above and below the threshold needed to appear in the consensus. Apparent loss and subsequent regain of mutations during the transmission of the infection across hosts (Carrillo, Lu et al. 2007b) can be explained with this mechanism, which is made more accessible to study by NGS. Moreover, the statistical characteristics of the SSPs we identified (κ, dN/dS) are very similar to those found previously (Cottam, King et al. 2009a), further corroborating the validity of our results. Finally, randomizations of the diversity measured in the capsid region allowed us to obtain simulated clones whose characteristics in terms of mutation were analogous to those found in (Cottam, King et al. 2009a). We conclude that NGS data can be used to examine the nucleotide diversity of each genome position at unprecedented resolution. Observing the mutant spectrum of the viral population at a fine resolution will provide a more sophisticated understanding of evolutionary processes shaping its variability.

Comparisons between the sequences recovered from the Inoculum and clinical lesions provide new insights into the impact of early replication events on viral evolution within a host. This study reveals that only a few sites displayed mutations present in a large fraction of the population, i.e. high frequency polymorphisms (>1%), while the vast majority of the polymorphisms were present at lower frequencies. We hypothesize that the high frequency polymorphisms have been selected over multiple rounds of replication within cells, and that the lower frequency polymorphisms most likely directly reflect the high rate of mutation experienced by these viruses, as our estimate of an upper limit of the genome-wide mutation rate suggests. In this study we used a cell culture adapted virus (as the Inoculum) which gave us the opportunity to monitor changes at specific loci associated with the HS binding site that were under selection pressure during initial replication in a mammalian host. Examination of these sites (collated in the Supplementary Table) reveals for the first time the presence of intermediate stages in the evolution of the viral population between a tissue culture adapted genome and a host-adapted genome.

Cordey et al. (2010) investigated the dynamics of Human Rhinovirus (HRV) during an infection experiment and in HeLa cells, and find results similar to ours in terms of number of mutations fixed at the consensus level (Cordey, Junier et al. 2010). However, while their approach identifies hot and cold spots in the HRV ORF, and some minority variants, the resolution is not sufficient to observe the micro-evolutionary processes whose signature lays in small fractions of the viral population (<2%). Moreover, their estimation of the substitution rate during the infection is based solely on the count of the nucleotides changed among those analyzed: although the value is compatible to our genome-wide, mutation rate, we believe that considering the cellular process of viral replication (and specifically assuming the minimum number or copying events in a cell) allows us to gain a better insight of the process generating variation in the viral population and obtain a more stringent upper bound.

Figure 3.5 reveals that the viral population sequences are highly heterogeneous supporting the findings of previous studies that have used cloning approaches (Domingo, Martin et al. 2006; Jridi, Martin et al. 2006). However, the massively increased coverage enabled by NGS enables the nature of this heterogeneity to be established at much greater resolution. This is important for understanding viral evolutionary processes because heterogeneity is a necessary but not sufficient condition (Holmes 2010a; Holmes 2010b) for the dominance of quasi-species dynamics (see Eigen 1971b; Domingo, Martin et al. 2006 and references therein). For quasi-species dynamics to dominate the micro-evolutionary process, the frequency of a dominant sequence must be maintained primarily by the back-mutation or recombination of closely related genetic variants, rather than the faithful replication of any single genome. This requires a balance of two qualities: genetic variants closely related to the master sequence must be maintained at sufficiently high prevalence; and that mutation and recombination rates must be sufficiently high to generate the observed prevalence of the dominant sequence from these variants. Previous studies have examined this question empirically and concluded that these conditions are indeed met in many RNA viruses, mostly through studies of mutational robustness as a selectable trait ("survival of the flattest" effect in which selection acts not on the dominant sequence, but on the swarm of viruses immediately mutationally adjacent to the dominant sequence) (Domingo, Sabo et al. 1978; Pfeiffer and Kirkegaard 2005), which has been reviewed by Fishman and Branch 2009 with particular focus on Hepatitis C virus. However, taking FMDV as an example, given that there are ~25,000 one-step mutant variants to any one sequence (3 alternative nucleotides at each position of the ~8,300 nucleotide genome), NGS approaches are clearly a powerful tool for examining directly whether viral populations are structured in a way that is consistent with a quasi-species dynamic.

NGS data can be coupled to evolutionary models to estimate parameters, such as the genome-wide mutation rate of FMDV. Here, we computed this number hypothesizing that the viral replication strategy followed the so-called "stamping machine" mode of replication, where all viral genomes leaving the cells are obtained as copies of "first generation" negative stranded genomes, which are in turn direct copies of the genomic RNA originally infecting the cells. For this reason, the estimate of 7.8×10^{-4} per genome per duplication round should be considered an upper bound on the mutation rate which is a tighter estimate than previous figures obtained for other RNA viruses (Drake 1993; Drake and Holland 1999) as a result of the deep coverage that NGS generates. Were the replication strategy "geometric" (i.e. including the possibility of several rounds of positive/negative strand copying before exiting the cell), the mutation rate would be several-fold (perhaps 3-6 times) lower (Thebaud, Chadoeuf et al. 2010). The assumptions that all nucleotide mutations at a site are equally likely, and that all

STOP codons are generated by a *de novo* mutation are also likely to lead to an overestimation of the mutation rate.

To present date, the analysis of the amount of complexity carried by a genome has mostly coincided with information-theoretical measures, aimed to quantify at the entropy and the frequency distributions of short oligomers (Holste, Grosse et al. 2001; Liu, Venkatesh et al. 2008). This approach looks at the “horizontal” complexity along a genome; with NGS we are now able to obtain the closely-related sequences for a whole viral population in a single experiment, thus enabling us to look at the “vertical” complexity of the viral variants, i.e. at the amount of variability present in the population at each site.

A viral population within a host undergoes complex processes, including the onset of infection, cellular replication, selection, and migration to different tissues. In particular, it is not clear how the diversity generated within a cell propagates through a host to give rise to the amount of diversity we observe. The data collected in studies like this can be used for building models aimed at understanding the link between the micro-evolution of FMDV at the cellular scale with the population heterogeneity at the host scale. We anticipate that a model of viral replication across several cell generations within a host will produce a more stringent upper bound to the genome-wide mutation rate.

Although further work is required, these findings strongly suggest that data generated through the use of this methodology can provide novel insights into viral evolutionary dynamics at a greater resolution than previously achieved for a positive-stranded virus such as FMDV. In particular, the genome wide assessment of polymorphic frequencies is likely to be an important asset in the parameterization of models that can evaluate the role of quasi-species dynamics in RNA virus evolution.

Chapter 4

Optimisation of the protocol for NGS template production with clonal control study

NGS data generated during the clonal control study will form part of a wider analysis to be submitted to the Journal of General Virology.

The analytical and statistical pipeline used within this chapter was as described in Chapter 3 (constructed by Dr Marco Morelli). Initial statistical analysis was performed by Dr Morelli and completed by Dr Richard Orton (University of Glasgow).

4.1 Summary

This chapter describes the development and optimisation of a practical method for the preparation of FMDV genetic material suitable for Next-Generation Sequencing (NGS) analysis on the Genome Analyser II platform (Illumina). This method used reverse-transcription (RT) PCR to produce products spanning almost the entire length of the viral genome and could be applied to a range of biological samples, including epithelium, serum and oesophageal-pharyngeal scrapings. Measures were taken to reduce the introduction of artefactual mutation (abbreviated to ‘artefacts’) and bias that can occur during RT and PCR amplification so that the final sequence diversity measured was as representative of initial viral diversity in a particular sample as possible. A ‘Clonal control’ study was also conducted to better understand the introduction of artefacts into viral genomic sequences during the experimental process, at the ultra-deep level by NGS. By more accurately quantifying total artefacts incurred during the production, and subsequent sequencing, of template DNA, a more precise measure of background sequence noise was established. Consequently, a mutation frequency threshold was set, above which there can be relative confidence that the polymorphisms observed are genuine viral mutations. Together, this improved protocol for NGS template production and quantification of background sequence noise contribute to overcoming analytical challenges created by the application of a novel technology.

4.2 Introduction

RNA viruses exist as complex, heterogeneous populations, otherwise known as viral swarms, in which almost every genome sequence contains a natural mutation. However, before these complex populations can be examined, viral genetic material must first be isolated and amplified. This is necessary, firstly, because in any given infected sample, the ratio of virus to host DNA is heavily weighted towards host genetic material. Assuming the following:

- The bovine genome is 2.87 Gb long (Elsik, Tellam et al. 2009) (totalling *1.148^{10}* base pairs for double stranded DNA and considering two copies of each chromosome)
- There are 10^{9} cells per gram of tissue (Alexandersen, Bloom et al. 1988)
- The FMDV genome is approximately 8.5 kb long
- There are between approximately 10^{7} and 10^{9} FMDV genomes per gram of tissue (Murphy, Bashiruddin et al. 2010)

We deduce that, on average, there are at least seven logs more host nucleotides than viral in a gram of infected tissue. Consequently, considering a total sequencing output of 20 GB (Illumina Genome Analyzer IIx specifications for read lengths of 100 bp), this would equate to only x1 coverage of the FMDV genome, if template, containing both host and viral material, was sequenced directly. To solve this problem, viral material can be targeted during PCR amplification, which is commonly achieved using sequence specific primers, but it can similarly be targeted during the preceding step of RNA reverse transcription (RT) in the same way. Alternative measures can be used for total mRNA enrichment, for example, by removing host DNA (by DNase digestion) or by rRNA depletion (using commercial kits such as Invitrogen's RiboMinus kit). Transcribed RNA (cDNA) can also be enriched by microarray-based methods, a technique increasingly used in conjunction with NGS (Chou, Liu et al. 2010; Hong, Doddapaneni et al. 2012). Secondly, the quantity of FMDV RNA that can be directly extracted from clinical tissues remains too low for direct RNA sequencing on any of the NGS platforms currently available. Moreover, the processes required to amplify nucleic acid prior to NGS analysis can introduce artefectual mutations (artefacts), and bias.

4.2.1 Experimental impacts on the viral swarm

The following sections will discuss the potential impacts of each of three fundamental features of RT-PCR on the viral swarm in regards to bias and artefact introduction (see Chapter 1, section 1.5.2b for a more detailed description of the current challenges of RNA sequencing). These features include i) Starting template copy number, ii) Primers and iii) RT-PCR enzymes. While discussed separately, the impacts of each of these factors are linked through the common influence of final sequence coverage.

4.2.1a Starting template copy number

Although what is commonly examined through the molecular sequencing of viruses is a sub-set of the total population that exists within, for example, a single lesion, measures should to taken to ensure this sub-set is a faithful representation of that population. Figure 4.1 demonstrates the importance of starting template concentration in terms of ensuring this faithful representation of the original population post PCR amplification. It was calculated that, for a coverage in the range of 10,000-20,000x, and a viral template copy number of >500,000, the probability of genome re-sampling during PCR was very low ($<10^{-5}$), as discussed in Chapter 3, section 3.4.3. In this instance, the PCR product provides a representative sample of the original population (scenario 'A' in Figure 4.1). Conversely, using the same number of PCR cycles, but with a template input of ~ 5 viral copies for the same sequencing coverage output, the probability of re-sampling during PCR is high (scenario 'B' in Figure 4.1). Scenario 'B' offers a misleading representation of the target viral population post PCR amplification with the over-representation of some variants and potentially the complete loss of others. Consequently, in order to conduct a more controlled comparison between viral populations from different samples, template input was standardised at > 500,000 viral copies.

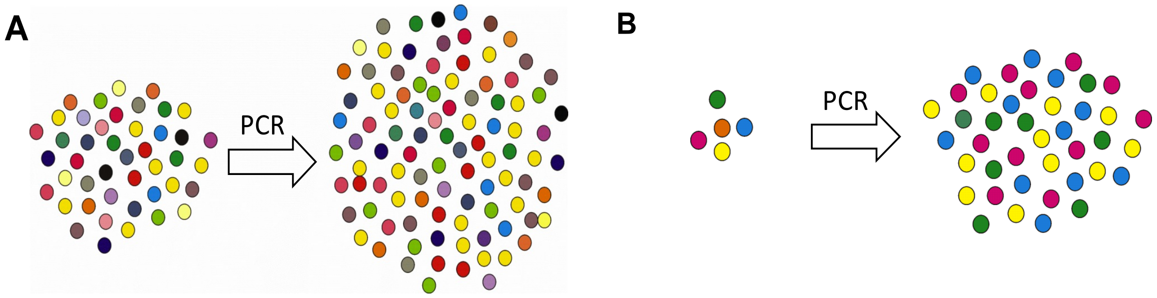


Figure 4.1

Schematic depicting the impact of starting template concentration. Coloured dots depict individual viral variants within a sample of **A**: High viral template concentration, or **B**: Low viral template concentration. In scenario ‘A’ the PCR product provides a representative sample of the original population whereas scenario ‘B’ does not.

4.2.1b Primers

Before PCR amplification, RNA must first be reverse transcribed using one, or a combination, of three common RT priming strategies plus an RT enzyme. There are oligo (dT) primers, which bind to the endogenous poly ‘A’ tail at the 3’ end of mRNA, often producing full-length cDNA. Alternatively, random hexamers (random primers) can be used, which bind to mRNA at a variety of complementary sites and lead to the generation of partial, short length cDNAs. Finally, specific oligonucleotide primers that selectively prime the mRNA of interest can be used. Different RT priming strategies can result in different biases in cDNA and subsequently PCR products. For example, random hexamers are less likely to give a 3’ end bias as compared to an oligo (dT) primer (Stangegaard, Dufva et al. 2006). Random hexamers can also be used to reduce the bias towards positive strand compared to negative stand RT and may overcome difficulties presented by extensive secondary structure in the template leading to bias.

Following RT, template cDNA is amplified by PCR. Primer specificity (ability to anneal to the correct template region) and sensitivity (ability to amplify all genomes in the population), influences the proportion of individual sequences amplified during this process. Variations in primer specificity and sensitivity may lead to some genomic regions being preferentially amplified over others, potentially impacting on estimates of mutation frequency within a viral population (as depicted in Figure 4.2). These primer features are controlled by many parameters, including buffer type, polymerase type/concentration, primer concentration as well as extent and stability of the primer/template match (Mathieu-Daude, Welsh et al. 1996), which is in turn strongly influenced by temperature (Reysenbach, Giver et al. 1992; Ishii and Fukui 2001; Li, Pei et al. 2006).

The melting temperature (T_m) of a primer, defined as the temperature above which the primer will dissociate from the DNA template, depends on the length of the primer designed. If the primers are designed too short, the probability of them annealing at different regions on the DNA template increases; whereas if primers are too long, their T_m would also increase, which may lead to insufficient primer-template hybridization resulting in low PCR product yield.

However, a balance needs to be struck in terms of primer specificity and sensitivity as natural mutations within a viral population that result in primer mismatches can lead to preferential amplification and bias. Molecular studies of bacterial communities have shown that the use of relatively low annealing temperatures can reduce preferential amplification (due to primer mismatches) while maintaining PCR specificity (Ishii and Fukui 2001; Sipos, Szekely et al. 2007). Such biases may also be reduced by the employment of different PCR priming strategies, for example, single poly ’A’ to poly ‘C’ priming or random primers, which can also be used on trace amounts of DNA (Peng, Isaacson et al. 1994; Wong, Stillwell et al. 1996; Zou, Ditty et al. 2003).

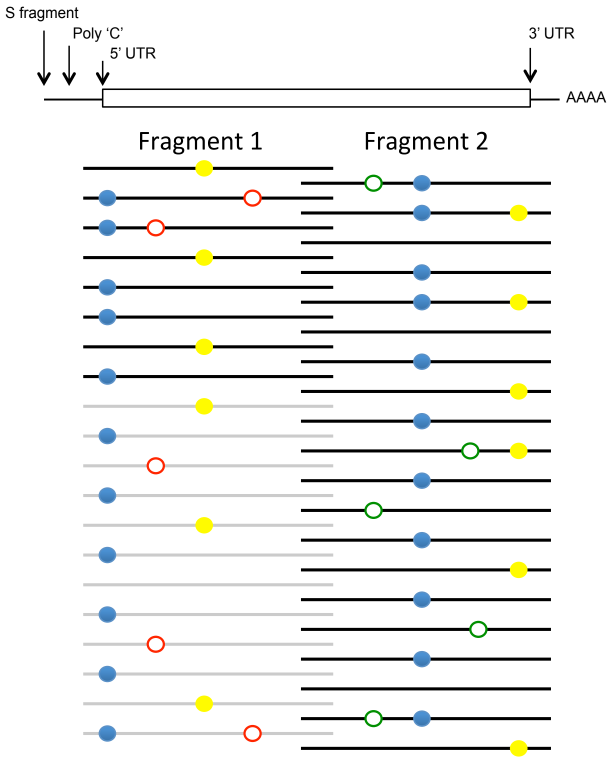


Figure 4.2

Schematic depicting the impact of unequal coverage of the FMDV genome amplified in two long overlapping fragments. On a sliding scale of mutation frequency, blue dots indicate higher frequency mutations (least impacted by fluctuations in coverage), yellow dots indicate intermediate frequency mutations (can be impacted by extreme fluctuations in coverage) and white dots indicate lower frequency mutations (most impacted my fluctuations in coverage), which can either be undetected due to the low depth of coverage achieved for fragment 1 (white dots with red border) or detected due to the higher depth of coverage achieved for fragment 2 (white dots with green borders).

4.2.1c RT-PCR enzymes

Introduction of errors into the target template sequence is predominantly due to the enzymes used during RT-PCR (Mullan, Kenny-Walsh et al. 2001; Malet, Belnard et al. 2003; Arezi and Hogrefe 2007; Domingo-Calap, Sentandreu et al. 2009).

As a non-expansive step, the conversion efficiency of the reverse transcriptase enzyme is an important intermediary in terms of how representative a PCR product is of an original viral population. The conversion efficiency of the Moloney murine leukemia virus (MMLV) reverse transcriptase has been calculated to be approximately only 20% (Curry, McHale et al. 2002). However, the same study showed that a low starting template concentration can negatively impact the RNA-to-cDNA conversion efficiency irrespective of the reverse transcriptase used. Therefore, in order to obtain high yields of quality cDNA, with full and accurate template representation, not only should a reverse transcriptase with high fidelity be used (Arezi and Hogrefe 2007) but, again, we see the importance of having a high starting template concentration. Bias can also be introduced if a reverse transcriptase without high temperature performance is used, especially when transcribing RNA with secondary structure or when working with specific oligonucleotide primers.

Conversely, PCR sees the exponential amplification of the template nucleic acids (cDNA), where errors can be introduced with every cycle. However, when considering first strand cDNA synthesis using a 3' – 5' exonuclease proofreading MMLV RT enzyme, followed by second-strand synthesis and DNA amplification with a proofreading DNA polymerase, RT error rate is higher than polymerase error rate (Arezi and Hogrefe 2007) (details of different RT and PCR enzyme error rates are given in Table 4.4 in the current Chapter, section 4.3.4). The combined masking effect of RT and PCR artefacts on the underlying heterogeneity of the original viral population is demonstrated in Figure 4.3. Here, if naturally occurring mutations occur at a frequency similar to or lower than that of the combined artefacts introduced by RT and PCR, this level of population heterogeneity will be lost to genetic analysis (as indicated by gaps and faint dots within the viral swarm depicted in Figure 4.3).

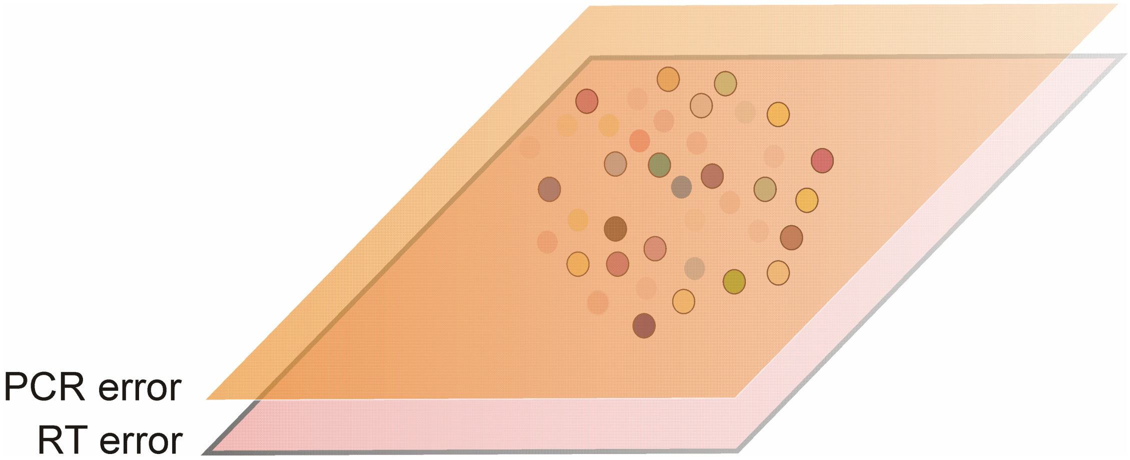


Figure 4.3

Schematic depicting the masking of authentic viral diversity (coloured dots indicate individual variants) by both RT (pink layer) and PCR (orange layer) introduced artefacts. Dot (variant) colour intensity indicative of variant frequency within the viral population: faintest dots equal lower frequency variants; brightest dots equal higher frequency variants. The fainter the variant, the more likely it is to be masked by RT-PCR artefacts.

4.2.2 Objectives

4.2.2a Protocol optimization objectives

The aim of this study was to therefore optimise the protocol used to generate sufficient FMDV genetic material for NGS on the Genome Analyser II platform (Illumina) by two long RT-PCR assays (PCR 1 and 2). The aim of this optimization was not only to improve protocol robustness but also to also minimise bias and experiment introduced artefacts. Unfortunately, it was neither financially nor practically possible to measure the impact of each element of protocol optimization by NGS. However, the alternative output of PCR product yield ('NGS template' yield) was used as an achievable measure of the success of each optimization step with the final goal of achieving an improved, robust assay detection limit. In terms of the minimization of artefact and bias introduction, this measure was based on the hypotheses that optimized production and standardization of NGS template (PCR product) would minimise both, for the reasons discussed in section 4.2.1.

4.2.2b Clonal control study objectives

The aim of this study was to better understand and quantify the cumulative introduction of artefacts into viral genomic sequences of FMDV during the experimental process, at the ultra-deep level, by NGS. Additionally, by more accurately quantifying the total artefacts incurred during the production, and subsequent sequencing, of template DNA, a more precise measure of background sequence noise could be obtained. Above this frequency threshold, there can be relative confidence that the polymorphisms observed are genuine viral mutations. The experimental process included the optimised protocol described in the current Chapter; section 4.3, as well as library preparation for sequencing on the Genome Analyzer II platform (Illumina), as detailed in Appendix 1.

4.3 Protocol optimization

The following section details an improved protocol for generating FMDV genetic material for NGS, taking into consideration the experimental features and impacts discussed in section 4.2.1. Protocol elements were optimized within the confines of the following three practical constraints:

i) <80% of intact FMDV genome can be transported outside the restricted area in any one container at any one time (IAH Biosecurity regulations)

ii) Sufficient length of product required for efficient fragmentation (as part of the Illumina platform library preparation)

iii) Sufficient quantity of total DNA product required for sequencing (minimum of 700 ng as requested by the Glasgow Polyomics Facility at the time of protocol optimization)

It should be noted that PCR priming strategies, such as random and single poly 'A' to poly 'C' priming, discussed in the current Chapter, section 4.2.1b, would not be possible within the confines of the above three practical constraints.

The measures previously taken to achieve the aims set out in section 4.2.2a (as described in Chapter 3; section 3.3.1) had limitations as, without standardization, the potential for variations in the amplification dynamic between samples still existed. Additionally, further consideration and improvement was required in order to achieve these aims using a protocol that was both robust and able to effectively process multiple samples in a realistic timeframe.

Elements of the protocol to be tested and optimized are broken down into four categories:

- Primers
- DNA purification
- Total RNA extraction
- RT and PCR enzymes
- Quantification

Each protocol element incorporates the previously optimized step. The 'Quantification' step was not optimized itself but was used to standardise starting template concentration. Figure 4.4 provides a basic schematic of the original protocol workflow.

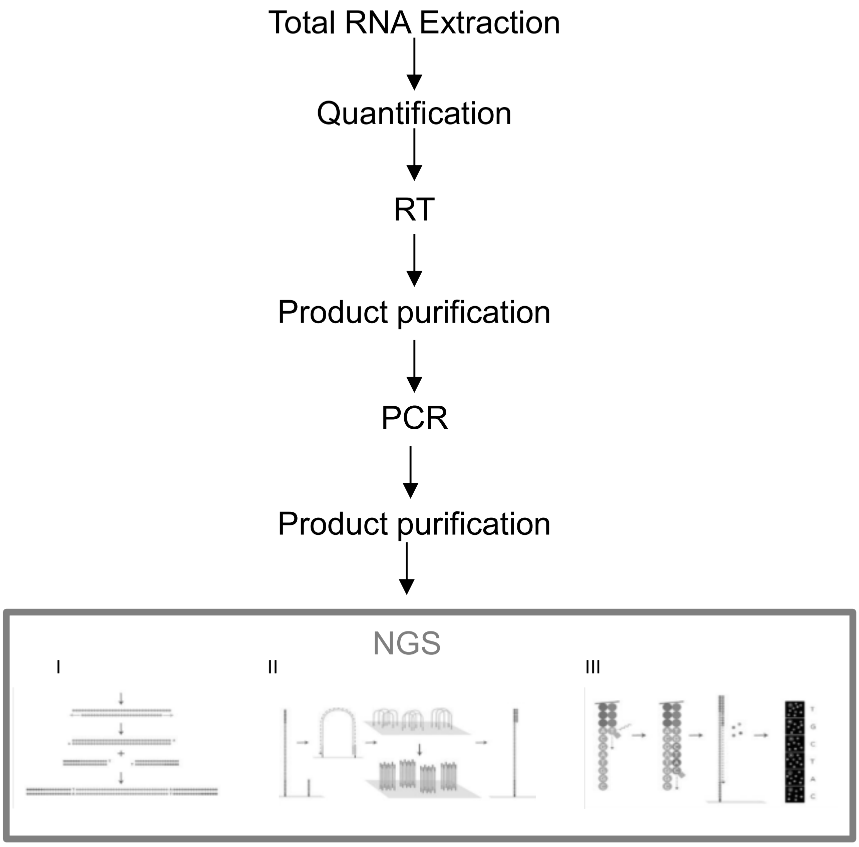


Figure 4.4
Schematic of the protocol workflow for NGS template production. ‘PCR’ corresponds to the two long PCR assays (PCR1 and 2 that equates to genome fragment 1 and 2 respectively. ‘NGS’, figures adapted from (Ansorge 2009), does not form part of the current optimization but is included as a reference to the entire process (NGS includes, **I** Library preparation **II** Cluster generation **III** Sequencing-by-synthesis).

Although the ‘NGS’ process itself (see Figure 4.4) did not form part of the current optimization, impacts of this process, including read coverage, error and bias introduction, on mutation detection, are discussed in Chapter 1, section 1.5.2b.

FMDV epithelium samples from the UK 2007 outbreak, held within the World Reference Laboratory (WRL) library at IAH, Pirbright, were used for protocol optimization. FMDV samples from this outbreak that had tested positive by ELISA (Ferris and Dawson 1988) were selected, as this virus was of the same strain as that studied within Chapter 3 and 5 of this thesis. As discussed previously, the amount of viral RNA in biological samples is often relatively low, therefore, in order to test the sensitivity of the RT-PCR assays, a decimal titration series was made of *test* samples (10^{-1} to 10^{-4}) using negative bovine epithelium suspension (provided by the WRL at Pirbright; also used as negative control during optimization) as a diluent. Therefore, unless otherwise stated, all optimisation steps following the first round of optimised PCR primer testing where carried out on diluted test samples, either at a single or multiple dilutions as stated above. The impact of each optimisation step on PCR product yield was assessed quantitatively by spectrophotometric analysis (Nanodrop, Labtech) and/or visually by gel electrophoresis. Although not a quantitative measurement, visualisation in this way provided a crude but cost effective way of demonstrating differences in yield as well as a means of assessing product size and integrity.

4.3.1 Primers

4.3.1a PCR primers

PCR primers used to amplify the two, long genome fragments (fragment 1 and 2) for the pilot study described in Chapter 3; section 3.3.1, were taken from those used by (Cottam, Wadsworth et al. 2008a). These two, long PCR assays (PCR1 and 2 for the amplification of genome fragment 1 and 2 respectively), individually produced single products of the correct size. However, primer pair specifications within the pilot study protocol, such as coordination of primer pair melting temperature (T_m), had not been optimised. Using quantitative RT-PCR (Callahan, Brown et al. 2002), the limit of detection for the original two assays was estimated at approximately 10^8 copies of FMDV RNA per µl. Therefore, in order to improve the analytical sensitivity of the protocol, whilst limiting potential bias between genome fragments, the decision was made to develop new and improved primers within the same assay conditions. Consequently, except for the individual PCR primers used, PCR1 and 2 shared a single thermal cycle and master mix.

Taking into account the practical constraints mentioned previously, suitable primer pairs needed to be designed so that as much of the FMDV genome was included, whilst continuing to omit the S fragment (as decribed in Chapter 3, section 3.3.1). The published genetic sequence of the inoculum used in the pilot study (GenBank sequence EU448369) was used to design new primers. OligoAnalyzer 3.1, a free online sequence analysis tool provided by Integrated DNA Technologies (http://eu.idtdna.com/analyzer/Applications/OligoAnalyzer/) was used to estimate the T_m of all candidate primers, which were designed such that the T_m was within 3°C of each other (within primer pairs).

In summary, factors taken into account during primer design included:

- Primer length
- Primer T_m (range: 52 – 58°C)
- Primer annealing temperature [T_a] (generally ~ 5 °C less than primer T_m)
- GC content (range: 40-60%)
- Presence of GC clamp
- Avoidance of primer secondary structure
- Avoidance of repeats (ATATAT) and runs (ACGGGGGG)
- Avoidance of template secondary structure

Deep sequence data, generated within Chapter 3, provided an opportunity to check mutation frequency at genomic sites within proposed annealing positions of all new primers. No mutations were found at frequencies >1% within any of the proposed annealing positions within this data set. However, this does not entirely preclude the possibility of mutations >1% ever occurring within these primer annealing regions or account for mutations <1% and therefore the introduction of bias via the selection of populations due to primer mismatches. However, in terms of what was practical and achievable, once the first round of candidate primer pairs had been designed, these were then tested as part of the PCR amplification process described in Chapter 3; section 3.3.1. Briefly, 3 µl of cDNA, was added to 47 µl of master mix (5 µl 10x buffer, 2 µl 50mM $MgSO_4$, 1 µl 10 mM deoxynucleotide triphosphate mix, 1 µl 10 mM forward primer, 1 µl 10 mM reverse primer, 0,25 µl Platinum Taq DNA Polymerase Hi-Fidelity [Invitrogen], 37 µl nuclease-free water). This master mix plus *test* cDNA were run on a PCR program cycle of initial denaturation at 94°C for 5 min and then 39 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 4 min, ending with incubation at 72°C for 7 min. Results of PCR reactions were visualized using a UV camera after running 2-10 µl of PCR product on a 0.7% Agarose gel (Severn Biotech) with 0.002% Ethidium bromide (Invitrogen) at 90v for approximately 40 minutes. The quantitative DNA ladder BenchTop 1 Kb DNA ladder (Promega) was run alongside the products for comparative quantification of product size.

Table 4.1 contains the details for the original primers used for PCR1 and 2 during the pilot study (Un-optimised) and those designed and chosen to be used for subsequent PCR assays (Optimised primers: PCR1i and PCR2i).

Table 4.1 Oligonucleotide primers used for the amplification of the FMDV genomes studied (Un-optimised and Optimised-1)

	PCR	Primer[1] (OBFS)	Location on genome (region)	Amplicon size (nt)	Primer Sequence (5' to 3')	T_m^{oC}	GC(%)
Un-optimised	1	370 F	370-386[2] (5'UTR)	4557	CCCCCCCCCCCCCTAAG	63	82
		4926 R	4908-4926[2] (2C)		AAGTCCTTGCCGTCAGGGT	59	58
	2	3876 F	3876-3893[2] (VP1)	4317	AAATTGTGGCACCGGTGA	55	50
		8193 R	8172-8193[2] (3'UTR)		TTTTTTTTTTTTTTGATTAAGG	43	14
Optimised	1i	516+F	499-520[2] (5'UTR)	4065	CCTTCGCTCGGAAGTAAAACGA	57	50
		4563 R	4545-4563[2] (2C)		CCCGCTGCTTTTCAAGGAT	56	52
	2i	4094 F	4094-4111[2] (2B)	4033	TCTCGACGAGGCCAAACC	58	66
		8126 R	8109-8126[2] (3'UTR)		CTCCTAAGGTGTCGCGCG	58	57

[1] The last letter indicates a Forward (F) or Reverse (R) primer

[2] Numbering according to GenBank sequence EU448369

After multiple primer pairs were tested (alternative primer details are included in Appendix 3), it became apparent that the success of the long PCR assays were dependent on template quality, since repeat freeze/thawing of test RNA resulted in a reduction in assay sensitivity. This consideration was therefore noted for all subsequent handling of biological samples.

The optimal T_a for the new primers was determined empirically by gradient PCR (data not shown). Taking into account the empirically tested T_a results, as well as the need to standardise the amplification process across assays and limit non-specific priming, a single annealing temperature of 60°C was chosen. In order to test detection limit equivalency between PCR assays, using the ‘optimized’ primers and annealing temperature, in isolation, PCR amplification was conducted on a titration series of linearized pT7S3 plasmid (as discussed in the current Chapter, section 4.4). The detection limit of both assays was found to be equivalent (Figure 4.5). All bands were at the correct size (4065 and 4033 nts for PCR 1i and PCR 2i, respectively).

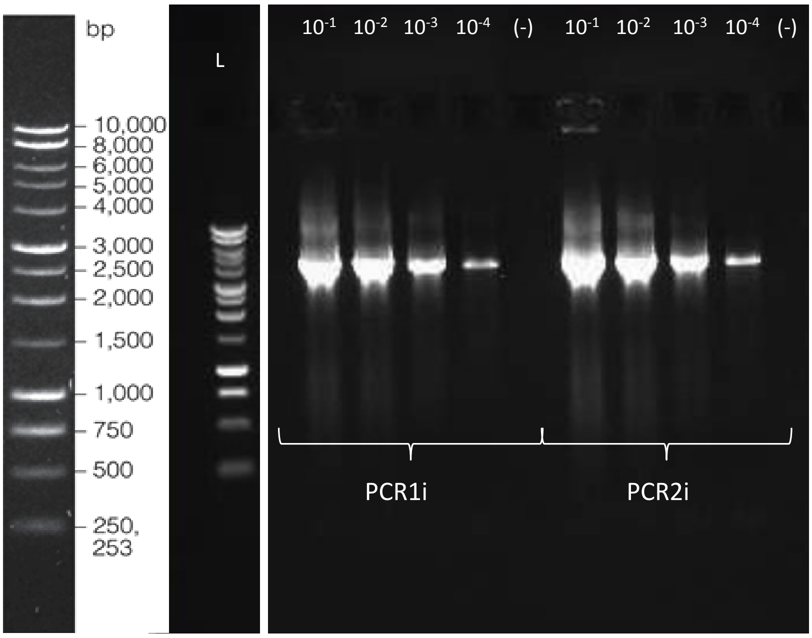


Figure 4.5

Agarose gel depicting PCR products on amplification of titrated, linearized O1Kaufbeuren plasmid. PCR1i and PCR2i primers (as Table 4.2).

4.3.1b RT primers

Previous studies (Stangegaard, Dufva et al. 2006; Domingo-Calap, Sentandreu et al. 2009) have shown that bias may be introduced during RT using the oligo (dT) priming strategy employed during the pilot study, as described in Chapter 3. Therefore, alternative strategies were tested. A combination of random hexamers (Promega) and the oligo (dT) primer (UKFMD/Rev6), used in Chapter 3, was not found to improve the overall detection limit or equivalency of the subsequent two PCR assays after testing in triplicate (data not shown).

Whereas random hexamers lead to the production of partial, short length cDNAs, it was hypothesized that two strategically placed FMDV specific RT primers may lead to the production of more 'fragment 1 and 2' length cDNAs. It was therefore also hypothesised that two FMDV specific RT primers may minimise the 3' bias associated with oligo (dT) RT priming. Table 4.2 provides genome position and details of the two RT primers tested for this 'Dual' priming strategy, which was compared to the oligo (dT) primer (UKFMD/Rev6) for the 'Singular' priming strategy.

Table 4.2 Oligonucleotide primers used for reverse transcription

RT Assay	Primer name[1]	Primer sequence (5' to 3')	Location on genome[2]
'Singular'	UKFMD/Rev6	GGCGGCCGCTTTTTTTTTTTTTTT	Poly 'A'
'Dual'	**OBFS-8193R**	TTTTTTTTTTTTTT*GATTAAGG*	8155-8176
	OBFS-4926R	*AAGTCCTTGCCGTCAGGGT*	4908-4926

[1] The last letter indicates a Reverse (R) primer

[2] Numbering according to GenBank sequence EU448369

Italicized nucleotides are FMDV specific

The 'Dual' RT priming strategy was not found to improve the overall detection limit or equivalency of the subsequent PCR1 and 2 assay after testing in triplicate (data not shown). However, the use of the two FMDV specific primers, detailed in Table 4.2, during RT was incorporated into the optimized protocol for the reasons stated above.

After multiple testing of titrated test samples, using the RT and PCR priming strategies discussed above, the consistent detection limit of the optimized protocol was measured as 10^6 copies of FMDV RNA/µl. The detection limit was assessed by simultaneously conducting the RT-PCR assay and qRT-PCR on the same titrated test sample. The resulting PCR product was then visualized on an agarose gel, whilst yield was measured on a Nanodrop Spectrophotometer and equated to starting RNA template concentration, measured by qRT-PCR (as detailed in Figure 4.6 and Table 4.3 respectively).

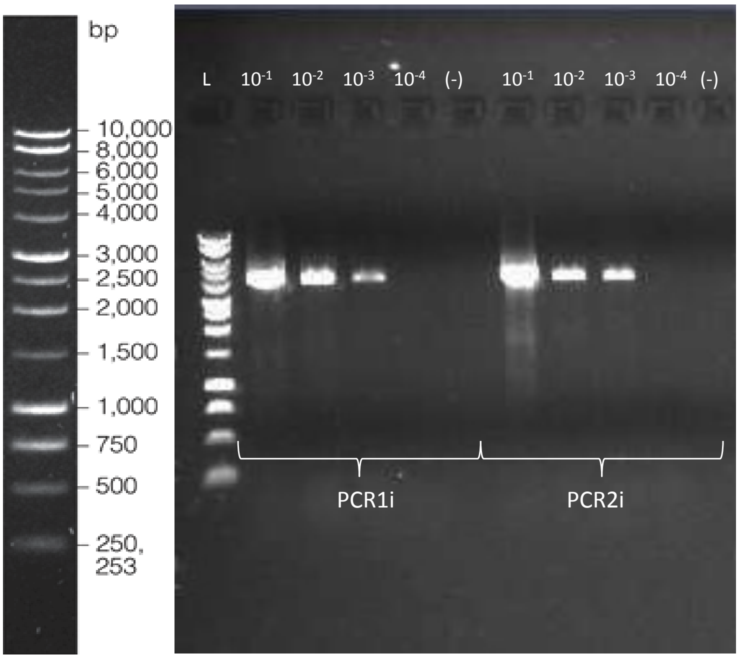


Figure 4.6

Agarose gel depicting PCR products following the final optimized RT-PCR strategy, using 2x RT primers, OBFS-8193R and – 4926R, followed by PCR amplification of FMDV genomic fragments 1 and 2 using PCR1i and PCR2i primer pairs.

Table 4.3 qRT-PCR and Nanodrop measurements for test sample titration series (*Rsq: 0.97, Efficiency: 86%)*

Dilution	**FMDV RNA copies/µl**	**PCR product yield (ng/µl)**	
		PCR1 (fragment 1)	PCR2 (fragment 2)
10^{-1}	3.5×10^{7}	12.0	13.4
10^{-2}	**3.4×10^{6}**	**9.7**	**8.2**
10^{-3}	3.9×10^{5}	7.4	8.4
10^{-4}	6.3×10^{4}	Not measured	Not measured

Above this threshold, a biological sample could be processed and consistently produce sufficient PCR product for sequencing, i.e. combined fragment 1 and 2 totalling at least 700 ng. However, the robustness of this limit of detection was very much dependent on the quality of template RNA; samples that were freeze/thawed more than twice resulted in decreased assay sensitivity. Therefore caution should be exercised when using this assay to process RNA of questionable quality, for example, partially degraded or old RNA. As well as standardization of template starting concentration, which will be discussed in the current Chapter, section

4.3.5, all template RNA was freeze/thawed once, in order to standardise template quality as much as possible between samples.

Optimized primer pairs, PCR1i and PCR2i, were used in all subsequent experimental work and are referred to as PCR1 and PCR2 (replacing the original PCR1 and PCR2 primers).

4.3.2 DNA purification

The DNA purification step is required to remove unincorporated primers after RT and PCR; however it is not a totally efficient process. The recovery of DNA is dependent on product size. The DNA recovery capacity of the illustra GFX PCR DNA Gel Band Purification Kit, by GE Healthcare, for template fragment sizes between 6,000 and 10,000 bp is quoted by the manufacturers as 68.1% and 42.6% respectively. Alternatively, the QIAquick PCR Purification Kit, for the purification of PCR products, 100 bp to 10 kb, by QIAGEN, quotes a DNA recovery capacity of between 90 and 95% for template fragments sizes between 100 and 10,000 bp. This element of the protocol presented a quick and easy step to test and could also increase protocol robustness in terms of achieving sufficient sequencing template. Therefore the two purification kits discussed above were tested to see which resulted in the greatest recovery of DNA. Figure 4.7 shows an improved assay detection limit, after both RT and PCR purification steps, using the QIAquick (QIAGEN) kit compared to the GFX kit (GE Healthcare). The 10^{-3} dilution of test sample was equivalent to an average of 8.1 and 5.7 ng/µl for PCR 1 and 2 respectively for the QIAGEN kit and 7.4 and 3.2 ng/µl respectively for the GFX kit, as measured using the nanodrop spectrophotometer. The gel images in Figure 4.7 are representative of a test performed in triplicate.

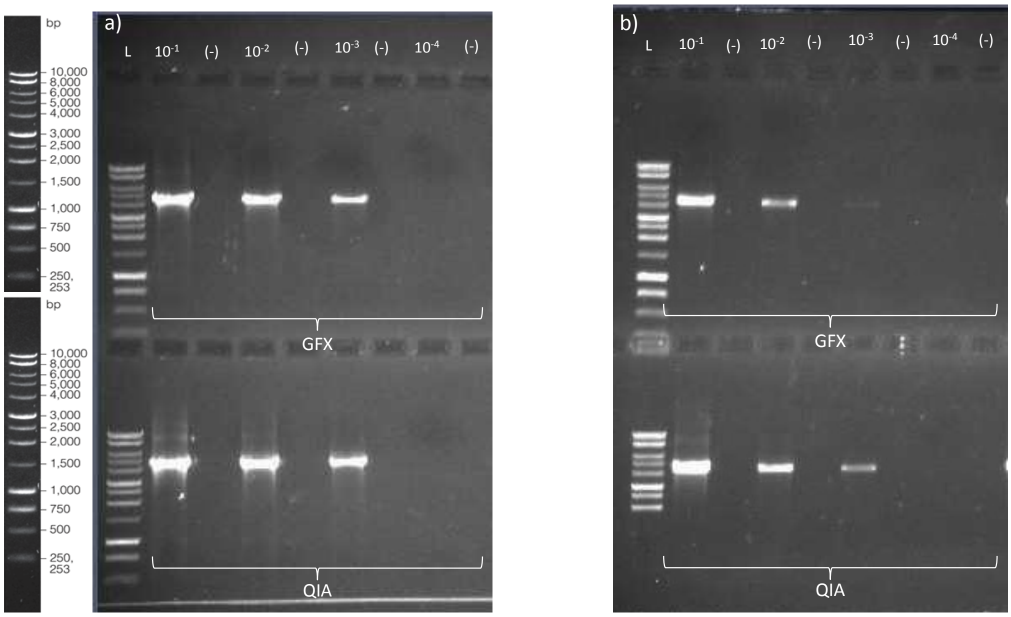


Figure 4.7

Representative agarose gel depicting alternative DNA purification kits; the illustra GFX PCR DNA Gel Band Purification Kit, by GE Healthcare (‘GFX’) and the QIAquick PCR Purification Kit by QIAGEN (‘QIA’) for a) PCR1 and b) PCR2.

4.3.3 Total RNA extraction

The limits of detection for two commonly employed RNA extraction techniques were compared. Total RNA extraction using TRIzol (Invitrogen, Paisley, UK), takes approximately 1.5 hours (operator dependent). In comparison, the equivalent extraction using the RNeasy Mini kit (Qiagen, Crawley, West Sussex), takes approximately 30 minutes (operator dependent). Combining greater sensitivity with a more rapid processing time would reduce the time taken to process multiple samples and potentially improve subsequent product yield from low viral load biological samples by the current assay. Therefore, the above RNA extraction methods were used on titrated test samples, in duplicate. Products of subsequent RT and PCR amplification were then compared, where the TRIzol extraction method lead to a higher limit of detection, for both fragments compared to the RNeasy kit method (Figure 4.8 a and b gel images are representative of duplicate test). The slightly brighter band at the 10^{-4} dilution for PCR1 compared to PCR2 (Figure 4.8a) was not a consistent observation and could have been cause by inherent tube-to-tube variation in PCR product yield or a pipetting error in this instance. The Future use of additional RNA controls to account for such variation is discussed in Chapter 7.

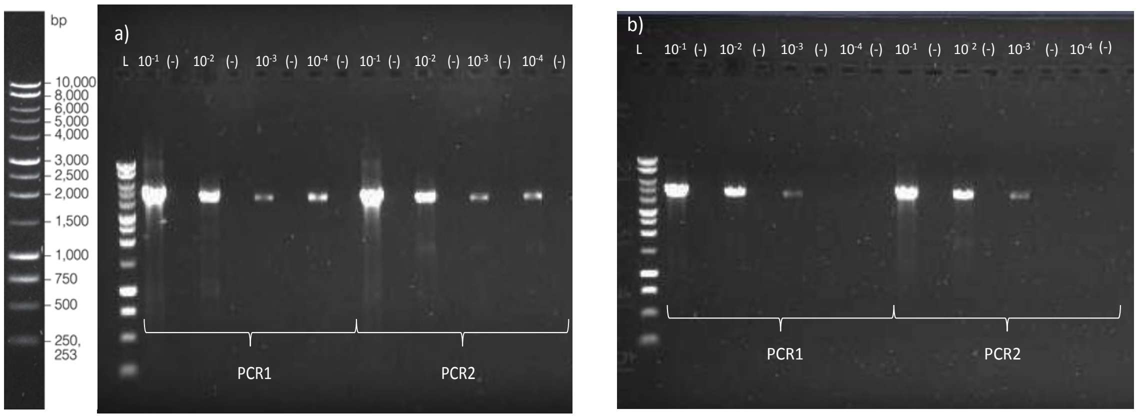

bp
10,000
8,000
6,000
5,000
4,000
3,000
2,500
2,000
1,500
1,000
750
500
250,
253
a)
L 10-1 (-) 10-2 (-) 10-3 (-) 10-4 (-) 10-1 (-) 10-2 (-) 10-3 (-) 10-4 (-)
PCR1
PCR2
b)
L 10-1 (-) 10-2 (-) 10-3 (-) 10-4 (-) 10-1 (-) 10-2 (-) 10-3 (-) 10-4 (-)
PCR1
PCR2

Figure 4.8

Comparison of total RNA extraction methods using either a) the TRIzol method or b) the RNeasy kit method. Representative of duplicate tests.

Therefore, although longer in processing time, the higher limit of detection provided by TRIzol extraction, meant this method was selected for use.

Automation by use of robotic extraction was disregarded due to decreased yield as compared to manual extractions (S Reid and A Shaw personal communication). Limiting contamination was also a priority and therefore samples were processed individually.

4.3.4 RT-PCR enzymes

A literature search of peer reviewed published literature and enzyme manufacturer's documentation, revealed RT and PCR enzymes with reportedly improved fidelity, compared to those enzymes used in the original pilot study (see Table 4.4 for details).

Table 4.4 Details of two of the highest fidelity RT and PCR enzymes (commercially available at time of protocol optimization) plus their non-proofreading counterparts

RT/PCR	Manufacturer	Enzyme name	Error rate
RT	Stratagene	Accuscript^{TM}	1 in 62,200 bases[2]
	Invitrogen	Superscript III®[1]	1 in 35,000 bases[2]
	-	MMLV	1 in 30,000 bases[3]
PCR	Stratagene	*PfuUltra*^{TM} II Fusion	1 in 2,500,000 bases[2]
	Invitrogen	Platinum® High Fidelity Taq[1]	1 in 550,000 bases[2]
	-	Non-proofreading *Taq*	1 in 37,000 to 125,000 bases[3]

[1] Used in the original pilot study

[2] Quoted by manufacturer

[3] Quoted in reference (Arezi and Hogrefe 2007)

The higher fidelity of both enzymes, combined with the reportedly improved processivity of *PfuUltra* II Fusion DNA polymerase, would satisfy the requirements necessary for the fulfilment of the study aims, within the constraints stated previously. The following section therefore details the testing of these alternative RT and PCR enzymes.

4.3.4a RT enzyme

Multiple attempts, using different RNA samples, were made to amplify both FMDV genome fragments, using the optimized protocol, but following RT with Stratagene's Accuscript enzyme (as per the manufacturer's instructions for both master mix components and thermal cycle). However, the subsequent detection limit of the PCR1i assay, following RT using the Accuscript enzyme, was 10 fold less compared to the same assay following RT using the Superscript III enzyme (Figure 4.9 a). This discrepancy was even more pronounced for the PCR2i assay, where the Accuscript enzyme was found to be 1000 fold less sensitive (Figure 4.9 a). All bands were at the correct size (4065 and 4033 nts for PCR1 and PCR2, respectively).

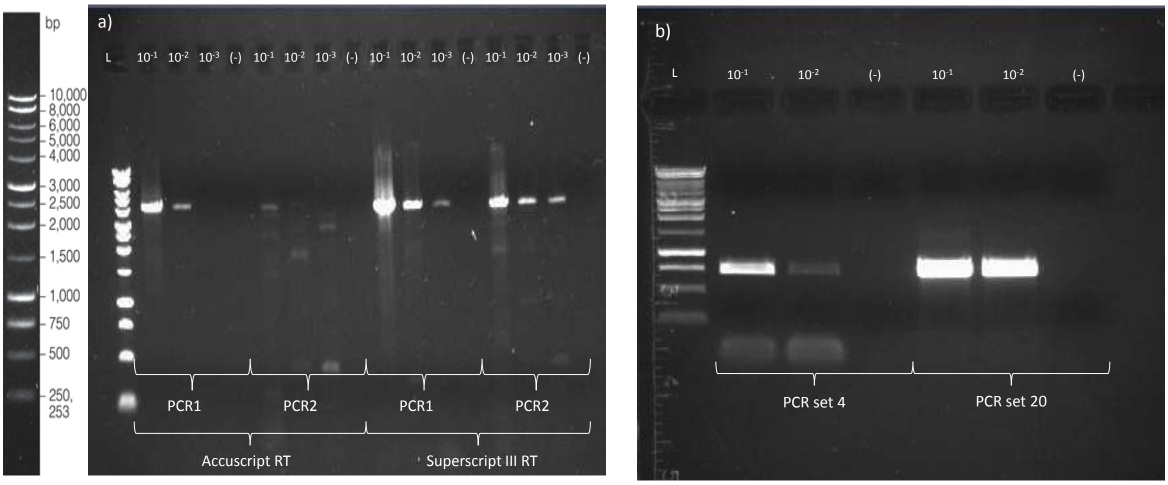

bp
10,000
8,000
6,000
5,000
4,000
3,000
2,500
2,000
1,500
1,000
750
500
250, 253
a)
L 10^{-1} 10^{-2} 10^{-3} (-) 10^{-1} 10^{-2} 10^{-3} (-) 10^{-1} 10^{-2} 10^{-3} (-) 10^{-1} 10^{-2} 10^{-3} (-)
PCR1
PCR2
PCR1
PCR2
Accuscript RT
Superscript III RT
b)
L 10^{-1} 10^{-2} (-) 10^{-1} 10^{-2} (-)
PCR set 4
PCR set 20

Figure 4.9

Agarose gel depicting a) PCR1 and 2 product following RT using either Stratagene's Accuscript or Invitrogen's Supperscript III RT enzymes or b) PCR set 4 and 20 (Cottam, Wadsworth et al. 2008a) product following RT using Stratagene's Accuscript RT enzyme only.

Interestingly, the detection limit was improved by using a PCR assay targeting a shorter fragment (700 bp), using primer set 20 (primer details in the supplementary material of (Cottam, Wadsworth et al. 2008a)), which sits near the 3' end of the PCR2 fragment, after using the Accuscript RT enzyme (Figure 4.9 b). The short PCR assay (700 bp) using primer set 4 (Cottam, Wadsworth et al. 2008a), which sits near the 5' end of the PCR1 fragment, also demonstrated an improved detection limit after using the Accuscript RT enzyme. These results indicated that successful RT using the Accuscript enzyme may have been more dependent on high quality RNA compared to the Superscript III enzyme. As no significant difference between mutations incurred by either RT enzymes tested here was found when studied previously (Cottam, King et al. 2009b), the Superscript III RT enzyme was selected for use.

4.3.4b PCR enzyme

Multiple attempts were also made to amplify both FMDV genome fragments, using the same optimized protocol as described above, but using Stratagene's PFU Ultra II Fusion polymerase during PCR assays (as per the manufacturer's instructions for both master mix components and thermal cycle). However, this polymerase enzyme consistently resulted in a high molecular weight smear across all three dilutions of initial virus tested (Figure 4.10). These results indicated that a substantial degree of non-specific amplification occurred using Stratagene's PFUUltra II Fusion polymerase. Conversely, Platinum High Fidelity Taq produced bands at the correct size (4065 and 4033 nts for PCR 1i and PCR 2i, respectively). It should be noted, the reduced limit of detection for both PCR assays (faint bands only visible down to 10^{-2} dilution) was due to a decrease in RNA template quality caused by repeat freeze/thawing.

PFUUltra II Fusion polymerase is reported as having a 4 fold higher fidelity, compared to Platinum High Fidelity Taq. However, once a PCR enzyme with 3' – 5' exonuclease proofreading capacity is being used, it is the proofreading RT enzyme that more significantly contributes to experimental error introduction (Arezi and Hogrefe 2007). Therefore, no more attempts were made to incorporate PFUUltra II Fusion polymerase into the protocol and Platinum High Fidelity Taq was selected for use.

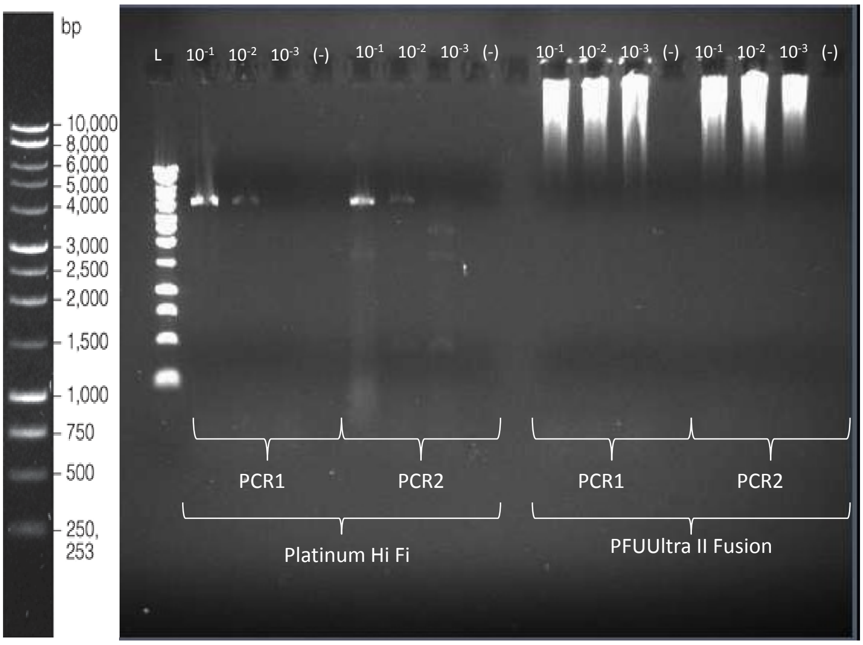

bp
10,000
8,000
6,000
5,000
4,000
3,000
2,500
2,000
1,500
1,000
750
500
250,
253
L 10^{-1} 10^{-2} 10^{-3} (-) 10^{-1} 10^{-2} 10^{-3} (-) 10^{-1} 10^{-2} 10^{-3} (-) 10^{-1} 10^{-2} 10^{-3} (-)
PCR1
PCR2
PCR1
PCR2
Platinum Hi Fi
PFUUltra II Fusion

Figure 4.10

Agarose gel depicting PCR1 and 2 products following amplification using either Invitrogen's Platinum High Fidelity Taq or Stratagene's PFUUltra II Fusion polymerase.

4.3.5 Quantification by qRT-PCR

In order to standardise starting template concentration across all samples (as discussed in section 4.2.1), initial FMDV RNA copy number was quantified using external standards. The FMDV RNA standard was synthesized *in vitro* as previously described (Quan, Murphy et al. 2004). Briefly, linearization of a pGEM®-T Easy plasmid vector (Promega), containing a 950 base pair insert from the 3D region of FMDV O/KUW/4/97, was achieved by Nde I digestion (Promega), as per the enzyme manufacturer's instructions. *In vitro* transcription to generate FMDV RNA was then performed using a MEGAScript T7 kit by Ambion, UK (protocol including in Appendix 3). Finally, plasmid DNA was removed by adding 1 µl of TURBO DNase (supplied with the MEGAScript T7 kit) and incubating the mixture for 30 min at 37°C. qRT-PCR, with and without RT enzyme, demonstrated that this product was 99.999% RNA. The quality and size of synthesised RNA was measured in duplicate using the RNA 6000 Nano Kit on the Agilent 2100 Bioanalyzer, where an RNA template pulsing time of 31.10 seconds was measured, equating to a fragment size of approximately 950 nt (see Figure 4.11 a and b).

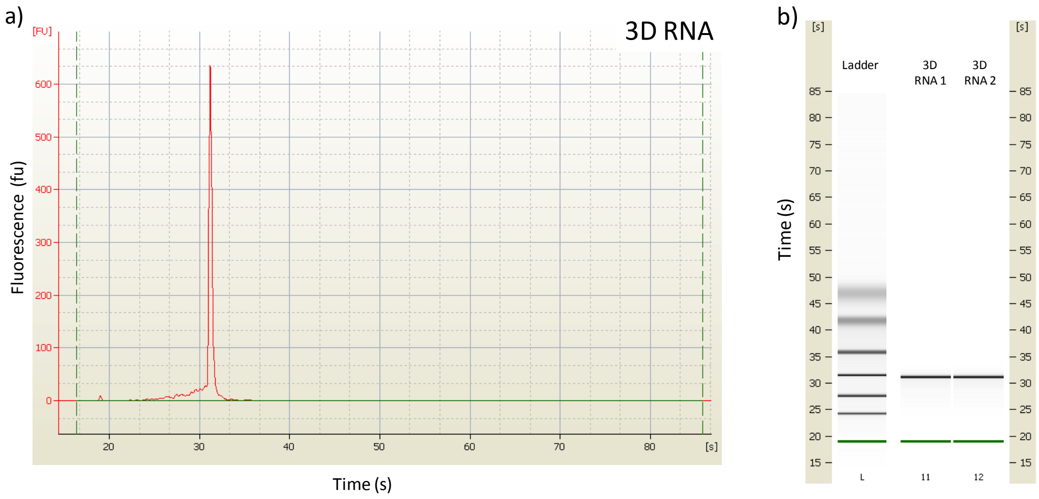


Figure 4.11

RNA quality analysis performed on the Agilent 2100 Bioanalyzer: a) Electropherogram, b) Reconstructed gel image.

Transcribed RNA was quantified using a Nanodrop spectrophotometer, giving an average (between the duplicates) of *774 ng/µl* (260/280 ratio: 2.15). The following equation was used to calculate the number of template RNA copies per ml of standard:

Copies = (6.023 x 10^{23} x weight g/ml of RNA) / molecular weight (ssRNA)

Note, this equation was adapted from (Yin, Shackel et al. 2001).

Aliquots of undiluted RNA standard were made and stored at -80^{o}C and a fresh decimal titration series made for each qRT-PCR quantification assay in order to minimise loss of template through repeat freeze/thawing. A 10 fold titration series from 10^{-5} to 10^{-11} was made for each qRT-PCR quantification by adding 5µl RNA to 45µl of sterile nuclease free water, which, using the above equation, equated to 1.3 x 10^{7} to 1.3 x 10^{1} FMDV RNA copies/µl of standard.

qRT-PCR was performed using the above synthesized RNA standard and an assay which can detect all serotypes of FMDV, as described previously (Callahan, Brown et al. 2002). Briefly, the SuperScript III and Platinum One-step qRT-PCR System from Invitrogen was used, where 5 µl of RNA was added to 20 µl of master mix (12.5 µl 2x Reaction Mix [a buffer containing 0.4 mM of each dNTP and 6 mM MgSO4], 1.5 µl nuclease free water, 2 µl 10mM Callahan 3DForward primer [5' ACT GGG TTT TAC AAA CCT GTG A 3'], 2 µl 10 mM Callahan 3DReverse primer [GCG AGT CCT GCC ACG GA 3'], 1.5 µl 5 mM Callahan 3DP Taqman probe [5' TCC TTT GCA CGC CGT GGG AC 3'], and 0.5 µl SuperScript^{TM} III / Platinum® *Taq* enzyme mix). The above master mix plus *test* RNA were run on a thermal program of a single cycle of 60^{o}C for 30 s, followed by a single cycle of 95^{o}C for 10 min, and then 95^{o}C for 15 s followed by 60^{o}C for 1 min for 50 cycles. The one-step qRT-PCR protocol is included in Appendix 3. qRT-PCR assays were performed on a Stratagene Mx3005P machine (Agilent Technologies, UK).

4.3.6 Optimization summary

A method to amplify the near complete genome of O1 BFS1860 FMDV through two overlapping PCR products, each ~ 4kb in length, was established (see Figure 4.12 for optimized PCR fragment positions and lengths). This optimized protocol was 2 logs more sensitive than that used during the pilot study and was standardized between FMDV genome fragments and across samples in order to minimise the introduction of sequence bias. Failure to incorporate new RT-PCR enzymes of increased fidelity meant the error rate during both of these processes would have remained the same between the pilot and 'Optimized' protocol.

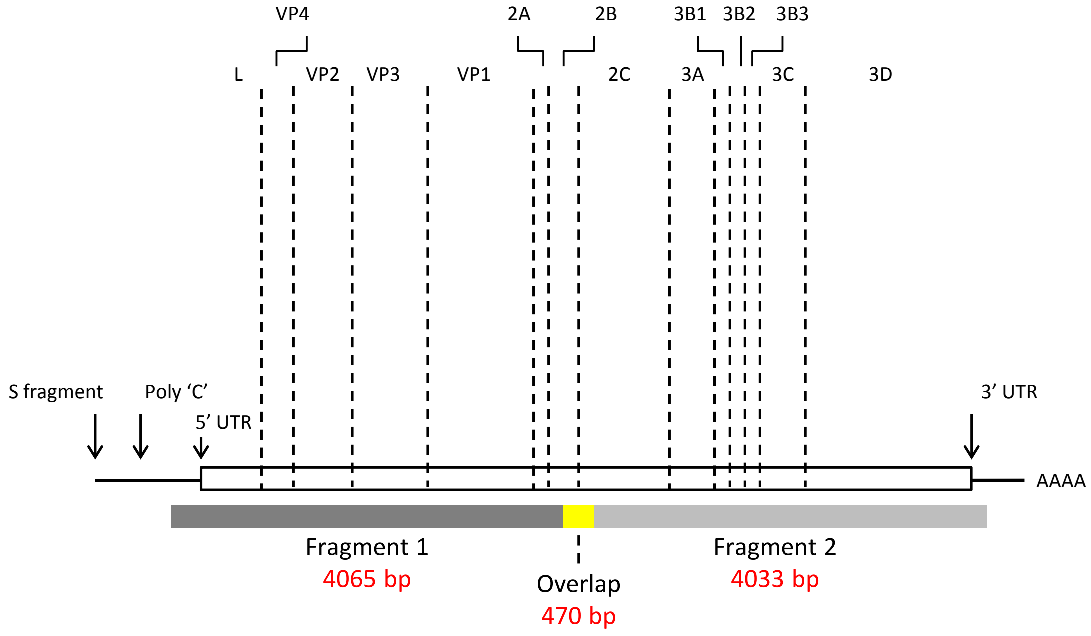
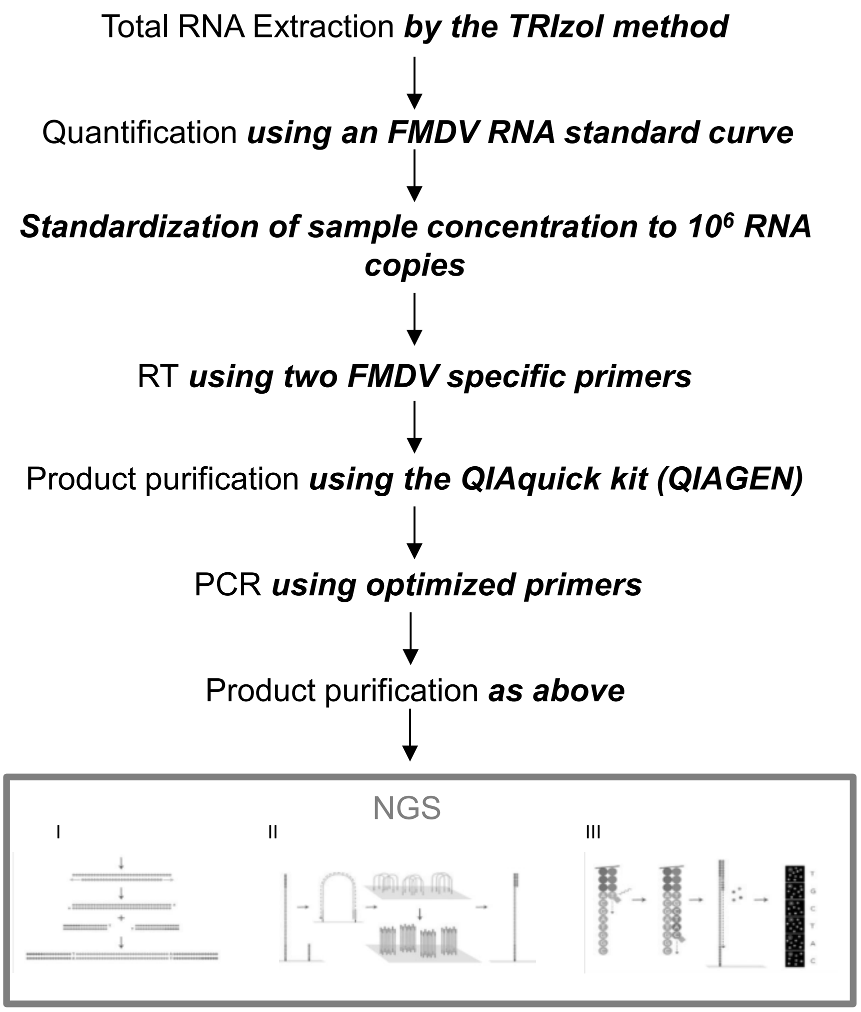


Figure 4.12

Schematic of fragment 1 and 2 positions relative to the FMDV complete genome, as amplified by optimised PCR1 and 2 assays respectively.

Figure 4.13 outlines the overall protocol for this optimized method.



Total RNA Extraction ***by the TRIzol method***

Quantification ***using an FMDV RNA standard curve***

Standardization of sample concentration to 10^6 RNA copies

RT ***using two FMDV specific primers***

Product purification ***using the QIAquick kit (QIAGEN)***

PCR ***using optimized primers***

Product purification ***as above***

Figure 4.13

Schematic of the optimized protocol workflow for NGS template production. Optimized protocol elements shown in **bold**. ‘PCR’ corresponds to the two long PCR assays (PCR1 and 2 that equates to genome fragment 1 and 2 respectively. ‘NGS’, figures adapted from (Ansorge 2009), does not form part of the current optimization but is included as a reference to the entire process (NGS includes, **I** Library preparation **II** Cluster generation **III** Sequencing-by-synthesis).

Figure 4.14 shows the resulting product yield from all PCR1 and 2 assays, using this optimized protocol, for calf 2 (A2) in the transmission chain discussed in Chapter 5. The total range of PCR product concentration, between genome fragments and samples, was 20 – 60 ng/µl, with an average PCR1 concentration of 39.8 ng/µl and an average PCR2 concentration of 36.9 ng/µl. No statistical difference was found between PCR1 and 2 product yields at the 95% CI using the non-parametric Mann-Whitney test implemented within Minitab 15. It is hypothesized that the variation in PCR product yield, between samples, was a result of variations in original RNA template quality. PCR product was also standardized across all samples before **'**Library preparation' (step **I** of 'NGS' in Figure 4.13).

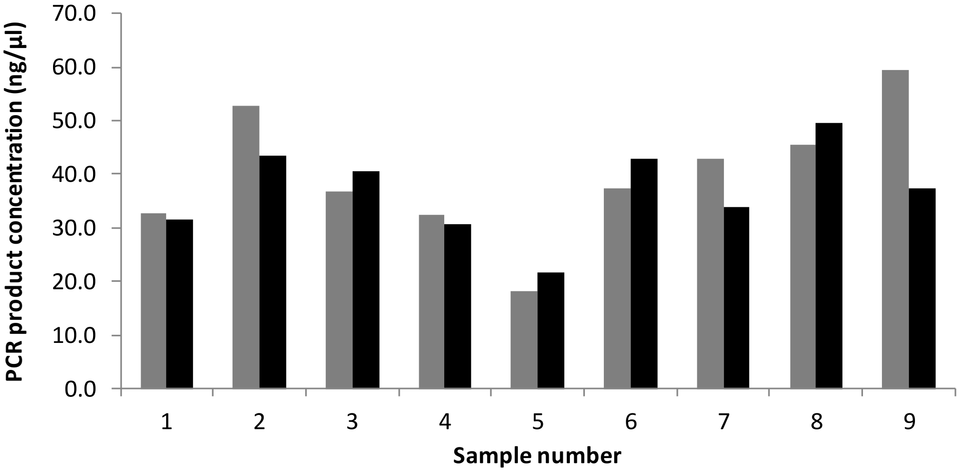


Figure 4.14

Distribution of PCR product concentrations for the final optimised PCR1 (grey bars) and PCR2 (black bars) for all nine samples from calf 2 (A2) described in Chapter 5.

4.4 Clonal control study

The following section describes the assessment of artefact introduction during sequencing template production. By more accurately quantifying the total artefacts incurred during production, and subsequent sequencing, of template DNA, a more authentic measure of background sequence noise could be obtained. This process can be broken down into three main steps, each of which involves the use of commercial enzymes, each having an inherent error rate, as detailed in Table 4.5.

Table 4.5 Details of enzymes and amplification cycle number used during sequencing template production

Step	Enzyme	Error rate	Manufacturer	Number of amplification cycles
RT	Superscript III®[1]	1 in 35,000 bases[2]	Invitrogen	-
PCR	Platinum® High Fidelity Taq[1]	1 in 550,000 bases[2]	Invitrogen	39
Library preparation PCR (Lp PCR)	Phusion® DNA Polymerase	1 in 2,500,000 bases[2]	Finnymes Oy	10

[1] Used in the original pilot study

[2] Quoted by manufacturer

- Non amplification step

In order to better understand the artefacts generated within NGS reads by this process, the cumulative effect of these three steps needed to be measured. To this end, the pT7S3 plasmid (Ellard, Drew et al. 1999) (as described in (Botner, Kakker et al. 2011) and also discussed in Chapter 6, was kindly provided by Veronica Fowler (IAH, Pirbright). From this plasmid, four ‘clone’ controls, spanning the three main areas of artefact introduction during sequencing template production (detailed in Table 4.5), were produced. Figure 4.15 depicts the controls used over this cumulative process.

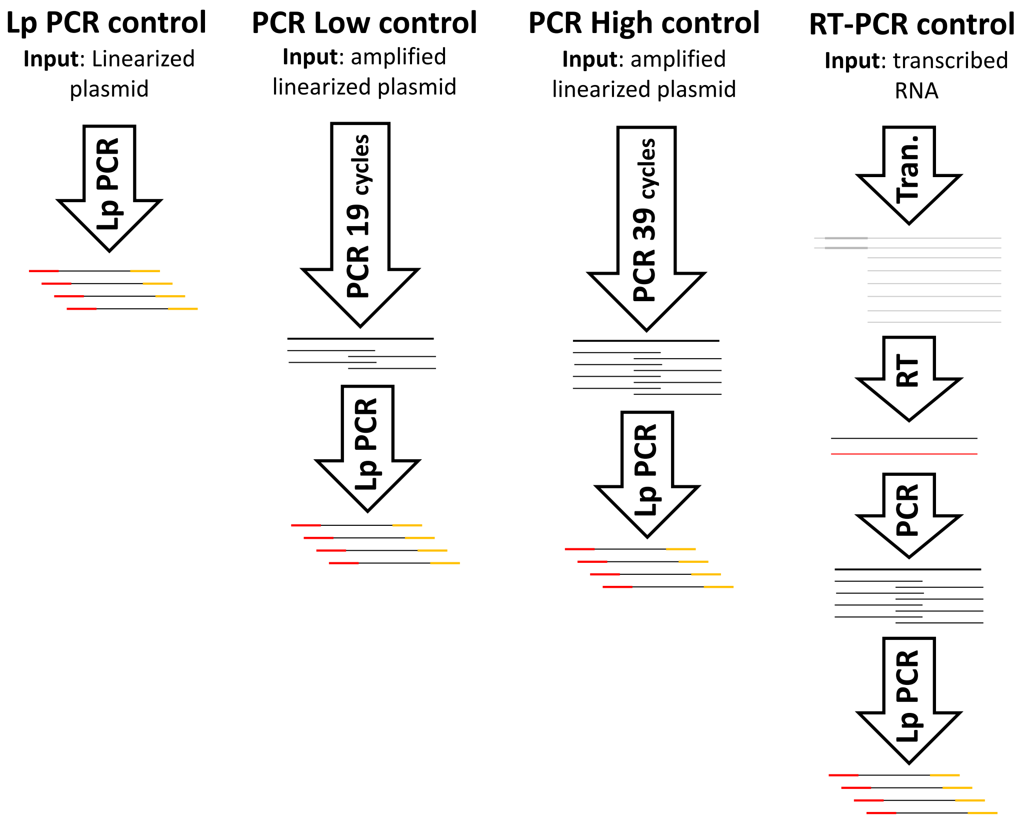


Figure 4.15

Schematic showing the controls used to measure the cumulative effect of the three main experimental steps in sequencing template production: RT, PCR and Library preparation (Lp) PCR. The 'RT-PCR control' contained T7 transcription (Tran.) as an additional point of error introduction

A standard protocol for the transformation, growth and purification of plasmid DNA was followed. Plasmid DNA that had been purified using a QIAprep Miniprep kit (QIAGEN) was pooled to account for any variation between different bacterial colonies. All controls therefore contained artefacts from the bacterial growth system used (E.coli), before the processing steps depicted in Figure 4.15. Studies of Escherichia coli replication, in the absence of DNA mismatch repair and external environmental stress, suggest an *in vivo* nt error rate in the range of 1 every 10^7 to 10^8 nt copied (Schaaper 1993). However, in comparison to a biological sample of FMDV, the starting template for each control represented a relatively clonal population. The 'RT PCR control' would have incurred additional artefacts during *in vitro* transcription of RNA by T7 polymerase, as part of the MEGAscript kit (Ambion), which has a reported error rate of 1 in 500,000 bases.

In addition, all controls would have incurred artefacts during cluster generation by bridge amplification and because of base miscalls during Illumina sequencing, post the processing steps depicted in Figure 4.15. However, the artefacts incurred during these final sequencing steps are addressed in the read filtering and trimming as part of the NGS analysis pipeline (detailed in Supplementary Table S2 in Appendix 4).

It should be noted that the number of amplification cycles quoted for the 'PCR High control' (39 cycles), was the number of cycles used throughout this study, unless otherwise stated. The total quantity of input DNA for the library preparation step was also standardized across all controls.

4.4.1 RT-PCR control

In vitro transcription of O1K B64 RNA was achieved using the same protocol as described in section 4.3.5, albeit following plasmid linearization with a template specific restriction enzyme (Hpa I from New England Biolabs). Again, any remaining plasmid DNA was digested using TURBO DNase, as per the MEGAScript T7 kit instructions (Ambion).Transcribed O1K B64 RNA (8322 nt in length) was quantified by qRT-PCR (as described in the current chapter, section 4.3.5), in order to ensure > 500,000 RNA copies were processed, limiting the probability of genome re-sampling. Following quantification, RT-PCR, according to the optimized protocol described in the current chapter, section 4.3, was performed on the transcribed RNA. However, a small number of changes were made within the PCR2 fragment primers (PT7S3 4094 F $^{5'}$ TCTCGACGA**A**GCCAAACC $^{3'}$ and PT7S3 8126 R $^{5'}$ CTCCTA**C**GGTGTCGC**A**CG $^{3'}$, changed nt highlighted in bold) and within the RT2 primer (PT7S3 8193 R $^{5'}$ GGAATT**G**GTTTTTTTTTTTTTT $^{3'}$, changed nt highlighted in bold). These changes were made to accommodate unique nt substitutions present in the clone. Primer T_m and amplicon size remained the same. Subsequent PCR1 and 2 amplification resulted in single products of the correct sizes (4065 and 4033 nts for PCR 1 and PCR 2, respectively), which were combined in equimolar amounts before sequencing, as described in Chapter 3.

4.4.2 PCR High and Low control

Both PCR controls came from the same pool of linearized pT7S3-O1K B64 plasmid as used for the production of the 'RT-PCR control'. Linearized plasmid was quantified using a Nanodrop spectrophotometer, giving an average of *44.8 ng/µl* (260/280 ratio: 1.96). The number of template DNA copies per ml was then calculated using the same equation as that given in the current Chapter, section 4.3.5 but according to the molecular weight for double stranded DNA (dsDNA).

Two 10 fold titration series were made by adding 5µl linearized plasmid to 45µl of sterile nuclease free water, one from 10^{-1} to 10^{-4} (Titration series 1) and one from neat to 10^{-4} (Titration series 2), which, using the above equation, equated to a total range of 10^{9} to 10^{5} plasmid DNA copies/µl. Linearized plasmid was also quantified by qRT-PCR, as described in the current chapter, section 4.3.5, which measured 2 logs more DNA per dilution. However, RNA standards were used and therefore over-estimation would be expected due to there being no equivalent RT of the template DNA.

PCR1 and 2, as part of the optimized protocol described in the current chapter, section 4.3, was then used to amplify Titration series 1 for 39 cycles and Titration series 2 for 19 cycles. Within Titration series 1, the 10^{-3} dilution resulted in single products of the correct size (4065 and 4033 nts for PCR 1 and PCR 2, respectively), after 39 cycles ('PCR High*'*). Within Titration series 2, neat linearized plasmid resulted in single products of the correct size (as above), after 19 cycles ('PCR Low*'*). Any remaining plasmid DNA was digested using the DPN1 restriction enzyme, as per the manufacturer's instructions (New England Biolabs). The two fragments of each PCR control were combined in equimolar amounts before sequencing, as described in Chapter 3.

4.4.3 Library preparation PCR control

The Library preparation PCR or 'Lp PCR control' came from the same pool of pT7S3-O1K B64 plasmid used to produce all three preceding controls. Biosecurity regulations held at the IAH stipulate that no more than 80% of the intact FMDV genome can be transported outside the restricted area of the Pirbright Laboratory in any one container at any one time. Therefore, NEBcutter V2.0 (available via the link: http://tools.neb.com/NEBcutter2/) was used to find an appropriate restriction enzyme to produce such a fragment from the circular pT7S3-O1K B64 plasmid. Two fragments were created following Pst I digestion. The target plasmid fragment, containing 80% of the FMDV genome (minus part of 3C through to poly 'A' tail), was separated from the second fragment using the QIAquick Gel Extraction Kit, as per the manufacturer's instructions (QIAGEN). In order to produce sufficient product for NGS, five gel extractions were performed and then pooled.

4.4.4 Biological samples

Two biological FMDV samples that had been through the same experimental process for sequencing template production were included in subsequent analysis.

4.4.5 Results

Four 'clone' controls, spanning the three main areas of error introduction during sequencing template production, were sequenced by NGS on the Illumina Genome Analyzer II platform. All validation and analysis of sequence diversity was as described in Chapter 3, section 3.3, albeit using the nucleotide sequence of the linearized pT7S3-O1K B64 plasmid (kindly provided by Veronica Fowler, IAH, Pirbright) as the reference sequence to which all reads were aligned.

Briefly, site-specific mismatch frequencies were grouped into discrete bins so that proportions of sites experiencing different mismatch frequencies could be examined, thereby obtaining a comprehensive picture of the mutation spectrum in each control population (Figure 4.16). This analysis clearly showed the accumulation of low frequency nucleotide artefacts incurred during the experimental process. The 'Lp PCR' control indicates the lowest mutation spectrum with a peak at around 0.05%. The mutation spectrums for the PCR Low and PCR high controls then shift towards higher error frequencies as more PCR cycles are added. The RT-PCR control shifts further towards higher error frequency with addition of the RT step (peak height between 0.05 and 1%).

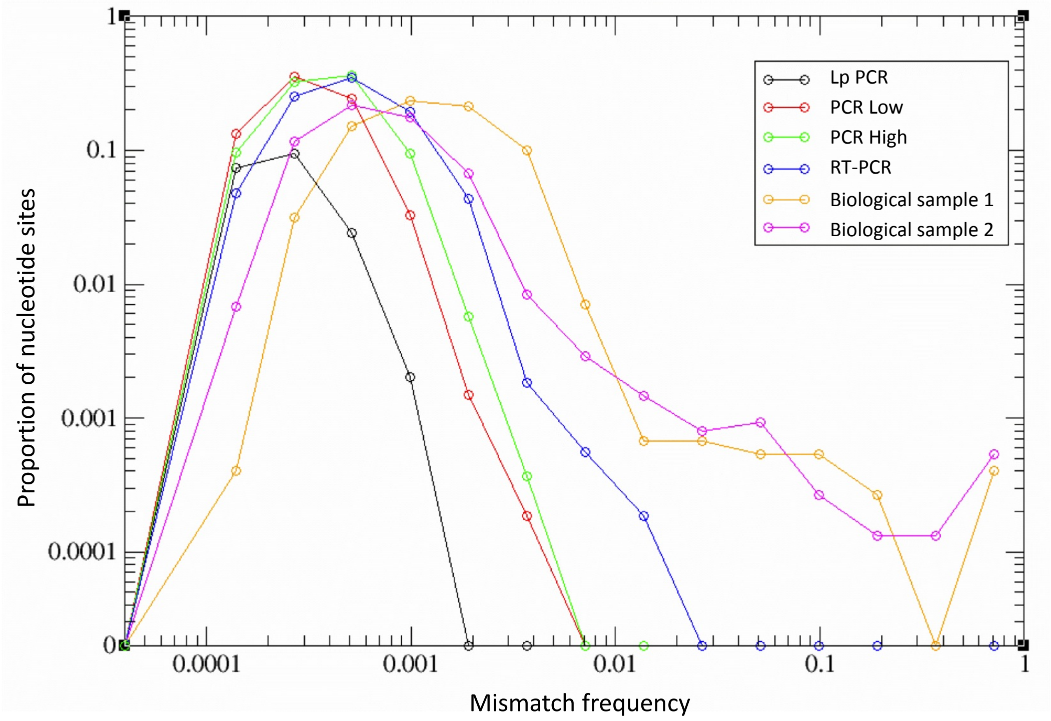


Figure 4.16

Graph showing the proportion and frequency of mismatches against the pT7S3-O1K B64 plasmid reference sequence for all four controls plus biological sample 1 and 2. Moving from the left to the right-hand side of the graph, the large peaks on the left indicate that the majority of sites exhibit low frequency mismatches (mutations), whereas fewer sites exhibit high frequency mismatches, as indicated by smaller peaks on the right, but these occur more frequently in the biological samples compared to the controls.

The majority of errors within the controls were at lower frequencies than those mutations exhibited within ‘Biological sample 1’ (Figure 4.16). However, the genetic diversity within ‘Biological sample 2’ was less than ‘Biological sample 1’, and therefore closer to that seen in the RT-PCR control, until a frequency of around 0.5% where the traces began to substantially diverge. When the mutation spectrum of both biological samples was compared to that of the RT-PCR control, it was found that 95% of mutations above 0.5% occurred within the biological samples and not in the RT-PCR control. The process of measuring genetic diversity within biological samples was as described in Chapter 3 and 5.

4.5 Discussion

An efficient and robust method was developed to generate sufficient FMDV genetic material for NGS on the Genome Analyzer II platform (Illumina), by two, long PCR assays. The improved sensitivity of the optimized protocol meant a greater number of biological samples could be analysed by NGS in the future.

Generally, the dependence of an RT-PCR assay on high quality RNA increases with amplicon length. Unfortunately, RNA is significantly more labile than DNA therefore often resulting in RNA template degradation. The combined requirements of the Glasgow Polyomics Facility for DNA fragmentation and IAH biosecurity, necessitated the use of *two* long PCR assays. Therefore, future projects of this type would benefit significantly from either an ‘in-house’ NGS service and/or alternative fragmentation methods, in terms of ease and robustness of sequencing template production from samples of low initial virus concentration.

In relation to quantifying the detection limit of this protocol, it is important to note that a discrepancy existed between the assay used for RNA quantification by the qRT-PCR method and that used for the two long PCR amplifications described in the current chapter. This discrepancy, in essence, is caused by the greater amplicon size of the two long PCR assays compared to that of qRT-PCR. The majority of RNA, including partially degraded RNA, would have been quantified by the qRT-PCR assay, with an amplicon length of 106 nt. Conversely, only high quality (almost full length) RNA would have been picked up by each of the two long PCR assays, both with an approximate amplicon length of 4000 nt. This discrepancy may therefore explain the relatively high apparent detection limit of the two long PCR assays (final starting template concentration set at 10^6 viral RNA copies/µl), determined simultaneously by the above qRT-PCR assay. This starting template concentration was chosen as it was the one most consistently detected by the two long assays. However, the detection limit of the assay is also very much dependent on RNA template quality.

While reducing the number of PCR cycles would have inevitably reduced the amount of sequence error introduced at the amplification stage, as demonstrated in the current chapter, section 4.4, it would also impact yield. Therefore, the decision was made not to reduce cycle number as this would lead to the exclusion of too many biological samples when using the two long PCR assays. The same RNA extraction protocol was used for all samples, rather than performing optimised extractions per sample type, in order to further standardise the protocol and limit the potential introduction of bias.

A better understanding of the introduction of errors into viral genomic sequences during experimental processing of FMDV samples, at the ultra-deep level by NGS, was obtained using four ‘clone’ controls. A single control, derived from relatively clonal, full-length infectious RNA (transcribed from the linearized pT7S3-O1K B64 plasmid), gave a measure of the total artefacts incurred during sequence template production. The mutation spectrum of this control was compared against that obtained for two biological samples, one of which contained less diversity, i.e. a narrower and lower mutation spectrum, than the other. A more conservative estimate of 0.5%, taking into account the sample containing less diversity (Biological sample 2), was made. Above this frequency threshold we can be confident that 95% of the polymorphisms observed are genuine viral mutations.

Previous work has been conducted, which also used plasmid controls to calculate such a threshold (Margeridon-Thermet, Shulman et al. 2009; Mitsuya, Varghese et al. 2008; Solmone, Vincenti et al. 2009; Varghese, Shahriar et al. 2009; Wang, Mitsuya et al. 2007) (as reviewed in Radford, Chapman et al. 2012). Some studies on HIV-1 (Mitsuya, Varghese et al. 2008; Varghese, Shahriar et al. 2009) and hepatitis B virus (Margeridon-Thermet, Shulman et al. 2009) used a similar technique, comparing mutation frequencies observed within the plasmid controls to those within biological samples. An alternative threshold of 2.0% has been established (Margeridon-Thermet, Shulman et al. 2009; Mitsuya, Varghese et al. 2008; Varghese, Shahriar et al. 2009), above which the authors calculated very low probabilities of artefactual mutations (equivalent to being confident that 99.5% of polymorphisms observed above this threshold are genuine viral mutations). Both the study by Solmone el al. (2009) and Wang et al. (2007) sequenced plasmid controls by both ultra-deep pyrosequencing and the Sanger method in parallel. Taking into account the error rates within both homopolymeric and non-homopolymeric regions and assuming pyrosequencing to be more error prone, any differences then observed between the two methods were considered a product of this method (Solmone, Vincenti et al. 2009; Wang, Mitsuya et al. 2007). Solmone et al. (2009) established a mutation frequency threshold of 1%. Conversely, Wang et al. (2007) doesn’t define the sensitivity of ultra-deep pyrosequencing due to sample specific variations in a number of influencing factors including number of starting templates, read coverage and enzyme introduced errors. All of these potential influences on the calculation of a mutation frequency threshold were taken into account here, as was RT introduced error, which was lacking in the aforementioned studies.

Future work would include repeats of this experiment to ascertain the robustness of this threshold and assess the need to include such a control with each new batch of samples. However, by more accurately measuring errors incurred during individual steps, by use of the three additional controls, three further estimations may be more precisely made: 1) starting population diversity (in this case incurred during bacterial growth and T7 transcription), 2) polymerase error rate and 3) RT error rate. To this end, the sequencing data from this clonal control study will be used as the basis for collaboration with a mathematical modeller, Richard Orton (University of Glasgow) to estimate mutation parameters associated with different parts of this sample preparation pipeline.

Measures have been taken to improve the faithful representation of viral populations for analysis by NGS. These measures become increasingly important as attention is focused on lower frequency polymorphisms in the viral swarm. In the absence of direct high fidelity RNA sequencing, further understanding and reduction of these artefactual mutations and bias is required, in order to investigate the diversity of a viral population below 0.5%.

Chapter 5

Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts

The work in this chapter has been published:

Caroline F. Wright[2,†], Marco J. Morelli[1,†,§], et al. (2013). Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. Vet Res **44**(1):12.

[1] Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom.
[2] Institute for Animal Health, Ash Road, Pirbright, GU24 0NF, United Kingdom
[†] These authors equally contributed to this work
[§] Current address: Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia at the IFOM-IEO Campus, Via Adamello 16, 20139 Milano, Italy

This work was supported by the Biotechnology and Biological Sciences Research Council, United Kingdom via a DTA PhD studentship, project BB/E018505/1, a SYSBIO postdoctoral grant, project BB/F005733/1, and the IAH's Institute Strategic Programme Grant on FMD. I would like to thank Dr N. Juleff for providing the samples used in the study and J. Galbraith at Glasgow Polyomics Facility, University of Glasgow, for sequencing the samples and for providing advice.

The analytical and statistical pipeline described in this chapter was constructed and performed by Dr Marco Morelli.

5.1 Summary

RNA virus populations within samples are highly heterogeneous, containing a large number of minority sequence variants which can potentially be transmitted to other susceptible hosts. Consequently, consensus genome sequences provide an incomplete picture of within- and between-host viral evolution and the dynamics of these viral populations during transmission. Foot-and-mouth disease virus (FMDV) is an RNA virus that can spread through the circulatory system of a host to create multiple lesions in distant epithelia, each of them potentially undergoing independent evolution and seeding subsequent transmission events. The Illumina Genome Analyzer platform was used to sequence 18 FMDV samples collected from a chain of sequentially infected cattle, to obtain snap-shots of the evolving population structures within these different hosts, and to understand how the population structures are influenced by transmission. Analyses of the mutation spectra of the samples reveal polymorphisms >0.5% at between 21 and 146 sites across the genome, while 13 sites acquire mutations in excess of consensus frequency (50%). These results highlight that a number of minority variants can be transmitted during host-to-host infection events, while the size of the bottlenecks appear to be narrower (i.e. tighter) between samples from the same host. This suggests strong intra-host founding effects and a rich within-host viral diversity. The dynamics of minority variants are dominated by genetic drift rather than a strong selective pressure, with the consequence that populations collected in the same host can be more divergent than populations observed in different hosts.

5.2 Introduction

Foot-and-mouth disease virus (FMDV) is a positive sense RNA virus, belonging to the *Picornaviridae* family, and the causative agent of the highly contagious and economically serious foot-and-mouth disease (FMD). RNA viruses evolve rapidly due to their large population size, high replication rate and poor proof-reading ability of their RNA-dependent RNA polymerase (quoted mutation rates commonly fall in the range of 10^{-3} – 10^{-5} per nt per transcription cycle (Duffy, Shackelton et al. 2008)). Within their hosts these viruses exist as complex, heterogeneous populations, comprising non-identical genome sequences (Eigen 1971c; Eigen 1978; Holmes and Moya 2002a). Much of the genetic variation within FMDV populations is thought to be driven by neutral selection or to be under varying levels of purifying selection, with evidence for positive selection observed in only a small fraction of capsid codons perhaps in response to interaction with the host immune system (Haydon, Bastos et al. 2001). To facilitate rapid replication and intra-host dissemination, FMDV has evolved specific mechanisms to evade the early innate and adaptive immune responses, as reviewed by Golde, de Los Santos et al. 2011. Infected hosts typically show clinical signs of FMD within 2-6 days post exposure that include vesicles on the coronary bands of the feet, in the mouth and on the tongue and teats (Alexandersen, Oleksiewicz et al. 2001). Although alternative primary sites of replication have been studied (for a review, see (Arzt, Juleff et al. 2011)) rapid dissemination of FMDV from host entry most likely follows initial replication in the pharyngeal area, passing into the systemic circulation (Burrows, Mann et al. 1981; Alexandersen, Zhang et al. 2002b; Alexandersen, Quan et al. 2003), where the virus is thought not to replicate and from which virus is transported to other distant, non-contiguous epithelia, including those of the feet, where the virus can once again replicate.

As a consequence of this transport to the discrete replication sites and subsequent establishment of new local foci, the initial viral population undergoes an intra-host ‘bottlenecking’ process, similar to that encountered during host-to-host transmission. The founder effects caused by these bottlenecks as the virus disseminates from the host inoculation site and replication in specific tissues have been observed by conventional sequencing during serial FMDV infection in pigs (Carrillo, Lu et al. 2007a) and by use of cDNA clones in poliovirus infection in mice (Pfeiffer and Kirkegaard 2006). Subsequent transmission of virus to a naïve host most frequently occurs shortly after the appearance of clinical signs (Charleston, Bankowski et al. 2011) when an infected individual can secrete large amounts of viral particles into the environment before developing a specific immune response.

An integral part of any disease control strategy is the epidemiological tracing of virus transmission, which, together with conventional field investigations, has largely been achieved with the application of molecular and phylogenetic methods (Samuel and Knowles 2001; Knowles and Samuel 2003; Cottam, Wadsworth et al. 2008a; Abdul-Hamid, Firat-Sarac et al. 2011; Kasambula, Belsham et al. 2011; Valdazo-Gonzalez, Knowles et al. 2011). Global tracing of FMDV movements have been successfully achieved using consensus sequences of the gene for one of the three surface exposed capsid proteins of the virus (VP1) (Samuel and Knowles 2001; Knowles and Samuel 2003; Kasambula, Belsham et al. 2011). However, at shorter ‘epidemic’ time scales, where the viral populations have not substantially diverged, VP1 sequencing cannot provide the required resolution. At this scale, complete genome consensus sequencing (CGCS) has proven to be a very powerful tool for transmission tracing (Cottam, Wadsworth et al. 2008a; Abdul-Hamid, Firat-Sarac et al. 2011; Valdazo-Gonzalez, Knowles et al. 2011). Both the heterogeneous nature of within host viral populations and the number of transmitted viruses between hosts may influence the rate of fixation of mutations (Kinnunen, Poyry et al. 1991; Villaverde, Martinez et al. 1991a); by only identifying the major viral sequence within a sample, CGCS masks the complex substructure of minority variants present and is therefore blind to subtle genetic differences between isolates that are closely related in space and time. Therefore, the level of resolution afforded by CGCS is inadequate to fully characterize single host-to-host transmissions and in particular to monitor the dynamics by which mutations accumulate over single transmission events. As a consequence, how variability generated at the intra-host scale is transmitted on to the inter-host scale is still poorly understood.

Next-Generation Sequencing (NGS) techniques provide the means for rapid, cost-effective dissection of viral population dynamics at an unprecedented level of detail (Hoffmann, Minkah et al. 2007; Wang, Mitsuya et al. 2007; Eriksson, Pachter et al. 2008b; Margeridon-Thermet, Shulman et al. 2009; Rozera, Abbate et al.

2009; Simen, Simons et al. 2009; Kampmann, Fordyce et al. 2011; Chapter 3). The resolution and high throughput nature of NGS platforms has the potential to allow differentiation between samples at the inter- and intra-host scale of infection. This technology has already been applied to compare 'longitudinal' samples of hepatitis C virus (HCV) and to study human immunodeficiency virus (HIV) infection and transmission (Fischer, Ganusov et al. 2010; Wang, Sherrill-Mix et al. 2010b; Bull, Luciani et al. 2011). These studies highlight the size of the population bottleneck during inter-host transmission as a likely influence on the long-term rate of nucleotide (nt) fixation. In contrast to both HIV and HCV, where typically only a few viral particles are transmitted to a naïve host (Fischer, Ganusov et al. 2010; Wang, Sherrill-Mix et al. 2010b; Bull, Luciani et al. 2011), investigations of the inter-host dynamics of equine influenza virus and norovirus have revealed inter-host transmission events to be characterized by a wide (i.e. loose) bottleneck (Murcia, Baillie et al. 2010; Bull, Eden et al. 2012). NGS platforms have been used for investigations over time scales sufficient to incorporate the influence of intra-host scale immune pressures on RNA virus population diversity and subsequent transmission (Fischer, Ganusov et al. 2010; Wang, Sherrill-Mix et al. 2010b; Bull, Luciani et al. 2011; Bull, Eden et al. 2012). However, much less is known about the insights that NGS technology can provide about the within and between host viral population dynamics of acute acting infections, particularly prior to the onset of a specific adaptive immune response.

Utilizing Illumina NGS technology, this study investigates the evolutionary dynamics of FMDV *intra*- and *inter*-host transmissions during serial, acute infections, both through time and between samples, prior to the onset of the adaptive immune response. Due to the greater resolution offered by NGS, we were able to characterize the polymorphic structure of viral populations within the samples collected from three hosts. These data were combined with those from a previous study of the inoculum and first bovine host in this chain (Chapter 3). We investigated the diversity and relatedness between these populations, the dynamics of polymophisms across the genome through time, and were able to make an assessment of *inter*- and *intra*-host bottleneck size.

5.3 Methods

5.3.1 Transmission experiment and sample collection

The samples analysed were collected during an infection experiment where FMDV was passaged in series through a group of four calves (Juleff, Valdazo-Gonzalez et al. 2013). Calf 1 (A1) was inoculated intradermolingually with a dose of $10^{5.7}$ 50% tissue culture infective doses (TCID50) of FMDV (O_1BFS 1860). The full-length FMDV genome sequence of this inoculum had previously been determined using Sanger sequencing (GenBank accession number EU448369). In addition, NGS data for selected samples originating from A1 have been previously described (Chapter 3). Twenty-four hours post needle-challenge, calf 1 (A1) was used to challenge naïve calf 2 (A2) by direct contact for a total of 4 days (transmission period 1 [T1] in the scheme in Figure 5.1). A1 was then removed from the experiment, and A2 was used to challenge naïve calf 3 (A3) by direct contact for 24 hrs (T2 in Figure 5.1). Following challenge, A2 was removed from the experiment. Successively, A3 was placed into direct contact with naïve calf 5 (A5) to be housed together until study termination (T3 in Figure 5.1). Calf 4 (A4) was an indirect contact challenge animal and therefore did not form part of the transmission chain and so was not included in this analysis. Sequenced samples are indicated in Figure 5.1.

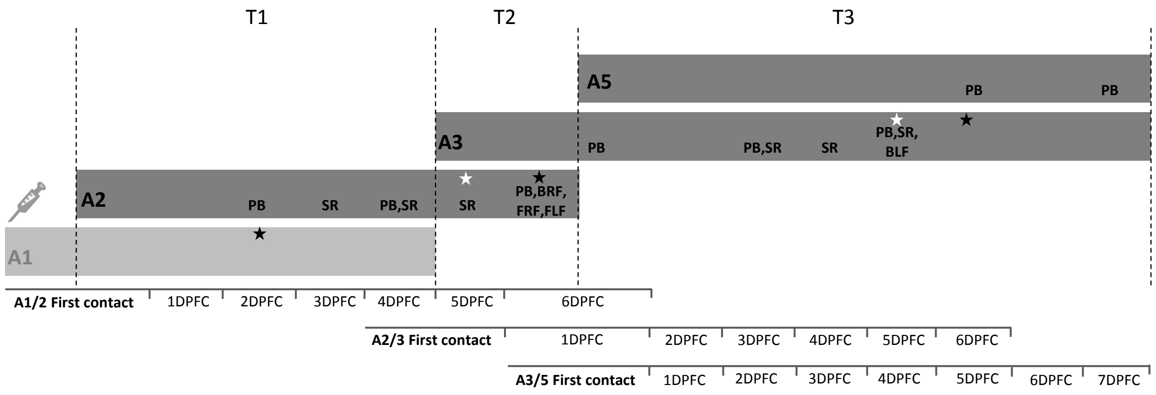


Figure 5.1

Temporal scheme of the transmission chain between calves 1 to 5 (A1, A2, A3 and A5) with the three transmission events (T1 to T3) indicated. Calf 4 (A4) did not form part of this analysis. Although calf 1 (A1) is included, only the 18 samples analysed here from calf 2 to 5 (A2, A3 and A5) are shown (serum [SR]; probang [PB]; front left foot [FLF] lesion; front right foot [FRF] lesion; back right foot [BRF] lesion). One timeline for each transmission event is indicated, where days post first contact (DPFC) applies to the naïve calf in that transmission event. A five-pointed black star indicates when lesions appeared on all four feet and the equivalent white star indicates when the first foot lesions appeared on the FR and BL for both calf 2 (A2) and calf 3 [A3]).

The sample types analysed here include blood serum (SR), oesophageal-pharyngeal scraping ('probang', PB) and foot lesion epithelium samples, indicated as *XY*F, where *X*= {B,F} for Back and Front, and *Y*= {L,R} for Left and Right, and F for Foot. The nomenclature for these samples followed the notation A*n*-*m*DPFC-*Z*, where *n*={2,3,5} represented the animal number in the chain, *m* was the number of days post first contact (DPFC) with an infected host for that particular animal, and *Z* was the sample type: for example, A2-4DPFC-SR corresponds to a serum sample taken from calf 2, 4 days after first contact with an infected host. Blood serum samples were taken daily and probang samples every other day. Foot lesion epithelium samples were collected within 24 hrs of first appearance. Daily rectal temperatures were monitored and clinical signs were defined here as any visible lesion or body temperature above 39.5°C.

5.3.2 Genome amplification

Total RNA was extracted (TRIzol, Invitrogen, Paisley, UK) from all biological samples collected from the experiment described above and quantified. Quantitative reverse-transcription polymerase chain reaction (qRT-PCR) was performed for quantification of FMDV genome copies in each of the samples, using an assay which can detect all serotypes of FMDV, as described previously (Callahan, Brown et al. 2002). rRT-PCR assays were performed on a Stratagene Mx3005P machine (Agilent Technologies, UK). For the generation of standard curves, a FMDV RNA standard was synthesized *in vitro* from a plasmid containing a 950 base pair insert from the 3D region of FMDV O/KUW/4/97 using a MEGAScript T7 kit (Ambion, UK) as described previously (Quan, Murphy et al. 2004).

FMDV concentrations in each of the samples (A2-A5) were normalized to 10^6 copies of FMDV RNA/µl prior to RT-PCR amplification (for reasons discussed in Chapter 4, section 4.3). Two genome fragments were amplified using a protocol modified from that previously described (Chapter 3). Briefly, two independent reverse transcription reactions were performed for each sample. An enzyme with high fidelity (Superscript III reverse transcriptase, Invitrogen) was used in each reaction plus two FMDV specific primers (see Table 5.1) in order to reduce RT-introduced artefacts and the risk of amplification bias (as described in Chapter 4, section 4.3).

Table 5.1 Oligonucleotide primers used for the amplification of the two large, overlapping FMDV genome fragments for both replicates. The fragments have the 5′ UTR S fragments omitted, up to and including the poly(C) tract, and overlap by 470 bp

PCR Set	Primer[1]	Primer Sequence (5' to 3')	Location on Genome[2]	Amplicon Size (bp)
1	OBFS-516+F	CCTTCGCTCGGAAGTAAAACGA	499-520	4065
	OBFS 4563 R	CCCGCTGCTTTTCAAGGAT	4545-4563	
2	OBFS 4094 F	TCTCGACGAGGCCAAACC	4094-4111	4033
	OBFS 8126 R	CTCCTAAGGTGTCGCGCG	8109-8126	
RT Set	Primer[a]	Primer Sequence (5' to 3')	Location on Genome	-
1	OBFS 8193 R	TTTTTTTTTTTTTTGATTAAGG	8155-8176	
2	OBFS 4926 R	AAGTCCTTGCCGTCAGGGT	4908-4926	

[1] Last letter indicates a forward or reverse primer

[2] Numbering according to Genbank sequence EU448369

For each of these replicas, two PCR reactions generating long overlapping fragments (4065 bp and 4033 bp respectively) were carried out using a proof-reading enzyme mixture (Platinum Taq Hi-Fidelity, Invitrogen). For biosecurity reasons these individual fragments comprised <80% of the complete FMDV genome, and corresponded to nts 499-4563 and 4094-8126 of EU448369 (see Table 5.1 for PCR fragment and primer details). This enabled the amplified DNA to be transported outside of the high containment FMD laboratory for sequencing.

The samples were amplified using the following cycling programme: 94 ˚C (5 min), followed by 94 ˚C (30 s), 60 ˚C (30 s) and 72 ˚C (4 min) for 39 cycles, with a final step of 72 ˚C for 7 min. Where a sample fell within half a log below the 10^6 copies of FMDV RNA/µl, neat sample was processed and sent for sequencing as long as it still yielded at least 700 ng of PCR product, samples below this threshold were not sequenced as indicated in Figure 5.2.

Two additional RT reactions were performed to yield enough cDNA to perform complete genome consensus sequencing (CGCS), via the Sanger method, on seven of the samples from calf 2 (A2) for validation purposes. The RT method was modified from that used in (Cottam, Haydon et al. 2006) and the PCR method from that used in (Cottam, Wadsworth et al. 2008a) (as described in Chapter 2, Section 2.3.3). These samples included the following: A2-2DPFC-PB, A2-2DPFC-SR, A2-4DPFC-PB, A2-4DPFC-SR, A2-6DPFC-PB, A2-6DPFC-FLF, and A2-6DPFC-FRF.

5.3.3 Illumina sequencing

The independently amplified replicates of each sample were sequenced with the Genome Analyzer IIx (Illumina) maintained by Glasgow Polyomics facility at the University of Glasgow, according to the protocol as detailed in (Chapter 3). Following the temporal order in the transmission chain, the first 12 samples were multiplexed on the same lane, while the corresponding duplicates were sequenced on a second lane, and ran on a different flow cell. The last 6 samples were multiplexed together on a lane belonging to a third flow cell. The 6 corresponding duplicates were multiplexed on a separate lane on the same flow cell.

5.3.4 Sanger sequencing

The protocol used for CGCS has been described (Cottam, Wadsworth et al. 2008a). However, sequencing reactions were performed using the Applied Biosystems BigDye Terminator V3.1 Cycle Sequencing Kit and an ABI 3730 genetic analyser.

5.3.5 Filtering and alignment

Single-end reads were 70 nt long for the first 12 samples, and 73nt long for the last 6. As commonly encountered with Illumina technology, the reads displayed a loss of quality for the nucleotides (nts) added last. A few reads were corrupted, with low quality throughout the whole length. Reads with unresolved nts or corrupted tags were removed from the analysis. We filtered the reads, removing any with an average probability of error per nt greater than 0.1% (probability of errors can be readily obtained from Illumina quality scores with the relation $p = 1/(1+10^{Q/10})$, where Q is the quality score and *p* is the probability of error). We observed that the same strategy removed about 20% of the reads for the first 12 samples, but over 30% for the last 6 samples (precise quantification can be found in the Supporting Table S1 in Appendix 4). Moreover, we trimmed the reads to 65 nt for the first 12 samples, and to 70 nt for the last 6.

The filtered, trimmed reads were aligned to FMDV genome O_1BFS1860 (the consensus sequence for the inoculum used to initiate the transmission chain) with a simple, custom-made scoring algorithm. No reads aligned ambiguously. For all subsequent analyses, we further trimmed the first and last 5 nts of each aligned reads, as they showed a higher number of mismatches to the reference sequence due to insertions or deletions close to the edges of the reads (Chapter 3), and we masked all nts whose individual probability of error was higher than 10^{-3} (corresponding to quality scores of 30 or lower). Primer regions (detailed in Table 5.1) were excluded from the analysis. Consensus sequences were always found to be identical between the two replicates for each sample. The genealogical relationships between consensus genomes were computed with the software package TCS (Clement, Posada et al. 2000) and reflected the most parsimonious genealogy. A schematic description of the steps in the analysis pipeline can be found in the Supplementary Table S2 (Appendix 4).

5.3.6 Validation of low-frequency polymorphisms

The frequency of a polymorphism at a particular position in the genome in a viral population was defined as the frequency of mismatches in the aligned reads relative to the consensus of the inoculum (GenBank accession no. EU448369). A proportion of these mismatches were expected to be artefacts, arising from base mis-callings in the sequencing process. In order to distinguish between real and artefactual variation, we extended the validation method described in (Chapter 3), summarized below. Under the assumption of independence, sequencing errors are binomially distributed, with the probability of observing xi or more mismatches given by Binom(xi; pi/3, ni), where xi is the number of nts bearing the most abundant mutation at site i, ni is the coverage, pi is the error probability computed from base qualities, and pi/3 represents the probability of the specific mutation observed in the reads. A score for site i was obtained, defined as si=1-Binom(xi; pi/3, ni). We defined si,1 to be the score obtained for the first replicate of the sample, and si,2 the score obtained for the second replicate. Only sites where the most frequent mutation was the same in the two replicates, and where si,1< θ and si,2<θ, with θ being a threshold chosen to be >0.05, were validated and used for successive analyses. Finally, in order to minimize artefacts introduced through RT and PCR error, we considered only mutations at frequencies above 0.5% (choice based on the analysis of control data generated using a FMDV cDNA clone, as described in Chapter 4, section 4.4). The second most abundant mismatched nt exceeded 0.5% in both replicates at only 1 site across the 18 samples so we focus here only on the most abundant mismatches.

From each alignment we constructed the ‘mutation spectrum’ which we define as a profile generated by the number of sites (y-axis) with a mismatch frequency of *x* (x suitably ‘binned’ on the x-axis). This was viewed as a log-log plot.

5.3.7 Genetic distance, entropy and dN/dS

Let $f_{i,A}$ be the frequency of the most abundant polymorphism at position *i* in sample A, obtained as a weighted average of the two replicates {1,2}:

$$f_{i,A} = (f_{i,A,1} * n_{i,1} + f_{i,A,2} * n_{i,2})/(n_{i,1} + n_{1,2})$$

where $n_{i,1}$ is the coverage of site *i* in the first replicate, and similarly for $n_{i,2}$. Genetic distance between two samples A and B was computed with a population-wide measure:

$$d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_{i,A} - f_{i,B})^2}$$

where *N* is the length of the sequence. Distances between samples were illustrated with a reduction to a two-dimensional space with classic (metric) multi-dimensional scaling, as implemented in the R software package; with this method, the distances between the points on the graph approximate the dissimilarities between the viral populations. Similarly, the complexity of the viral populations was characterized by computing their Shannon entropy at each site, and then averaging over every site in the sequenced genome for sample A:

$$S_A = \frac{1}{N}\sum_{i=1}^{N}[f_{i,A}\ln f_{i,A} + (1 - f_{i,A})\ln(1 - f_{i,A})].$$

The genome-wide entropy measures the amount of "disorder" in the population, and it is maximum when all sites have perfectly balanced polymorphisms (i.e. $f_{i,A}$=0.5 for all *i*). In order to estimate the synonymous to non-synonymous ratio dN/dS, for each codon *i* in the ORF, we first computed the expected number of synonymous (s_i) and non-synonymous (n_i) sites. Then, for each read *j* covering entirely codon *i*, we counted the number of observed synonymous (s^O_{ij}) and non-synonymous (n^O_{ij}) substitutions with respect to the consensus sequence of the inoculum. Using all codons where s_i>0 and $\sum_j s_{ij}^O > 0$, we obtained an estimate for the number of synonymous substitutions per synonymous site, p_S, and for the number of non-synonymous substitutions per non-synonymous site, p_N, using the following equation:

$$p_S = \frac{1}{n_{cod}}\sum_{i=1}^{n_{cod}}\frac{1}{r_i}\sum_{j=1}^{r_i}\frac{s_{ij}^O}{s_i}$$

where n_{cod} is the number of codons where the conditions above are met and r_i is the number of reads spanning entirely codon *i*. p_N was determined analogously. dN/dS was determined from p_N and p_S as described in (Nei and Gojobori 1986).

5.3.8 Complete genome consensus sequence analysis

Raw sequence data from the seven validation samples were assembled using SeqMan Pro™ 10.0.1 (DNASTAR, Madison, WI) followed by BioEdit v7.1.3.0 (Hall 1999) for all subsequent sequence comparisons (as described in Chapter 2, section 2.3.5).

5.4 Results

In this section, we discuss the results of the Illumina sequencing of 18 FMDV positive samples, distributed as follows: 9 from A2, 7 from A3 and 2 for A5. As the progenitor of this transmission chain, 2 samples from A1 plus the original inoculum (derived from a bovine tongue vesicle that had been extensively passaged in cell culture and used to artificially infect A1), previously described in (Chapter 3), were also included in analyses and discussed where appropriate. Short read data for all 18 samples were submitted to the EBI short read archive (SRA), which will be available from 01.05.13.

5.4.1 Quantification of viral titres

FMDV genome copies quantified by rRT-PCR of all the samples collected from A1-A5 (including the 18 samples analyzed in this study by NGS) are shown in Figure 5.2 (A-D). During these early stages of disease higher concentrations of viral RNA were measured in probang samples compared to serum samples. Viraemia*,* at 1-2 days post first contact, coincided with the clinical phase of disease. For A2 and A3 this correlated with the onset of fever and lasted up to 6 days after first contact with an infected host. As a consequence of being needle inoculated, the clinical phase of disease in A1 was shorter than that seen in subsequent animals. Conversely, the clinical phase of disease in A5 appeared elongated and less pronounced, as demonstrated by epithelial lesions not appearing on the feet until 8 and 9 days post first contact (not sequenced), as well as reduced fever and vireamia.

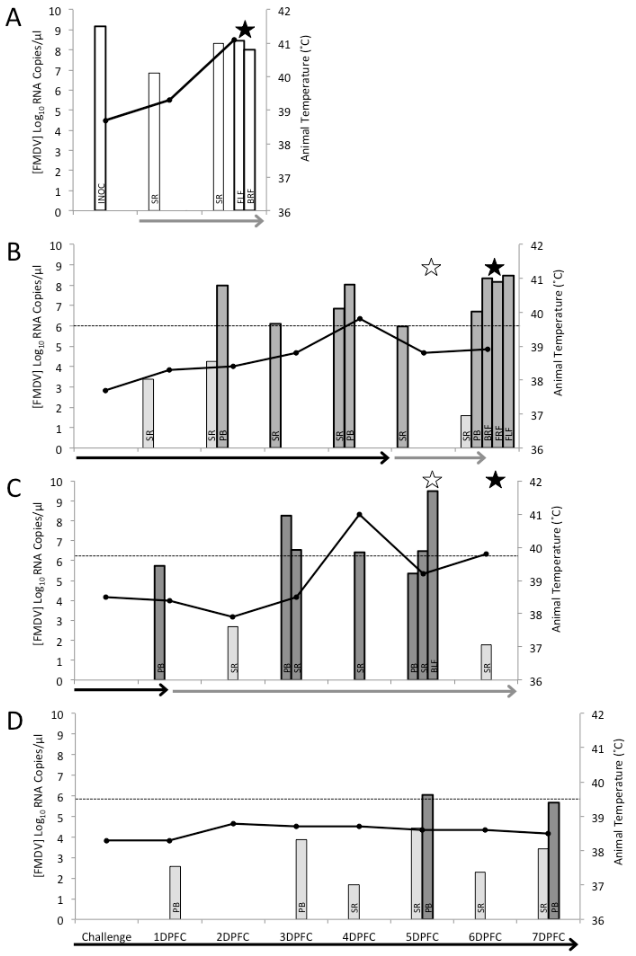


Figure 5.2

Quantification of viral RNA copy number and clinical signs (temperature) of infected hosts. FMDV RNA load in all samples collected during the serial passage of FMDV through four calves, detected by quantitative reverse-transcription polymerase chain reaction (qRT-PCR). Graph A-D, calf 1, 2, 3 and 5 (A1, A2, A3 and A5) respectively. A (A1) previously discussed in Chapter 3, sequenced samples in white with thick border and non-sequenced samples in white; B-D (A2, A3 and A5), sequenced samples in dark grey with thick border and non-sequenced samples in light grey. Inoculum (Inoc [A1 only]); serum (SR); probang (PB); front left foot (FLF) lesions; front right foot (FRF) lesions; back left foot (BLF) lesion; back right foot (BRF) lesion. Dashed lines indicate the minimum initial viral load to be amplified (106 copies of FMDV RNA/µl of sample) for A2, A3 and A5. Grey arrows indicate the time the calf spent in contact with the next calf, while black arrows indicate the time spent in contact with the previous calf in the transmission chain. Animal temperatures are shown on the same graphs (black solid line). White stars indicate the day when the first foot lesions appeared (FRF and BLF for both A2 and A3 [note, only BLF material from A3 was available to perform qRT-PCR]), while black stars indicate the day at which lesions appeared on all four feet.

5.4.2 Coverage and consensus genomes

Reads that passed the quality test were aligned to the consensus genome sequence of the original inoculum (FMDV strain O_1BFS1860). The coverage of the different samples were influenced by the different multiplexing of the Illumina lanes, and ranged from 11605x (A2-4DPFC-PB, first replicate) to 32208x (A3-5DPFC-BLF, second replicate); precise figures can be found in the Supporting Table S1 (Appendix 4). We computed the average frequency, for each mutation, that was weighted on the coverage received in the two replicates of each sample. We define consensus-level mutations as polymorphisms that appeared in more than 50% of this weighted average, with respect to the original inoculum.

We found a total of 13 consensus-level mutations across the samples in calves A2-A5, summarized in Table 5.2. The Sanger and NGS method both identified the same nine sites containing high frequency mutations within the same seven samples analysed; five mutated sites were identified at the consensus level by both methods. The remaining four mutated sites were identified at consensus level by NGS but remained as mixed populations (ambiguous sites) by Sanger sequencing (as detailed in Table 5.2). Figure 5.3 shows example chromatogram for two sites exhibiting mixed populations.

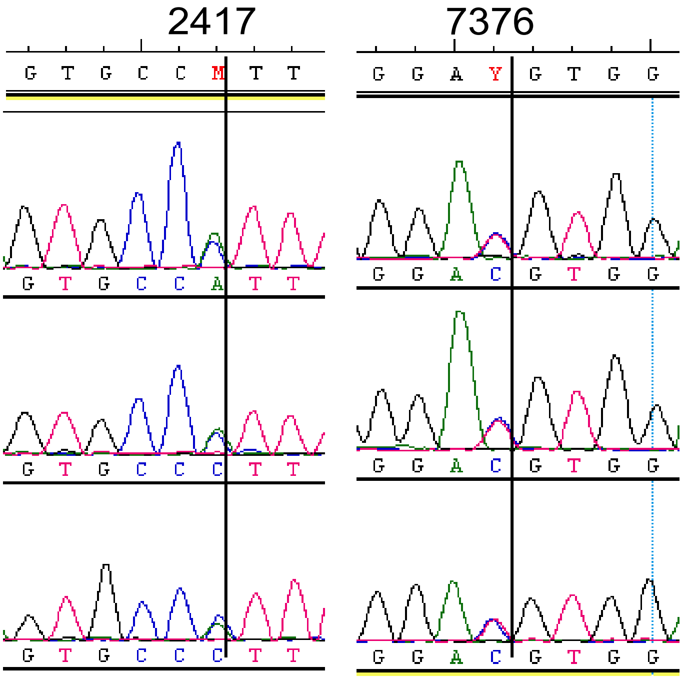


Figure 5.3

Chromatograms showing mixed populations within Sanger sequences for sample A2-6DPFC-PB at site 2417 and 7376 (numbering according to GenBank sequence EU448369)

Table 5.2 Consensus-level mutations, and their characterization. The mutation frequency is the weighted average frequency over the two sequencing runs for each sample

Position	Mutation	Frequency in sample	Gene	Syn/ Nonsyn[1]	Ts/ Tv[2]	Codon position	Sample[3]
1087	C->T	54.4%	Leader	N: T->I^{362}	Ts	2	A2-2DPFC-PB[4]
1164	A->G	63.9%	Leader	N: K->E^{388}	Ts	1	A2-6DPFC-P[5]
2417	C->A	51.1%	VP2	S: P->P	Tv	3	A2-6DPFC-PB[5]
		52.8%					A3-3DPFC-PB
2754	C->T	> 60%	VP3	N: R->C^{918}	Ts	1	ALL BUT A1[6]
2767	G->A	64.1%	VP3	N: G->D^{922}	Ts	2	A1-2DPFC-FLF
2768	C->T	52.8%	VP3	S: G->G	Ts	3	A3-3DPFC-PB
5435	C->T	> 55%	3A	S: G->G	Ts	3	ALL BUT[7] A1 & A2-2DPFC-PB A2-4DPFC-SR A2-5DPFC-SR A2-6DPFC-BRF A2-6DPFC-PB A3-3DPFC-PB
5669	T->A	99.0%	3A	S: L->L	Ts	3	A2-6DPFC-FLF[4]
5933	A->G	50.4%	3B2	S: K->K	Ts	3	A5-7DPFC-PB
6065	C->T	56.2%	3C	S: G->G	Ts	3	A3-1DPFC-PB
		99.7%					A3-3DPFC-SR
		99.3%					A3-4DPFC-SR
		75.6%					A3-5DPFC-PB
		99.6%					A3-5DPFC-SR
		99.7%					A3-5DPFC-BLF
		93.8%					A5-5DPFC-PB
		99.9%					A5-7DPFC-PB
6167	C->T	77.0%	3C	S: F->F	Ts	3	A2-6DPFC-FRF[4]
7355	C->A	58.0%	3D	S: A->A	Tv	3	A2-2DPFC-PB[5]
7376	T->C	54.4%	3D	S: D->D	Ts	3	A2-3DPFC-SR[5]
		68.5%					A2-6DPFC-PB[5]
		54.6%					A3-3DPFC-PB
7964	T->C	96.6%	3D	S: S->S	Ts	3	A3-3DPFC-SR
		97.6%					A3-4DPFC-SR
		53.4%					A3-5DPFC-PB
		99.1%					A3-5DPFC-SR
		99.9%					A3-5DPFC-BLF
		91.2%					A5-5DPFC-PB
		99.8%					A5-7DPFC-PB

[1] Synonymous or Non-synonymous mutation with associated amino acid change and position

[2] Transition or Transversion

[3] Sample notation as described in the Methods section

[4] Mutated site identified by both the Sanger method and NGS at consensus level

[5] Visible as mixed population via the Sanger method but identified at consensus level by NGS

[6] All 9 mutated sites

[7] Mutated site identified at consensus level by both the Sanger method and NGS for A2- 6DPFC-FLF and –FRF.

Previous analysis of the samples collected from the inoculated calf A1 (Chapter 3) identified one consensus-level mutation at position 2767, unobserved at this level in subsequent animals. Furthermore, two additional consensus-level mutations found in calf A1 in the 3' UTR region (position 8134 and 8140) could not be followed in this study, as the modified RT-PCR fragments ended at position 8126 (omitting 36 nt of the 3' UTR). Among the 13 mutations, one was present in every sample (site 2754, C->T). This mutation changes an amino acid residue in capsid protein VP3[56] associated with heparan sulphate (HS) binding, as does position 2767 in A1 (Chapter 3): the inoculum used in this experiment had undergone extensive cell culture passage and, in common with other in-vitro adapted viruses, utilizes HS as a cellular receptor (Sa-Carvalho, Rieder et al. 1997a; Fry, Lea et al. 1999a). Subsequent replication in mammalian hosts drives the reversion of positively charged amino acid residues at specific sites in the viral capsid, which is then fixed in the host chain.

Seven of the 13 mutations appeared only once across the samples (see Table 5.2), while several mutations (positions 5435, 6065, 7964) were fixed in the populations after the second transmission event. All the mutations appeared in the coding region of the genome: the majority were transitions (11/13), synonymous (10/13) and appeared at third codon positions (10/13), suggesting that most of these mutations did not confer any obvious selective advantage to the virus, but were likely close to neutral and subject to drift in the populations. When mutations were close enough to be spanned by a single read, we could check their co-occurrence (linkage): in the case of sites 2754 and 2768 in A3-3DPFC-PB, almost all the reads showed the former, but only half the latter, suggesting the co-circulation of two different viral variants, one of them acquiring the second mutation later in time. Moreover, two samples showing mutations at position 7376 (A2-3DPFC-SR and A2-6DPFC-PB) also exhibited a number of reads showing a mutation at position 7355 (~12% and 1% respectively), but almost no reads showed both sites mutated. We also interpret this finding as demonstrating the co-circulation of two different variants in the population, with two alternative mutations.

Figure 5.4 depicts the genealogy of the samples, based on statistical parsimony analysis of consensus sequences, and obtained with the software package TCS (Clement, Posada et al. 2000 961): the consensus genomes of the samples are very similar, as they were obtained within short time intervals of each other. Three samples in A2 have identical consensus genomes, and the same situation is found for 4 late samples in A3. The consensus of these samples further coincides with one sample in A5. The network shows accumulation of mutations through the chain, yet the structure is not simple: every host harbors multiple populations differing at one or more sites, and samples obtained from different hosts often fail to segregate (they can even display the same consensus, as discussed above). At the consensus level, the network appears to show several evolutionary "dead-ends", i.e. mutations that did not transmit further down the chain.

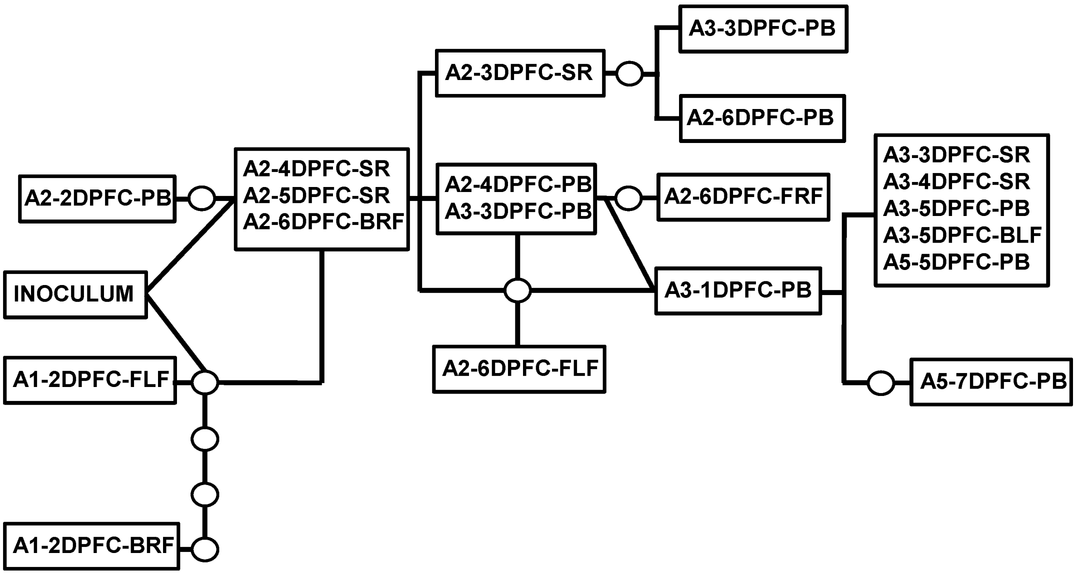


Figure 5.4
Genetic network of the samples based on statistical parsimony and obtained with the software TCS (Clement, Posada et al. 2000).

Finally, we saw no evidence at the consensus level of mutations within the non-structural genes that would suggest attenuation of the virus, as previously demonstrated during serial passage of FMDV in pigs (Carrillo, Lu et al. 2007a), to explain the observed elongated incubation period in calf A5. Although impacts on genome secondary structure cannot be ruled out with such data, due to lack of polymorphism linkage, this elongated incubation is more likely a result of reduced infective dose, indirectly indicated by the reduction in viral RNA copy number within samples from this host.

5.4.3 Sub-consensus mutations

Using the high coverage obtained with deep sequencing and the validation procedure described in the Methods section, we determined minority variants at each genomic site. First, we looked for the presence of the 13 consensus-level mutations in all samples (A2-A5). We found that many were present at sub-consensus levels in several samples. In particular, Figure 5.5 shows the “time series” of nine of these mutations, grouped by their respective dynamics: the first group includes mutations that were lost through the chain (sites 1087 and 7355, Figure. 5.6, top panel); we note that polymorphism frequencies at these two sites were tightly correlated along the chain, suggesting that they were both mutated on the same group of genomes. Other sites were mutated at sub-consensus levels only in some samples (Figure 5.5, middle panel): for example, site 2417 was found mutated in A3-5DPFC-PB and in A5-5DPFC-PB, but not in other late samples in A3. Finally, some mutations were fixed through the transmission chain (Figure 5.5, bottom panel). The dynamics of four additional consensus-level mutations are displayed in Supporting Figure S1 (Appendix 4), together with the single consensus-level mutation previously found in host A1 at site 2767. Supporting Figure S2 (Appendix 4) depicts the frequencies of the polymorphisms across the genome, for all the samples.

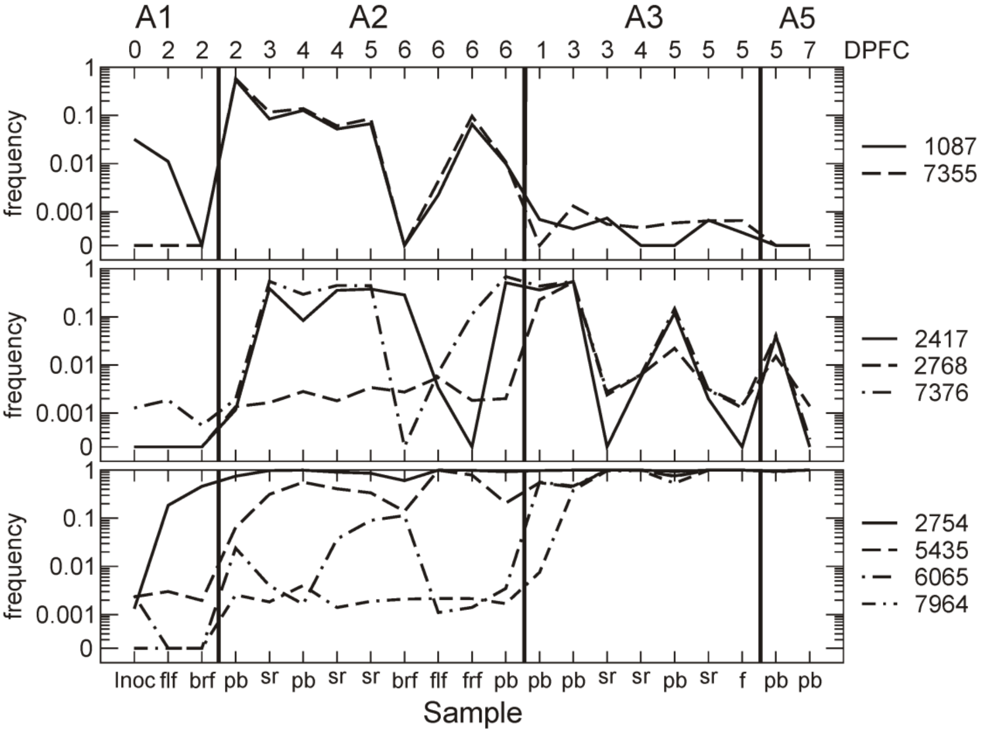


Figure 5.5

Frequencies across samples of 9 out of the 13 mutations reaching consensus in at least one sample, divided according to patterns. Top panel: Mutations present in A2 and then gradually lost in the next hosts. Middle panel: Mutations prevalently present in probang samples and sera, across all hosts. Bottom panel: Mutations reaching fixation.

Next, we obtained mutation spectra for all samples, defined as the collection of mutated sites, segregated into individual bins according to their frequencies. We computed the distance matrix between all the samples, displayed in Figure 5.6A: host boundaries are marked, although they did not always correspond to a sudden increase in the distance measures. In particular, early samples of A3 are more related to samples in A2 than to later samples in the same host. Late samples in A3, in turn, are very similar to samples in A5. Finally, samples like A2-6DPFC-FLF are very different from everything else, suggesting an independent evolutionary dynamic, which did not propagate through the infection chain.

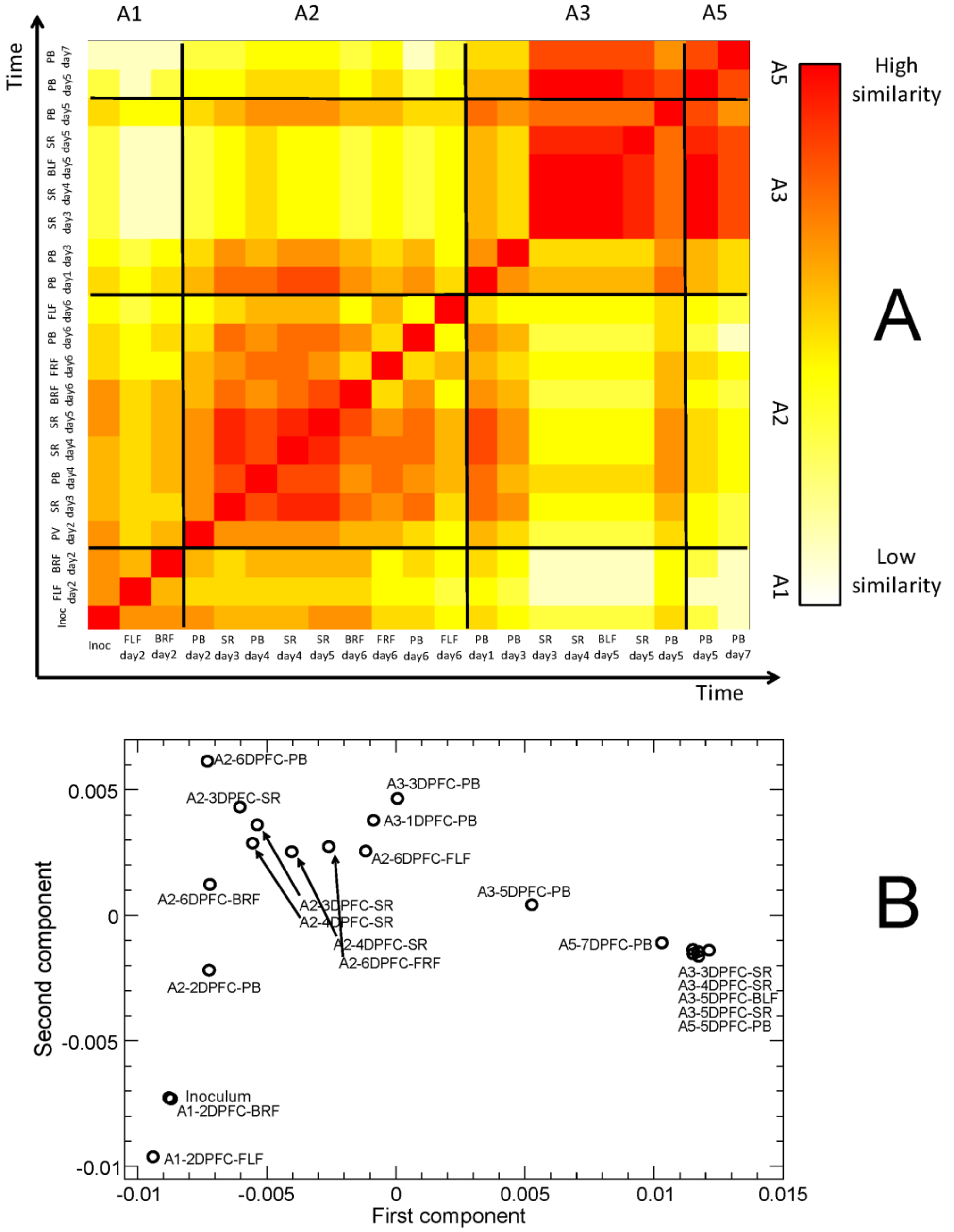


Figure 5.6

Panel A: Distances between viral populations collected in hosts A1-A5, obtained considering all validated mutations at frequencies above 0.5%. A2 presents a large heterogeneity, with the FLF samples being very different from all others. Conversely, A3 shows remarkably similar late samples, while the early probangs bear a larger similarity with the A2 samples. Samples in A5 are very similar to several late A3 samples. Panel B: Metric two-dimensional multidimensional scaling analysis of the distance matrix: the data formed the characteristic horseshoe pattern, sign of a latent order in the data.

The minimum distance between A3 and samples of A2 collected at 6DPFC is found between samples A2-6DPFC-FRF and A3-1DPFC-PB: based solely on this observation we would conclude that the viral population transmitted to A3 derived from the A2 FRF lesion. However, a closer inspection of the time series shows that the minimum distance between hosts A2 and A3 is found between A2-5DPFC-SR and A3-1DPFC-PB. Moreover, sample A3-1DPFC-PB has a comparable low distance from samples A2-4DPFC-SR, A2-4DPFC-PB and A2-3DPFC-SR. Finally, the presence of a consensus level mutation at site 6167 in A2-6DPFC-FRF, which was not found at any significant frequency in any A3 samples analysed here, reduces the probability that the transmitted viral population was seeded directly from this foot lesion. Considering all these observations, a possible scenario is that infection occurred around day 5 through a viral population originating from the upper oesophagus and pharynx of A2. Around the same time, other subpopulations originating in the palate seeded the feet lesions, where the virus underwent independent replication and diverged from the sample passed on to A3.

Moving on to the infection from A3 to A5, we noticed a more blurred situation: A5-5DPFC-PB was close to a number of A3 samples, including two serum samples, the back right foot lesion and, to a lesser extent, a late probang (the absolute minimum found with A3-3DPFC-SR). As samples are very similar to each other, resolution is limited and we cannot disprove either a direct infection route originating from a foot lesion in A3 or an infection originating from a population compatible to that found in the probang.

An easier visualization of the relationships between samples can be obtained with a standard metric multi-dimensional analysis in two dimensions, displayed in Figure 5.6B: the observed “horseshoe” pattern is typical of dimensionality reduction techniques, and is the sign of a latent ordering of the data, namely the accumulation of mutations along the transmission chain (Diaconis, Goel et al. 2008).

5.4.4 Inter-and intra-host bottlenecks

If a bottleneck is narrow (tight), only a few viral particles found a new population. Consequently, mutations included in the founding population will be likely fixed in the new population. Early replication cycles will introduce new variants which can spread in the population if neutral or advantageous. However, if the population is only recently founded, it is unlikely that these new mutations will reach significant frequencies. A population founded as a result of a narrow bottleneck could therefore be recognized by a depletion of sites with intermediate polymorphic frequencies in the mutation spectrum. Conversely, in the case of a wide bottleneck, the diversity of the founding population is a good representation of the diversity of the ancestral population, and we should then expect to see the mutations at intermediate frequencies well preserved in the new population. This criterion was used to qualitatively assess the size of the founding population in each of our samples. Here, we considered both intra-host bottlenecks (i.e. events leading to the founding of a new lesion in a distant epithelium) and inter-host bottlenecks (i.e. events leading to a host-to-host transmission).

Figure 5.7 displays the mutation spectra for all samples in calves A2-A5. We observed a typical pattern, depleted of intermediate frequencies, in a number of samples, and in particular in all the feet lesions in A2 and A3. This observation supports the hypothesis that these populations underwent a narrow intra-host bottleneck. The pattern is roughly U-shaped and originates from the combination of low-frequency mutations created in recent rounds of replication and mutations at consensus level, present in the founding population, and fixed by genetic drift. A3-1DPFC-PB, the earliest sample in A3, does not show this depletion, suggesting that the transmission to A3 arose as a result of the transfer of a sizable viral population from A2. A probang sample taken 5 days post first contact was the earliest sample to contain the minimum initial viral load of 10^6 copies of FMDV RNA/µl of sample from calf 5 (A5). A5-5DPFC-PB shows again the typical pattern corresponding to narrow bottlenecks. Unexpectedly, the viral population had not recovered sufficiently to demonstrate a full range of mutation frequencies 5 days post first contact; however, the elongated incubation period observed in A5, together with the observation that the calf showed no vireamia until 4DPFC, support the hypothesis of transmission to A5 through a severe bottleneck.

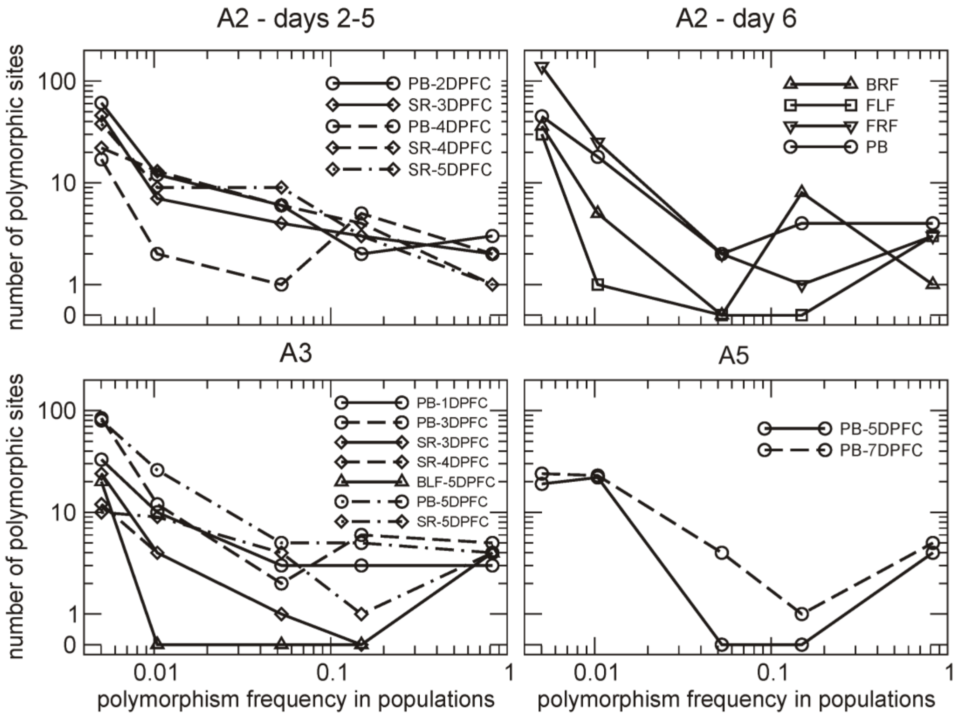


Figure 5.7

Mutation spectra representing the abundance of mutations at frequencies above 0.5% across the different samples: in some cases (typically probangs and sera) the mutation spectrum smoothly decreases in abundance as the frequency of mutations increases. However, in some samples (typically feet), the intermediate frequency region is depleted, suggesting narrow bottlenecks.

5.4.5 Entropy and dN/dS

Shannon entropy is a measure of the complexity of a population. Complexity can be acquired in two complementary ways: 1) through the presence of many low frequency polymorphic sites across the genome, where a nucleotide is largely dominant, and 2) through fewer but more balanced polymorphic sites where the nucleotides are equally represented.

Samples founded by a small initial population typically have not recovered from the loss of complexity associated with a narrow bottleneck (although vigorous replication could lead to high entropy through route 1). Conversely, samples founded by a large seeding population should display only a mild decrease in entropy. Figure 5.8A shows entropy for all the samples. The values fluctuate considerably: the lowest values are observed in the feet (host A2 and A3), reinforcing the hypothesis that these are “young” populations that have experienced a narrow bottleneck. However, the entropy of foot lesion A2-6DPFC-FRF is high: this value is reached through the very large number of polymorphic sites at frequencies around 0.5% found for this sample (see Figure 5.7) suggesting that this lesion was founded by a slightly larger population, and that early replication introduced numerous new mutations at low frequencies.

Early probang samples in A3 and the first probang in A5 available for sequencing show intermediate values of entropy. For A3, where the probang sample was taken only 1 day post first contact, the value observed, together with the absence of depletion in the mutation spectrum discussed above, supports the hypothesis that this complexity was inherited from an ancestral population through a wide bottleneck.

Finally, we computed the non-synonymous to synonymous ratio (dN/dS) for all the samples in this study (see Figure 5.8B). We found a monotonic reduction in dN/dS through the transmission chain, across all the samples collected from all tissues. While the values of dN/dS were close to 1 in A2, suggesting dominant random drift, it steadily decreases in A3 and A5, where the viral populations appear to undergo a continuous purifying selective pressure.

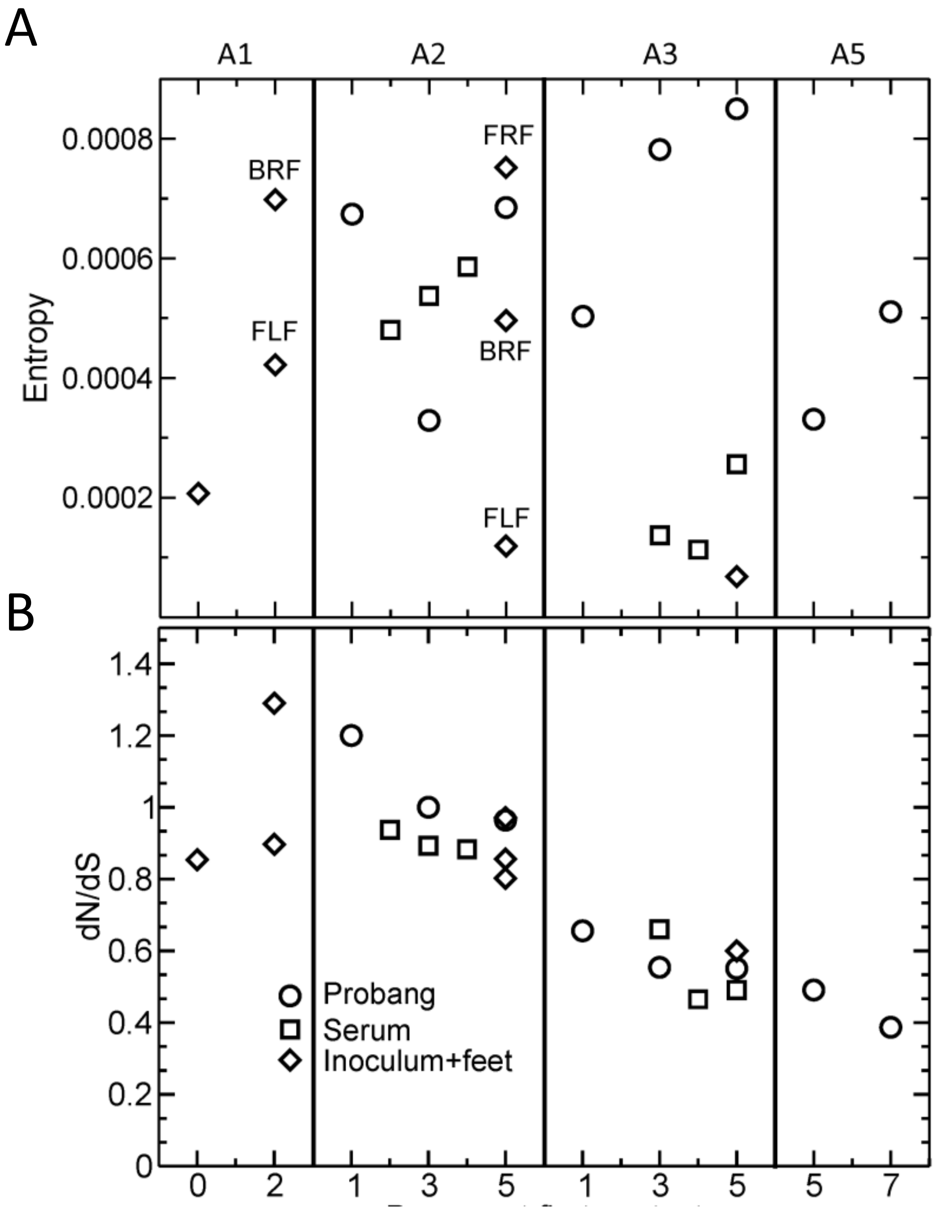


Figure 5.8

Shannon entropy (A) and dn/ds (B), across all samples, computed with validated mutations at frequencies above 0.5%. The complexity of viral populations fluctuates across samples, with lower values often found in correspondence of foot lesions. On the other hand, dn/ds ratios show a clear decreasing trend along the transmission chain.

5.5 Discussion

A total of 21 FMDV samples from a sequential infection experiment were analyzed using Illumina technology. These comprised 18 samples from calves A2-A5 and three samples (the inoculum and two foot samples from calf A1) discussed in (Chapter 3). There was good correlation between samples that were sequenced by both the Sanger and NGS method, NGS demonstrating improved sensitivity at sites identified as ambiguous by Sanger sequencing. No additional ambiguous sites were identified by Sanger sequencing. However, before interpretation of subsequent NGS viral population profiles, the genetic data should be put into context of the biological samples from which they were derived. For example, the different samples analysed are not anatomically equivalent. While foot lesions comprised a relatively spatially-discrete source of virus, probangs (oesophageal-pharyngeal scrapings) are thought to be composed of several infection foci (as well as those infected earliest), including the oesphagus, pharynx and oral cavity, and therefore are often more heterogeneous than samples taken from feet lesions. However, in addition to reconstructing a network at the level of viral consensus sequences, deep-sequencing has provided an unprecedented profile of the intra-sample evolution of the disease within and across hosts.

The consensus sequence network is informative and can be used to reconstruct the sequential accumulation of nt substitutions between hosts and provide evidence for the transmission of two separate viral populations from calf 2 (A2) to calf 3 (A3). However, this approach has limited resolution to differentiate between samples collected at the intra and inter-host scale as shown by the presence of identical consensus sequences within the same host (A2, 3 samples and A3, 4 samples) and between hosts (A2 and A3; A3 and A5). Deep sequencing allowed us to characterize samples collected from the transmission chain at a much deeper resolution than consensus sequencing. These data monitored low-frequency variation at specific sites in early samples prior to their appearance as consensus-level substitutions in later samples. For example, the advantageous mutation at genome position 2754 ($VP3^{56}$) related to the switch in receptor binding was rapidly fixed early in the transmission chain. NGS data also revealed patterns of apparently neutral mutations which were sometimes observed at lower frequencies but drifted over and under the consensus threshold through time. This study identified 13 consensus-level mutations (A2-A5) that were generated during the transmission chain. Of these: four were fixed at the level of the consensus by the end of the experiment; two were lost and a total of seven exhibited a 'drifting' pattern, appearing successively fixed and lost in the population. Additionally, the frequency profile of the two lost mutations (at sites 1087 and 7355) closely 'mirrored' each other so, although the short length of the NGS reads prevented the systematic reconstruction of viral haplotypes, linkage between mutations may be inferred by this tightly correlated pattern of mutation frequency over time. Moreover, by checking the linkage of significant mutations spaced on the genome less than the length of a single read, we were able to demonstrate that several viral genotypes can co-circulate in a lesion (as suggested by previous work (Cottam, King et al. 2009b)).

Investigation of the mutation spectra provided evidence for variation in the polymorphic structure of viral populations. In particular, we found indications of two types of founding events: intra-host, when the infection reaches a distant epithelium through the blood stream, and inter-host, when the infection is transmitted to the next host. In this experiment, several related lines of evidence point toward narrow bottlenecks during the process of virus dissemination during intra-host infections and a wider bottleneck for the inter-host transmissions. These include: 1) distances between viral populations were sometimes larger within hosts compared to between hosts; 2) the mutation spectra of populations sampled early during the infection of a host exhibited polymorphisms across a range of frequencies, while those of newly-formed lesions at the end of the clinical phase displayed a depletion of polymorphisms with intermediate frequencies; and 3) the Shannon entropy of populations did not drop substantially across hosts but was often low in samples recovered from 'younger' feet lesions. Where wide inter-host bottleneck transmissions have also been demonstrated in both Equine (Murcia, Baillie et al. 2010;) and avian-like swine (Murcia, Hughes et al. 2012) influenza virus, inter-host transmissions for HIV, for example, have been shown to be via extremely narrow bottlenecks of only a few particles (Fischer, Ganusov et al. 2010; Wang, Sherrill-Mix et al. 2010b; Bull, Luciani et al. 2011; Bull, Eden et al. 2012). In contrast to wide bottlenecks, where a more faithful representation of the diversity within a donor host is transmitted, extreme bottlenecks, such as those experienced during HIV transmission, further call into question whether there are biologically meaningful features of these few transmitted/founder viruses that facilitate their transmission.

The lack of clear boundaries between FMDV populations collected from different hosts highlights the importance of founder effects and subsequent tissue/organ-specific amplification during viral spread in an individual host. Following host entry and dissemination, a distant epithelial lesion can be considered as hosting a viral population that is relatively distinct from the systemic circulation, and founded by a subsample of an ancestral population. However, mixing of these populations through interchanges of viral particles via the blood stream may blur population ‘boundaries’ at finer scales. Nonetheless, the presence of large differences between populations within a single host suggests that the size of this founding population may be relatively small. Some populations detected across different hosts were surprisingly similar, both at the consensus and sub-consensus sequence level. This scenario is compatible with some host-to-host transmission events seeded by large viral populations, where a rather faithful representation of the diversity in the ancestor population is passed on to the next host. Analysis conducted with mutation spectra, at the host-to-host scale, showed a strong trend in dN/dS towards an increased purifying selective pressure along the chain. If a role for the adaptive immune response is ruled out so early in infection, we can hypothesize that the declining dN/dS ratio results from the elimination of mildly deleterious mutations generated early in the chain. We conclude that host-to-host transmissions can be seeded by populations of different sizes, while in all cases examined, seeding of a distant host epithelium lesion occurred via a small founding population.

In the present study, we considered only polymorphisms at frequencies higher than 0.5%. The coverage obtained by NGS allowed us to investigate lower frequencies, but at the likely price of introducing significant numbers of artefactual mutations into the analysis. Accordingly, we note that Shannon entropy was computed in (Chapter 3) for A1 samples in a slightly different manner: to avoid contamination by low-frequency artefactual mutations, we considered here only the contribution deriving from the dominant polymorphism at each site. The entropy of the original inoculum, computed according to the method used in this work then becomes 2.07x10^{-4}, while we obtain 4.22x10^{-4} and 6.98x10^{-4} for the A1 FLF and BRF lesions, respectively. These values are compatible with those found later in the transmission chain, confirming that a single host passage results in a cell-cultured population acquiring complexity equivalent to a natural *in vivo* infection. While polymorphisms at frequencies below 0.5% are unlikely to change the conclusions of the present study, a more comprehensive understanding of the population genetics of acute RNA virus infections will require quantifying polymorphic frequencies well below this threshold. Such understanding will require either direct high fidelity sequencing of RNA without amplification, or more detailed study and reduction of the errors introduced by the RT-PCR process and sequencing reactions themselves.

Taking multiple samples from the different hosts allowed us to see a host as a collection of potential sources of infection rather than harboring a single heterogeneous population. The different populations, while clearly related, can differ at several consensus positions (in accordance with previous studies (Carrillo, Lu et al. 2007a)), and showed different levels of heterogeneity, potentially caused either by tissue/organ-specific amplification or bottlenecking and founder effects during intra-host viral spread. While the ability to recognize a single lesion as a source of infection is limited to the samples available and by the extent of mixing between populations via the blood stream, characterizing multiple potential source populations is a clear advancement. This information could be a powerful tool to reconstruct more refined transmission trees and develop a more sophisticated understanding of how viral genetic differences accumulate with transmission events.

Chapter 6

The effects of sequential bottlenecks on foot-and-mouth disease virus population diversity in vitro

The analytical and statistical pipeline used within this chapter was as described in Chapter 5 (constructed by Dr Marco Morelli). Statistical analysis was performed by Dr Richard Orton (University of Glasgow).

6.1 Summary

This chapter describes the optimization and results of a novel experimental design, the aim of which was to further investigate the impact of bottleneck size on the acquisition and fixation of mutations within FMDV populations, characterised at the ultra-deep level using NGS. Rescued virus from a full-length FMDV cDNA clone was subjected to serial passage *in vitro* to better characterise viral population diversity generated from a more defined clonal starting material. NGS successfully demonstrated that mutation frequency in the population increases more rapidly during small population passages and provided evidence for positive selection during the passage of large populations. The novel experimental design described provides a potential resolution for such investigations negatively impacted by background sequence noise by use of an evolutionary *marker.*

6.2 Introduction

RNA viruses exist as swarms, or populations, of closely related viral genomes as a consequence of the poor proofreading ability of the viral RNA dependent RNA polymerase. The error rate of this polymerase is in the order of 10^{-3} to 10^{-5} misincorporations per nucleotide (nt) copied (Batschelet, Domingo et al. 1976; Drake 1993; Domingo and Holland 1997; Drake and Holland 1999), and confers on RNA virus populations a high degree of genetic heterogeneity, which is thought to favour adaptability to different environments. The serial passage of relatively large proportions of this heterogeneity (i.e. transferring large viral populations) is generally accompanied by an overall gain in population fitness within the environment in which replication takes place (Novella, Duarte et al. 1995; Escarmis, Davila et al. 1999) referenced in (Escarmis, Lazaro et al. 2008). Competitive optimization between different mutants within the viral swarm may explain this observation (Escarmis, Davila et al. 1999). Conversely, the serial passage of a small proportion of this heterogeneity (i.e. transferring small viral populations), by subjecting the viral population to sequential *bottlenecks,* results in reduced mutant spectrum diversity and has been demonstrated for a range of plant, animal and human RNA viruses (Li and Roossinck 2004; Ali, Li et al. 2006; Jridi, Martin et al. 2006; Murcia, Baillie et al. 2010; Boeras, Hraber et al. 2011). The serial bottleneck passage of small viral populations results in average fitness loss (Escarmis, Davila et al. 1999). Next-generation sequencing (NGS) has been used to provide initial evidence for bottleneck driven diversification of Norovirus populations (Bull, Eden et al. 2012), as well as for similar studies of Hepatitis C virus (Wang, Sherrill-Mix et al. 2010b; Bull, Luciani et al. 2011), and HIV-1 (Fischer, Ganusov et al. 2010) diversification, *in vivo,* as was discussed for FMDV in Chapter 5.

Viral fitness loss, associated with serial bottleneck events, may be explained by the predominant effect of genetic drift, where the probability of replicative optimization is restricted to competition between founding variants. As a result of genetic drift, mutations may become more rapidly fixed in small founding populations by chance and, if deleterious in nature, will often lead to the decline in replicative ability of that population. Bottleneck associated decline in viral population fitness has been demonstrated experimentally *in vitro*, where a range of

RNA viruses have been subjected to serial bottleneck events via plaque-to-plaque passages (Duarte, Clarke et al. 1992; Novella, Elena et al. 1995; Escarmis, Davila et al. 1996; Yuste, Sanchez-Palomino et al. 1999). Bottleneck size has been controlled experimentally by varying MOI, between 0.01-0.1 (low MOI) and 1-10 (high MOI) (Escarmis, Davila et al. 1999; Escarmis, Lazaro et al. 2008). This bottleneck effect was mimicked during serial contact transmission of FMDV in pigs (Carrillo, Lu et al. 2007a).

Similar cell-culture based, serial passage, experiments have been conducted to investigate the impact of viral load (in the form of MOI) on FMDV populations undergoing enhanced mutagenesis (Sierra, Davila et al. 2000; Moreno, Tejero et al. 2012). A study by (Sevilla, Ruiz-Jarabo et al. 1998) looked at the effect of MOI on the relative fitness of two competing viral subpopulations of FMDV *in vitro*. In this study, the subpopulation with increased affinity for *heparan sulphate* was found to out-compete the other viral subpopulation when passaged in BHK-21 cells at low MOIs.

The current study used NGS to dissect the evolutionary progression of rescued virus subjected to two different bottleneck regimes *in vitro*, at the ultra-deep level. The ultra-deep coverage provided by NGS reveals mutations present in only a small fraction of the population. Therefore, NGS may potentially provide information about the fine-scale impacts of these different bottleneck regimes on the viral swarm, which would have been missed by less high-throughput techniques such as cloning. The aim of this study was to test the hypothesis that mutations, irrespective of selective value, become more rapidly fixed in the population during more severe bottleneck transmissions.

6.3 Experimental design & optimization

Previous studies have conducted serial, extreme bottleneck transmissions as plaque-to-plaque transfers in which each infection is initiated by a single infectious particle. Such experiments have compared the pattern of mutations and their distribution along the FMDV genome of clones subjected to serial plaque-to-plaque transfers to those observed within FMDV clones subjected to serial large population passages (Escarmis, Davila et al. 1999; Escarmis, Lazaro et al. 2008). Other studies have looked at the impact of bottleneck size on the fitness of vesicular stomatitis virus during serial cell-culture passages. One such study varied the number of infectious particles sampled from the parental population during serial plaque-to-plaque transfers (Novella, Elena et al. 1995), another varied bottleneck size by using flasks of an appropriate size for each population size (Novella, Dutta et al. 2008).

Within this experimental design, MOI was also kept constant by varying the number of cells between the two passage series. Small viral populations were then sequentially transferred between an accordingly small number of cells and large viral populations between an accordingly large number of cells, as demonstrated in Figure 6.1. Therefore, serial passage of large viral populations would simulate serial *wide* bottleneck transmission and small viral populations, serial *narrow* bottleneck transmission (series ‘A’ and ‘B’ in Figure 6.1 respectively). Conducting serial cell-culture passage in liquid culture medium, rather than semi-solid culture medium, would allow the inclusion of the entire viral population at each passage, not just fit viral particles.

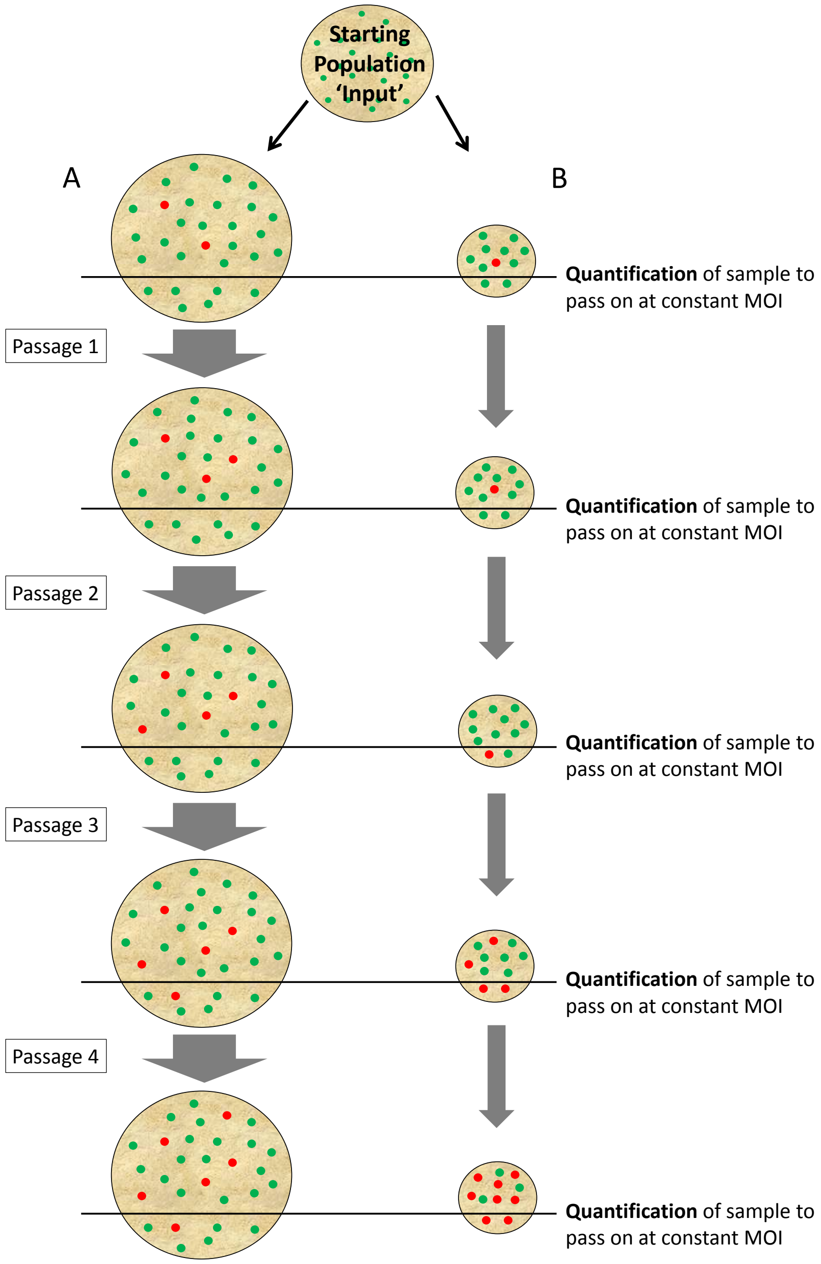

Starting Population 'Input'
A
B
Passage 1
Passage 2
Passage 3
Passage 4
Quantification of sample to pass on at constant MOI
Quantification of sample to pass on at constant MOI
Quantification of sample to pass on at constant MOI
Quantification of sample to pass on at constant MOI
Quantification of sample to pass on at constant MOI

Figure 6.1

Schematic depicting the theoretical principle behind the experimental design of this study. Green dots represent the viral swarm and red dots represent a specific mutation that either slowly increases in frequency within a *large* viral population (series A, *wide* bottleneck transmission) or, by chance, reaches fixation by virtue of being passaged from one *small* viral population to the next (series B, *narrow* bottleneck transmission).

The main experimental variables to be kept constant between the passage series were MOI and incubation period. In order to keep MOI constant, confluency of the cell monolayer needed to be consistent. To this end, passages were only conducted when both the 'Large' and 'Small' viral population (population L and S respectively) were at 100% confluency. An additional consideration was the relative homogeneity of the starting viral population with regards to observing the potential impact of bottleneck size, which was achieved by use of an infectious clone. Additionally, the minimum amount of virus propagation was conducted to infect population L at the highest of two MOIs tested (MOI of 1).

A major concern of this study was that a large number of passages would be required before mutation frequency reached levels above background sequence noise (set at 0.5%, as discussed in Chapter 4, section 4.4). Therefore, the virus and cell line used were chosen in order to induce some degree of selective pressure at known sites of biological significance associated with a switch in cellular receptor usage (see below). If appropriate selective pressure was exerted on the viral population to induce mutations at these sites, the hope was that these mutations may then reach fixation more rapidly. Such mutations would therefore act as a form of evolutionary *marker.*

An important factor which determines the infectivity of a number of animal viruses is the presence of suitable cellular surface receptors for attachment and internalization. The epithelial cell expressed heterodimer, integrin, has been shown to be the cellular receptor for FMDV *in vivo (Jackson, Clark et al. 2004; Monaghan, Gold et al. 2005; O'Donnell, Pacheco et al. 2009).* Although several integrins are known to bind to the conserved RGD amino acid motif found on the VP1 capsid protein of FMDV, including αvβ8 and αvβ3, the integrin αvβ6 is considered the main receptor of wild-type FMDV. However, Jackson *et al.* (1996) observed that the glycosaminoglycan, heparan sulphate (HS), could mediate the interaction of FMDV serotype O with cells in culture. Nine motifs, associated with the subtype O1 FMDV-HS receptor complex, namely residue 134, 135 and 138 of VP2, residues 56, 59, 60, 87 and 88 of VP3, and residue 195 of VP1 are discussed in (Fry, Lea et al. 1999a). Amino acid residue 56 of VP3, an arginine in cell-culture-adapted viruses and commonly a histidine or cysteine in ‘field’ strains of FMDV, is critical to virus/receptor recognition (Fry, Lea et al. 1999a; Borca, Pacheco et al. 2012). The HS-binding phenotype has been linked to attenuation of a serotype O genetically engineered virus in cattle (Sa-Carvalho, Rieder et al. 1997a). However, the progression of two amino acid reversions associated with the switch from a cell-culture to host adapted virus (HS to integrin receptor usage), was observed in cattle at the ultra-deep level using NGS, as discussed in Chapter 3. Although HS binding has been observed for cell-culture adapted FMDV serotype C viruses (Baranowski, Ruiz-Jarabo et al. 2000), serotype A viruses (Fry, Newman et al. 2005) and SAT-type viruses (Maree, Blignaut et al. 2010), it is not clear whether HS can be utilized by all serotypes of FMDV as an alternative cell receptor (Botner, Kakker et al. 2011).

Experimental design was consistent between the two passage series. Therefore, time, virus and reagent expenditure was minimised by conducting optimisation steps on a single population size, where appropriate. Total RNA extraction, RT, PCR and product visualization was as described in Chapter 4, section 4.4, including the same RT and PCR primers. As the FMDV genomic region of interest, in terms of selective pressure during cell-culture passage, was incorporated within fragment 1 (PCR1) amplification, this assay alone was performed. qRT-PCR was as described in Chapter 4, section 4.3.5.

6.3.1 Bottleneck size

Bottleneck size was controlled by varying the size of the viral population transferred at each cell-culture passage. However, to maintain constant MOI, the number of cells available for infection needed to be varied accordingly, which was achieved by using different sized polystyrene cell-culture vessels. The aim was to demonstrate the impact of bottleneck size on the mutation frequency distribution of a viral population. The priority was therefore to achieve as large a differential in bottleneck size as possible and, in order to do this at constant MOI, as large a differential between cell-culture vessels needed to be achieved. The two vessels tested provided an approximate 3 log differential (as detailed in Table 6.1).

Table 6.1 Cell-culture vessel details for population L and S

Population	Cell-culture vessel	Manufacturer	Surface area (cm^2)	Cell count (100% confluent monolayer)
L	T175 flask	Greiner Bio-One	175	2.0 x 10^{7} (1)
S	96-well (3)	NUNC	0.3	3.4 x 10^{4} (2)

(1) Cell count obtained from an average within this size vessel using a haemocytometer

(2) Cell count extrapolated from that given for the T175 flask according to surface area

(3) A single well of a 96 well, flat bottomed ELISA plate

6.3.2 FMDV full-length cDNA plasmids

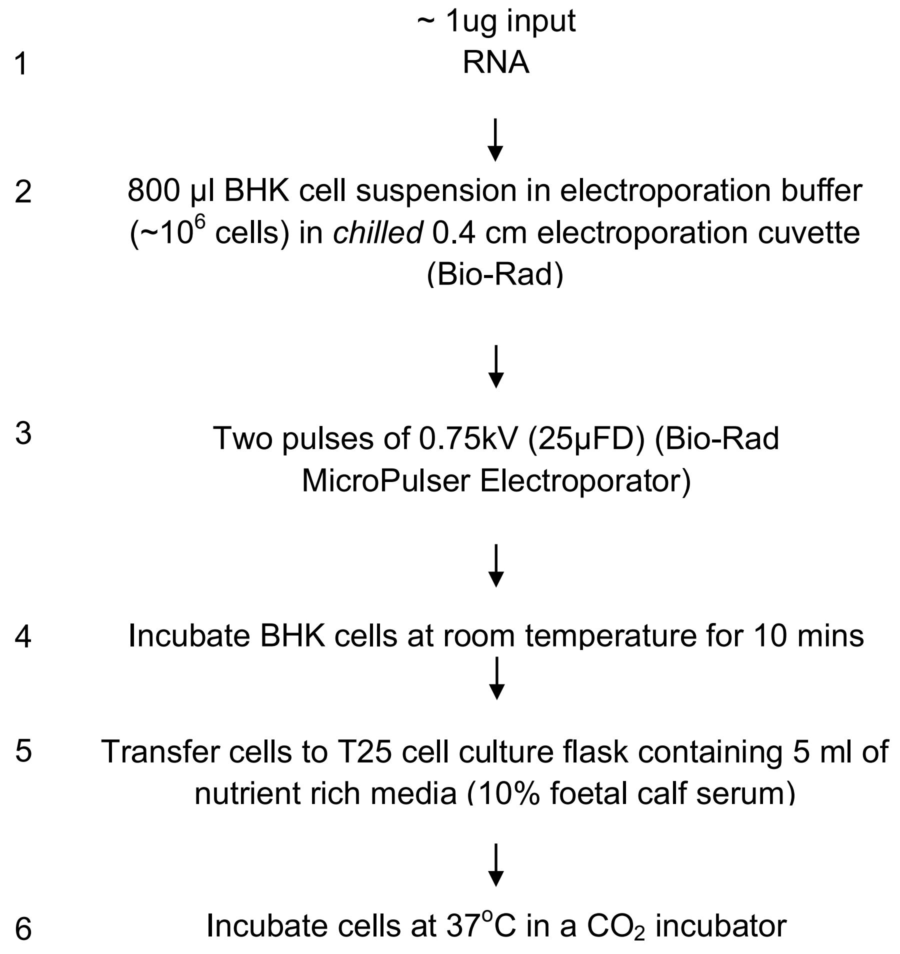
The virus studied had been rescued from a plasmid containing full-length FMDV O_1Kaufbeuren cDNA (pT7S3) (Ellard, Drew et al. 1999). This infectious copy was a cell-culture-adapted B64 strain of the O_1Kaufbeuren virus (O_1K B64) (Ellard, Drew et al. 1999), and therefore has the ability to bind HS, producing observable cytopathic effect (CPE) in both the first and second passage in BHK cells (Botner, Kakker et al. 2011). A study by Botner *et al.* (2011) found that the O_1K B64 infectious copy grew well in both BTY cells and a goat cell line (ZZ-R 127), producing complete CPE in each cases (Botner, Kakker et al. 2011). The pT7S3 plasmid, with its nucleotide sequence, was kindly provided by Veronica Fowler (IAH, Pirbright).

6.3.3 Cell-culture cell line

The ZZ-R 127 fetal goat tongue epithelium cell line was developed from the Friedrich-Loeffler-Institute (FLI) Collection of Cell Lines in Veterinary Medicine (CCLV). An evaluation study by (Brehm, Ferris et al. 2009) found that the sensitivity of ZZ-R 127 cells to infection by both wild-type and cell-adapted FMDV strains was only slightly inferior to that of primary BTY cells, the most sensitive cells for FMDV isolation. Preliminary studies by (Brehm, Ferris et al. 2009) indicated that > 90% of ZZ-R 127 cells expressed the αvβ6 integrin receptor, potentially explaining the higher sensitivity of this cell line to field strains of FMDV compared to other permanent cell lines, including BHK-21. However, in contrast to primary BTY cells, ZZ-R 127 cell lines maintain their polymorphic epithelium-like morphology and sensitivity to FMDV after multiple passages. The decision was therefore made to use the more stable ZZ-R 127 cell line for this study. Additionally, by using this cell line, genomic sites associated with the switch in cellular receptor usage could potentially provide a genetic marker in testing the study hypotheses. Using these sites in this way required the assumption that there may be some increasing level of selective pressure for the *cell-culture* adapted infectious copy to switch receptor usage to integrin with passage.

6.3.4 Rescue and growth of virus from full-length cDNA plasmids

In vitro transcription of full-length FMDV RNAs was achieved using the same protocol as described in Chapter 4, section 4.4.1. After quantification and checking transcribed RNAs by gel electrophoresis, the RNAs were introduced into BHK cells by electroporation, essentially as described previously (Nayak, Goodfellow et al. 2006) and depicted in Figure 6.2.



1 ~ 1ug input RNA

2 800 µl BHK cell suspension in electroporation buffer ($\sim10^6$ cells) in *chilled* 0.4 cm electroporation cuvette (Bio-Rad)

3 Two pulses of 0.75kV (25µFD) (Bio-Rad MicroPulser Electroporator)

4 Incubate BHK cells at room temperature for 10 mins

5 Transfer cells to T25 cell culture flask containing 5 ml of nutrient rich media (10% foetal calf serum)

6 Incubate cells at 37^{o}C in a CO_2 incubator

Figure 6.2

Schematic of the electroporation process for the rescue of O_1K B64 and O-UKG virus

An initial experiment was conducted in order to demonstrate the degradation of RNA left on the outside of cells by natural RNases with time (Figure 6.3). The same quantity of RNA was either i) introduced into cells and incubated at 37^{o}C for 6 hrs, as above, (RNA ‘Inside’ cells) or ii) introduced to cells *after* they had been electroporated and given time for the membrane pores to close so that RNA sat on the outside of cells at 37^{o}C for 6 hrs (RNA ‘Outside’ cells). Electroporated cells in both scenarios were added to 5 ml of nutrient rich media for incubation (as above). PCR1 amplification was performed, as described in Chapter 4, section 4.4. All bands were at the correct size (between 4 and 5 kb).

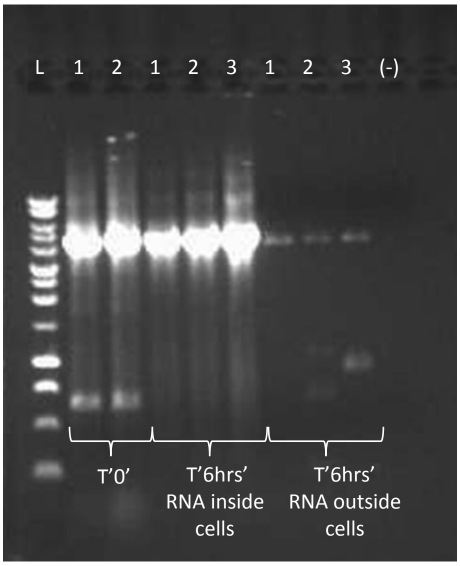


Figure 6.3

Agarose gel depicting PCR1 products for O_1K B64 rescued virus at time zero (T'0') in duplicate, followed by after 6 hrs incubation at 37°C inside BHK cells (T'6hrs' RNA inside cells) in triplicate, and after 6 hrs incubation at 37°C outside BHK cells (T'6hrs' RNA outside cells) in triplicate.

A time course was also conducted to demonstrate replication. Electroporated cells containing O_1K B64 RNAs were incubated for 2 hrs, 3hrs, 4 hrs, 5 hrs, 6 hrs, 8 hrs and 12 hrs. Following each incubation period, the appropriate T25 flask, containing media/cells and virus, was frozen at - 20°C for at least 2 hrs to detach and burst any adhered cells, creating a 'virus/cell' suspension. Two hundred microliters of this suspension was then added directly to 800 µl of TRIzol (Invitrogen, Paisley, UK) for total RNA extraction. PCR product visualization and corresponding quantification using a Nanodrop spectrophotometer (Figure 6.4a and b respectively) demonstrated an increase in product from 8 to 12 hrs. All 260/280 ratios were between 1.80 and 1.87. All bands were at the correct size (between 4 and 5 kb).

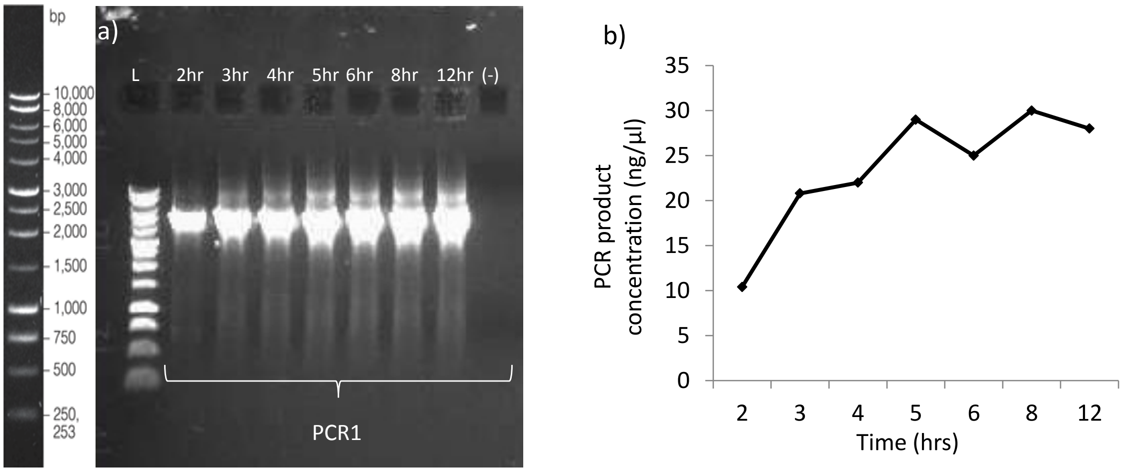


Figure 6.4

O_1K B64 time course results **a)** agarose gel depicting PCR1 product and **b)** graph showing PCR1 product yield (ng/µl), as quantified on a Nanodrop spectrophotometer, with trendline fitted.

In order to generate enough initial 'Input' virus to infect population L at the highest MOI tested (MOI 1), a virus titre of at least 2.0 x 10^7 plaque forming units (PFU)/ml needed to be achieved. This was done by overlaying the electroporated cells onto a fresh monolayer of BHK-21 cells for 24 hrs and conducting a single passage onto the same cell line until CPE was observed.

After growing sufficient virus in BHK cells, the actual study would be conducted using ZZ-R 127 cells therefore duplicate virus titres were measured by plaque assay using this cell line (method modified from (Dulbecco and Vogt 1954)). Briefly, infected cells were incubated at 37^{o}C, under a Noble agar (Sigma-Aldrich) overlay and stained with crystal violet at 72 hrs post infection. O_1K B64 ($VP3^{56}$-Arg) demonstrated a small plaque morphology (Figure 6.5a and b) as described previously (Borca, Pacheco et al. 2012). In total, O_1K B64 virus 'Input' had a total of 33.5 hrs of replication in BHK-21 cells (achieving a final average titre of 4.4 x 10^7 PFU/ml).

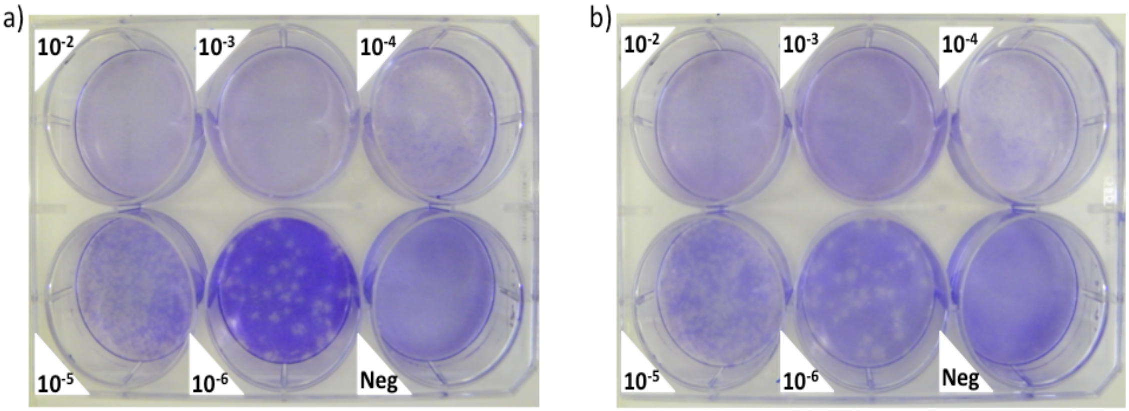


Figure 6.5

Plaque assay quantification of O_1K B64 rescued virus a) and b) represent duplicate plates. Plaques from appropriate dilutions are shown.

6.3.5 Measuring viral replication & infectivity

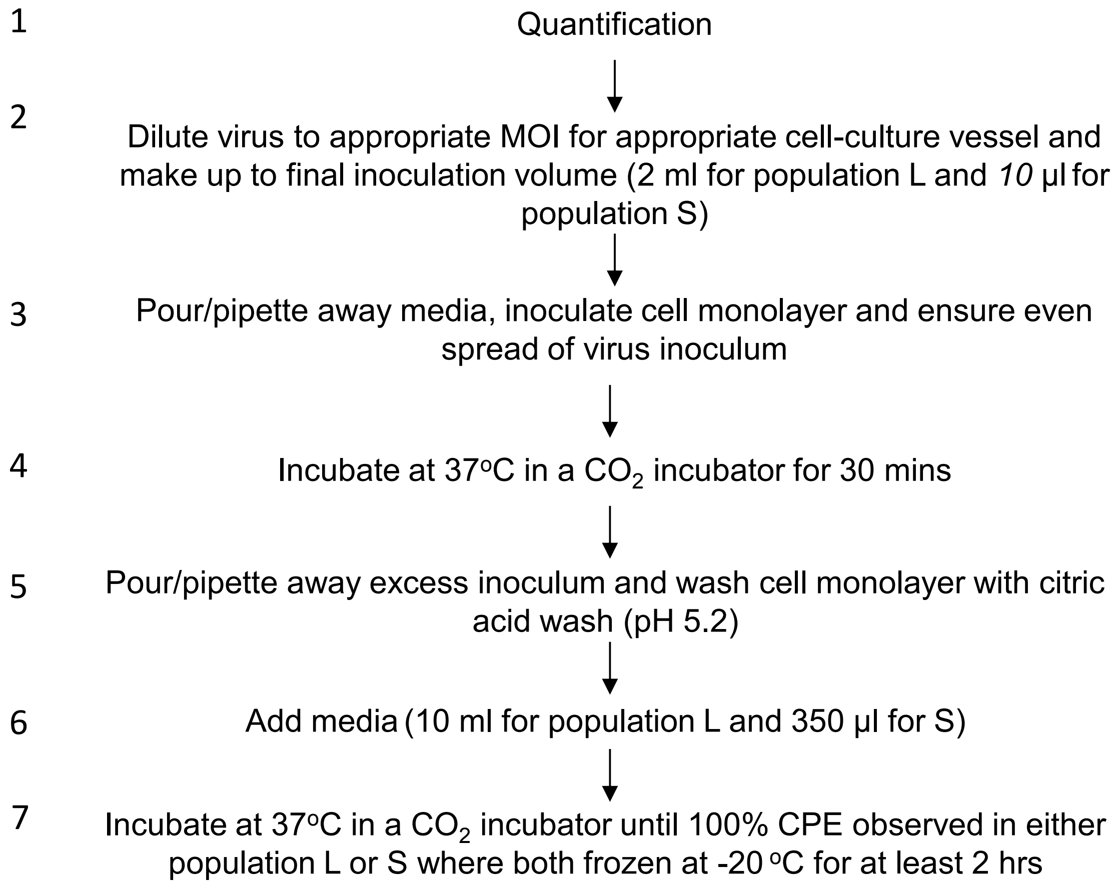
An increase in virus *titre* at each passage, from input to output, can be used to provide a quantitative measure of viral replication. However, traditional methods, such as the plaque assay discussed in section 6.3.4, or endpoint dilution, measuring the tissue culture infective dose that produces CPE in 50% of an inoculated cell culture ($TCID_{50}$/ml) (Khatib, Chason et al. 1980), are relatively insensitive and time-consuming.

6.3.5a PFU to FMDV RNA copy number comparison

A literature search revealed qRT-PCR as potentially providing a more rapid means of viral quantification. (Jonsson, Gullberg et al. 2009) observed a strong linear correlation between Ct (threshold cycle) value obtained by a two-step qRT-PCR method and viral titre (PFU/ml) obtained by the plaque assay method when quantifying the enterovirus EV7W. However, the calculated number of RNA genomes ('viral particles') needed to generate a plaque was not consistent across enterovirus serotypes, which ranged from a mean value of 94 to 3552 (Jonsson, Gullberg et al. 2009), with an average of 1001. Estimates for the PFU to FMDV particle ratio varies considerably and have been quoted as ranging between 1:7 x 10^3 and 1:1 x 10^4 (reference (Verdaguer, Fita et al. 1997) within (Sevilla, Ruiz-Jarabo et al. 1998), and between 1:8.3 x 10^1 and 1:2.5 x 10^3 (Bachrach, Trautman et al. 1964). Duplicate estimates of the PFU to FMDV particle ratio were made for O_1K B64 (mean value of 1:2.9 x 10^3). Consequently, this value of FMDV RNA copy number was used to calculate the number of RNA copies to be added at each passage, equating to a constant MOI, for both population sizes. These calculations were according to the T175 (population L) and 96-well (population S) cell counts given in table 6.1. RNA genomes were quantified by the one-step qRT-PCR, as described in Chapter 4, section 4.3.5, and PFU/ml was quantified using ZZ-R 127 cells.

6.3.5b Initial test at MOI 1

An initial test was conducted, in duplicate, to measure the increase in RNA copy number from time zero until 100% CPE was observed within population L infected at an MOI of 1. Figure 6.6 summarises the overall workflow for this procedure within *either* population L or S. It should be noted that inoculation volume for population S at step 2 of the workflow was dependent on the quantification of RNA copy number at step 1. The citric acid wash (AW) at step 5 was modified from that used in (Jackson, Ellard et al. 1996a), and contained citric acid crystals (Sigma), sodium citrate powder (Sigma) in saline solution. Media used for all experiments using ZZ-R 127 cells, as described in (Brehm, Ferris et al. 2009), although only 1% foetal calf serum used. The calculation for the average difference in PFU to RNA copies to infect population L at an MOI of 1, is not shown.



1 Quantification

2 Dilute virus to appropriate MOI for appropriate cell-culture vessel and make up to final inoculation volume (2 ml for population L and *10* µl for population S)

3 Pour/pipette away media, inoculate cell monolayer and ensure even spread of virus inoculum

4 Incubate at 37°C in a CO$_2$ incubator for 30 mins

5 Pour/pipette away excess inoculum and wash cell monolayer with citric acid wash (pH 5.2)

6 Add media (10 ml for population L and 350 µl for S)

7 Incubate at 37°C in a CO$_2$ incubator until 100% CPE observed in either population L or S where both frozen at -20 °C for at least 2 hrs

Figure 6.6

Schematic of experimental workflow for the infection of either population L or S

100% CPE was observed within duplicate populations after 7 hrs incubation of infected ZZ-R 127 cells, at which time they were both frozen at -20°C for 2 hrs (step 7 Figure 6.6). After thawing, 200 µ/ of 'virus/cell' suspension was added directly to 800 µl of TRIzol (Invitrogen, Paisley, UK) for total RNA extraction. Subsequent quantification by the one-step qRT-PCR, as described above, revealed only a tenfold increase in RNA copy number from input to output therefore the decision was made to test an alternative MOI.

6.3.5c Initial test at MOI 0.01

A literature search revealed that, when an excess of virus is present at high MOI, binding to surface polymers may approach saturation and tends to become independent of the affinity of the virus for those polymers (Sevilla, Ruiz-Jarabo et al. 1998). The same study hypothesised that, if binding to heparan sulphate, or other negatively charged cell-surface polymers, is a first step in the interaction of

FMDV with BHK-21 cells (Jackson, Ellard et al. 1996a), it may be possible that a low MOI can lead to selection of viral subpopulations with increased affinity for such charged polymers. Following this rational, a low MOI may therefore favour the inducement of selective pressure upon the cell-culture adapted virus studied here for the alternative cell surface receptor expressed by ZZ-R 127 cells (integrin). The hope was that infecting cells at a lower MOI might drive beneficial mutations towards fixation above background sequence noise more rapidly. It was also hypothesised that, if lower MOIs equate to more viral multiplication cycles, then this may result in a more pronounced increase in RNA copy number from input to output. Therefore, an inoculation MOI of 0.01 was tested. The average difference in PFU to RNA copies was then calculated for both population L and S, to give the total number of FMDV RNA copies added at every passage, equating to an MOI of 0.01 (Table 6.3).

Table 6.3 FMDV RNA copies/PFU comparison for population L & S (MOI 0.01)

				Total RNA copies ('Input') to achieve an MOI of 0.01	
Virus	RNA copies/ml[1]	PFU/ml[1]	Difference[1]	Population L	Population S
O_1K B64	1.3 x 10^{11}	4.4 x 10^7	2.9 x 10^3	5.7 x 10^8	9.7 x 10^5

[1] Average of two measurements

The test described in section 6.3.5b was repeated but for both population L and S. Cell monolayers were checked for CPE every 2 hrs. After 12 hrs of incubation at 37°C (step 7, Figure 6.5), no CPE was observed and so both populations were frozen at -20°C for 2 hrs and total RNA extraction, followed by quantification by qRT-PCR, as above.

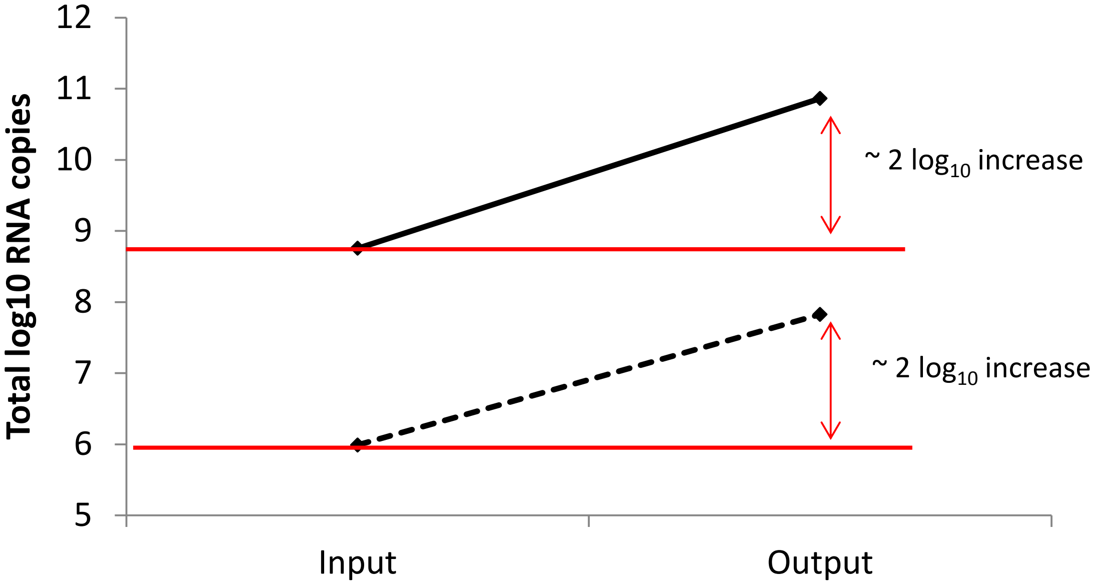


Figure 6.7
Quantification of FMDV RNA copy number by qRT-PCR for population L (solid line) and population S (dashed line) infected at MOI 0.01: Increase in RNA copy number from 'Input' to 'Output' (following incubation at 37°C for 12 hrs) is indicated by a red arrow.

After 12hrs of incubation, and although no visible CPE was observed, both population L and S demonstrated an approximate hundred-fold increase in FMDV

RNA copy number (Figure 6.7). The decision was therefore made to conduct the main experiment at an MOI of 0.01 with a fixed incubation period of 12 hrs.

6.3.5d Additional attempts to standardise the two population sizes

A number of measures were taken to standardise the relative rate of virus attachment and entry into cells within population L and S. As well as keeping incubation time constant, where possible, the difference in inoculation volume to monolayer surface area ratio was kept within a log between the two series. Attempts were also made to ensure even spread of virus inoculum within both cell-culture vessels, before incubation, through tilting (T175 flask) and rapid, horizontal movement’s back-and-forth (96-well).

6.4 Results

6.4.1 *In vitro* serial passage

Rescued infectious virus from the PT7S3 plasmid (Ellard, Drew et al. 1999) was subjected to two serial passages in vitro. One passage series was through simulated serial *narrow* bottleneck transmissions and, the other, *wide* bottleneck transmissions. Viral replication was measured by increases in viral RNA copy number, quantified by qRT-PCR (as described previously). Although the incubation period was constant (12 hrs), increase in viral RNA copy number decreased with passage, within both population L and S (Figure 6.8a and b)

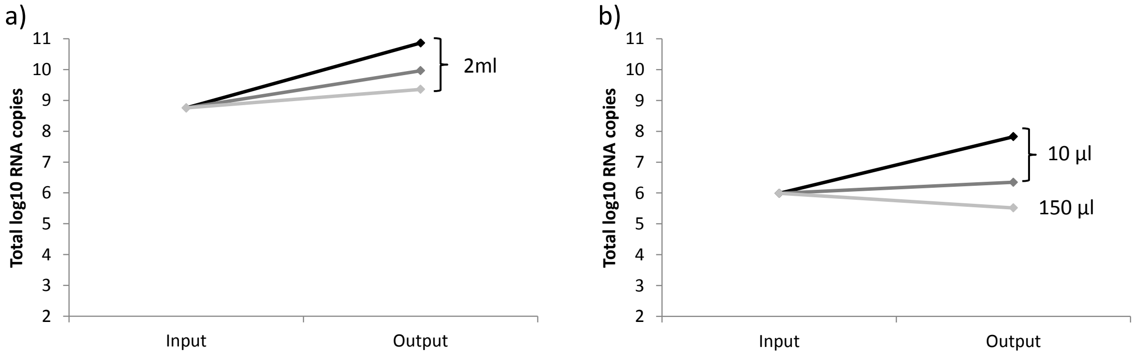


Figure 6.8

Quantification of O_1K B64 RNA copy number by qRT-PCR across 3 passages in cell culture for **a)** population L and **b)** population S. Both populations were infected at an MOI of 0.01. Passage 1 'Input' to 'Output' indicated by a black solid line, passage 2 by a dark grey solid line and passage 3 by a light grey solid line. Inoculation volumes are indicated and were modified for population S for passage 3 in order to account for FMDV concentration.

However, this decrease in RNA copy number was more pronounced within population S, compared to population L. Inoculation volume remained constant within population L but, where required to compensate for decrease in RNA copy number, was increased in population S.

6.4.2 Quantification of PCR products prior to NGS analysis

Following quantification by qRT-PCR, as described previously, starting template copy number was standardised across samples (for the reasons given in Chapter 4, section 4.2.1). One sample from passage 1 to 3 for O_1K B64 population S and L was diluted to a total of $5.0x10^5$ RNA copies before RT-PCR was performed, as described previously, before sequencing on the Illumina Genome Analyzer IIx platform. PCR products were visualized on a 0.7% Agarose gel, as described previously, where single products of the correct size (between 4 and 5 kb) were demonstrated (data not shown). PCR1 product yield was quantified using the Nanodrop spectrophotometer (Table 6.4).

Table 6.4 Nanodrop spectrophotometer quantification of PCR1 products for O_1K B64

Pass	Virus	Population size	PCR1 product yield (ng/µl)	260/280
1	O_1K B64	Large	23.6	1.86
		Small	18.5	1.74
2	O_1K B64	Large	13.6	1.91
		Small	20.1	1.90
3	O_1K B64	Large	21.4	1.76
		Small	12.6	1.83
'Input'	O_1K B64	-	22.9	1.84

6.4.2 NGS analysis

NGS workflow was as described in Chapter 3, although the Genome Analyzer II(x) platform was used (Glasgow Polyomics, University of Glasgow). Read filtering and trimming was essentially as detailed for the analysis pipeline discussed in Chapter 5 (Supplementary Table S2, Appendix 4). However, validation of observed mutations was achieved by use of the background sequence noise threshold of 0.5% (discussed in Chapter 4, section 4.4), plus the qualitative method described in Chapter 3 and 5, alone as no duplicate sequencing run had been performed. The nucleotide sequence of the linearized pT7S3-O_1K B64 plasmid was used as the reference sequences to which all reads were aligned.

All samples received coverage between X16,000 and X65,000 and shared similar regions of *low* and *high* coverage, as was found across samples discussed in Chapter 3 and 5. Figure 6.9 shows the coverage for all samples and indicates the over-represented primer regions that were omitted from analysis.

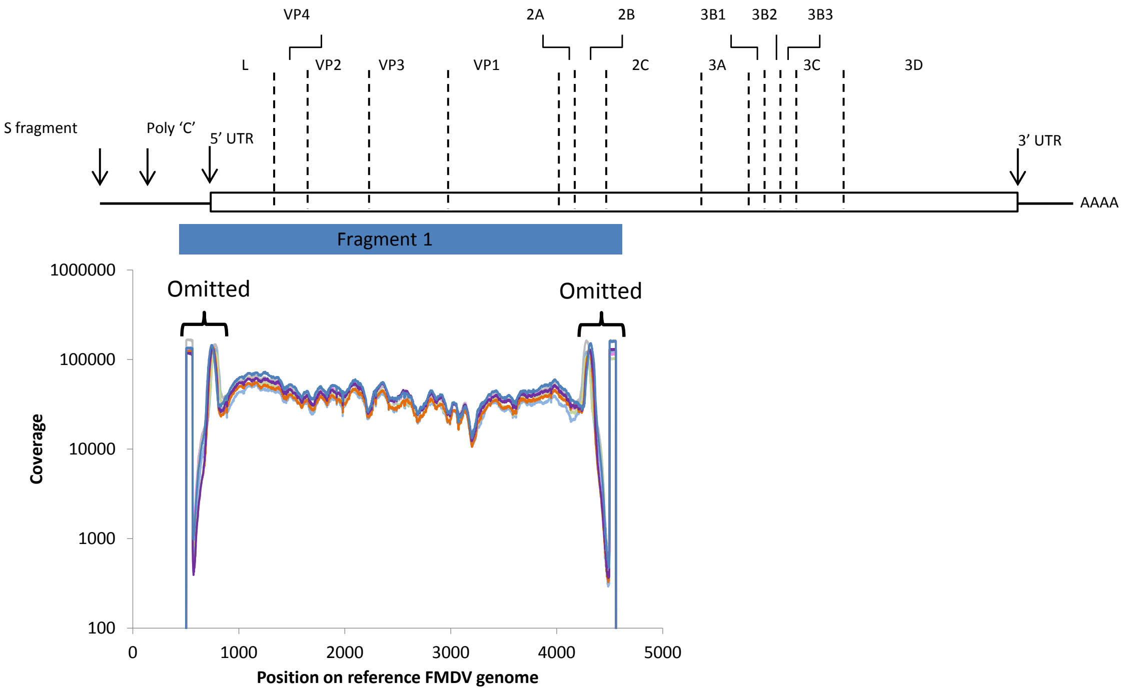


Figure 6.9

Coverage of the filtered and trimmed reads for O_1K B64 virus 'Input' plus population L and S (passage 1 to 3). Numbering is according to GenBank sequence EU448369. Over-represented primer regions, omitted from the analysis, are indicated.

The pattern of mutation frequency (> 0.5%) from 'Input' through to passage 3 (P3) revealed an observable difference between population L and S for O_1K B64 (Figure 6.10a and bi). Eight of a total of 40 sites where mutation frequency reached above 0.5% were omitted from analysis for population L and three of 18 sites from population S due to presence within primer associated regions.

The highest mutation frequency found within population L (9.7%) and S (12.9%) were found at site 2334 ($VP2^{134}$) and 883 (5' UTR), respectively (numbering according to GenBank sequence EU448369). The majority of mutations that occurred above 0.5% occurred below 1.0%. However, whereas no mutation reached above 1% within population L until P2, six mutations were observed above this frequency by P1 within population S. Of these six more dominant mutations, four were found within the polyprotein, the frequency pattern of which is shown in Figure 6.10 bii. Additionally, whereas all mutations (n=15) within population S were transitions (Ts), a relatively high number of transversions (n=7) were found within population L.

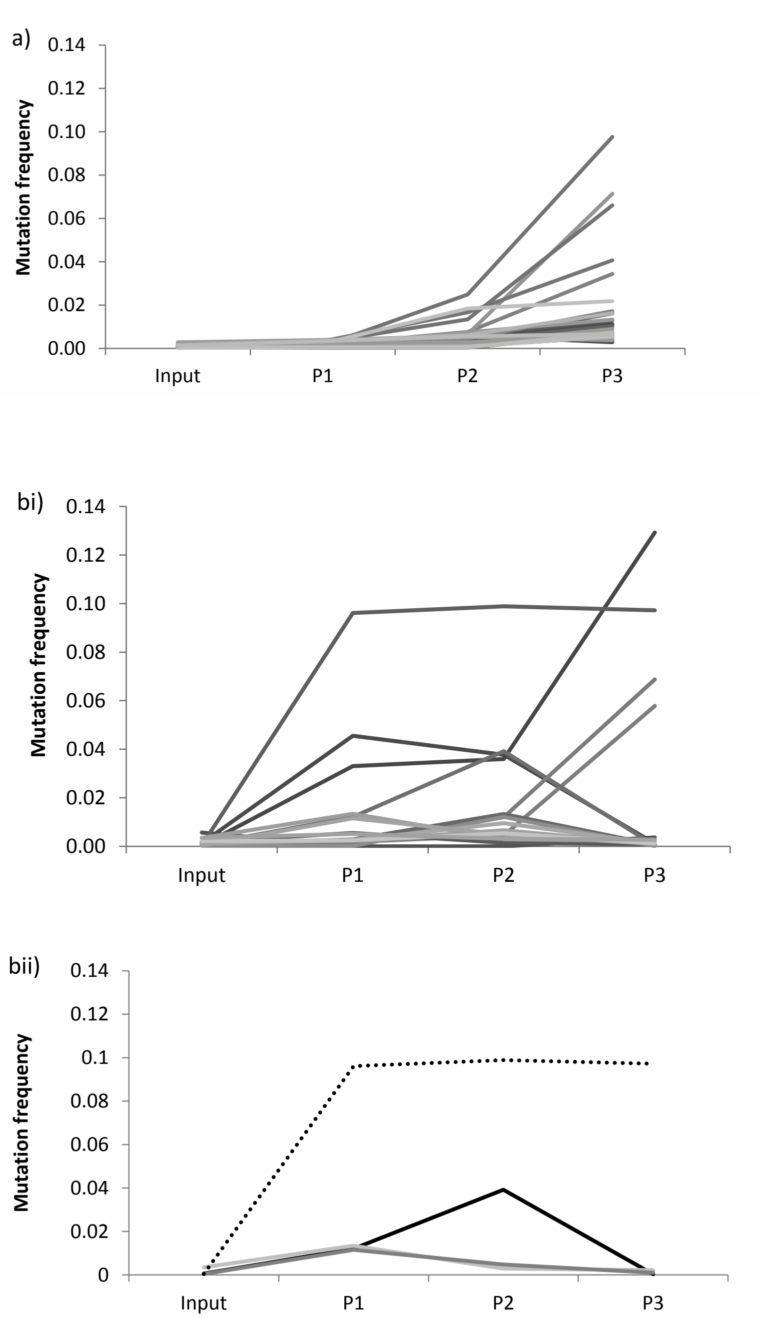


Figure 6.10

Mutation frequency within O_1K B64 virus 'Input' and **a)** at all 32 sites within population L (passage 1 to 3). **bi)** at all 15 sites within population S (passage 1 to 3). **bii)** at four highly dominant sites within the FMDV polyprotein of population S (solid lines indicate non-synonymous mutations and the dashed line a single synonymous mutation). Only mutations that reached above 0.5% were included in this analysis.

Population L and S of O_1K B64 were also found to differ over a very narrow range (between 1.25 and 1.04) when comparing the ratio of non-synonymous to synonymous mutations within the open reading frame (*dN/dS*) (Figure 6.11 a). *dN/dS* was calculated as described in Chapter 5.

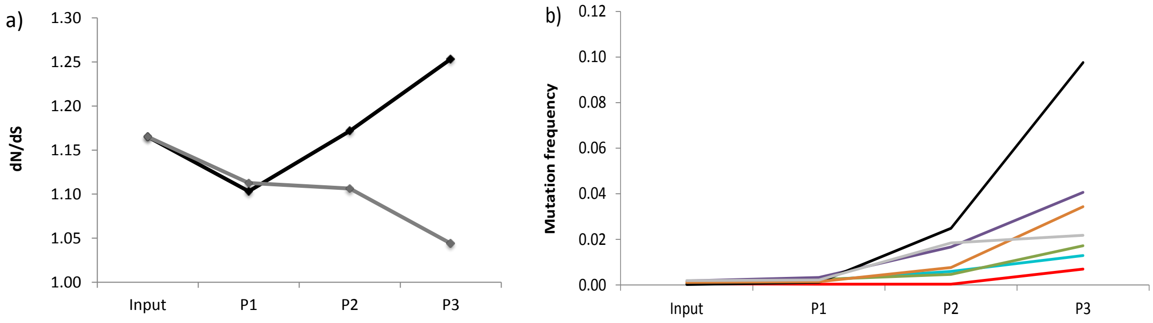


Figure 6.11

Evidence of selective pressure. **a)** Ratio of non-synonymous to synonymous mutations (dN/dS) within O_1K B64 virus population L (black solid line) and population S (dark grey solid line) from 'Input' through to passage 3. **b)** Mutation frequency within O_1K B64 virus 'Input' and population L (passage 1 to 3) for non-synonymous mutations associated with the subtype O1 FMDV-HS receptor complex only: VP2^{134} (black line), VP3^{88} (purple line), VP2^{138} (gold line), VP1^{195} (grey line), VP3^{60} (green line), VP3^{56} (turquoise line), VP3^{59} (red line).

Additional evidence for positive selection was provided by increasing frequencies of non-synonymous mutations associated with the subtype O1 FMDV-HS receptor complex in population L (Figure 6.11 b). Increased mutation frequency was found within seven of the nine motifs identified previously (Fry, Lea et al. 1999a) (Table 6.5 provides details of the substitutions and subsequent amino acid property changes at these sites). No mutations occurred above 0.5% at these sites within population S.

Table 6.5 Details of non-synonymous mutations associated with the subtype O1 FMDV-HS receptor complex in population L (O_1K B64 virus)

Genome position[1]	Ligand	Nt change	Amino acid change	Amino acid Property change	Maximum mutation frequency (%) observed
2334	$VP2^{134}$	A → C	Lys(K) → Gln(Q)	Polar/positive → /neutral	9.8
2346	$VP2^{138}$	T → C	Tyr(Y) → His(H)	Polar/neutral → /sometimes positive	3.4
2754	$VP3^{56}$	C → T	Arg(R) → Cys(C)	Polar/positive → /neutral	1.3
2764	$VP3^{59}$	G → C[2]	Gly(G) → Ala(A)	No change	0.7
2767	$VP3^{60}$	G → A	Gly(G) → Asp(D)	Non polar/neutral → /negative	1.7
2851	$VP3^{88}$	A → G	Asn(N) → Ser(S)	No change	4.0
3832	$VP1^{195}$	A → C[2]	His(H) → Pro(P)	Polar/sometimes positive → non polar/neutral	2.2

[1] Numbering according to GenBank sequence EU448369

[2] Alternative Nt substitution found within 'Input' and P1 but below 0.5% threshold limit

The A → C mutation at $VP2^{134}$ reached the highest frequency through the passage series within population L of O_1K B64, whereas the C → T mutation at $VP3^{56}$ thought of as the most critical motif in terms of virus/cell receptor recognition, only reached a frequency of 1.3% by P3. The mutation with the second highest frequency within O_1K B64 population L was T → A mutation at $VP3^{61}$.

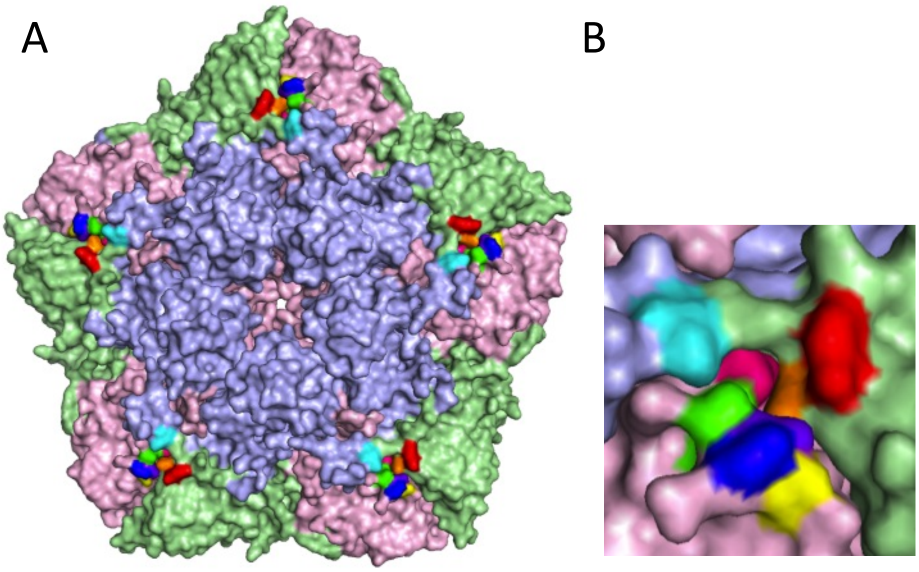


Figure 6.12
Predicted structure of the molecular surface of the FMDV pentamer. **A** Front on image showing repeated elements, VP1 (washout blue); VP2 (washout green); VP3 (washout red). HS associated motifs are highlighted: VP2^{134}and VP2^{138} in red and orange respectively; VP3^{56},VP3^{59}, VP3^{60} and VP3^{88} in green, blue, purple and hot pink respectively; VP1^{195} in cyan. The additional motif at VP3^{61} is highlighted in yellow. **B** Enlarged image of HS associated motifs. Predicted structure made using Pymol version 1.5.0.4 from Schrodinger LLC.

The additional VP3^{61}motif (highlighted in yellow in Figure 6.12) clusters with the other seven HS associated motifs on the FMDV capsid surface. However it is not known if this motif has a role to play in the O1 FMDV-HS receptor complex.

Additionally, the complexity of both O$_1$K B64 populations was found to differ through the passage series, measured by Shannon entropy (as described in Chapter 5). A two fold difference in entropy was observed between population L-P3 and population S-P3). After an initial increase from ‘Input’ to P1, for both populations, entropy was found to remain relatively constant within population L and decrease within population S (Figure 6.12).

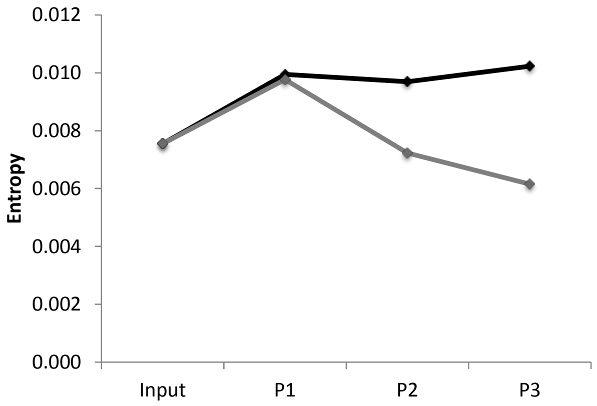


Figure 6.13

Pattern of population complexity (measured by Shannon entropy) within O_1K B64 virus population L (black solid line) and population S (dark grey solid line) from ‘Input’ through to passage 3.

6.5 Discussion

A study was conducted to better characterise viral population diversity generated during serial passage of FMDV *in vitro* from a more defined starting material. Performing the experiment with two different viral population sizes tested the hypothesis that mutations, irrespective of selective value, become more rapidly fixed in the population during more severe bottleneck transmissions.

No mutation, in either passage series, was observed above 13%. In contrast to the amount of viral replication that occurs at the intra-host scale, where consensus sequences can remain invariant (Murcia, Baillie et al. 2010; Bull, Luciani et al. 2011; Bull, Eden et al. 2012), relatively little viral replication had occurred by cell-culture passage 3; therefore it was not surprising that no mutations had reached consensus level. Consequently, the current study provided evidence in support of the hypothesis that mutations become more rapidly fixed in the population during more severe bottleneck transmissions. However, there was no *genetic* evidence to suggest whether mutations observed in population S were either advantageous or deleterious, which will be addressed later in the discussion.

The decreasing level of population complexity observed during serial *small* but not *large* population passage of FMDV was comparable to that demonstrated within published *in vitro* studies (Novella, Duarte et al. 1995; Escarmis, Davila et al. 1996; Escarmis, Davila et al. 1999; Escarmis, Gomez-Mariano et al. 2002; Domingo, Pariente et al. 2005; Escarmis, Lazaro et al. 2008) and reviewed by Domingo, Escarmis et al. 2005. An additional comparison can be drawn between the observations made here and those by Escarmis *et al.* (2008) where evidence for positive selection was also demonstrated during large population passages but not during ‘bottleneck transfers’, or small population passages (Escarmis, Lazaro et al. 2008). However, the number of biologically significant sites at which the cellular receptor associated mutations occurred and rate at which they accumulated, was surprising. The *in vitro* system used was a proxy for the host environment because a large proportion of ZZ-R 127 cells are known to express αvβ6 integrin, compared to other cell lines (for example, BHK-21 cells) that express heparan sulphate. However, this system comprised polymorphic but

mainly epithelium cell types and, as such, was a relatively simple system when compared to, for example, a bovine host.
The agreement between these and published *in vitro* results, plus observations made of mutations at biologically significant sites, provide a degree of confidence that these observations were not all artefacts of the experimental design. The increase in mutation frequency at the biologically significant sites, coupled with the relatively constant entropy measures for population L, provides additional evidence that virus was not being 'diluted out' through the passage series. However, although 'infective' dose (RNA copy number) and incubation period was kept constant, there was still an apparent decrease in productive viral replication, within both population S and L with viral passage.

No evidence of lethal mutation (STOP codons) or 'defector' genomes (non-synonymous mutations within highly conserved regions of the polyprotein) was found. *Defector* genomes are defined as any type of replication-competent genome (dependent or not on the standard virus for replication) with the potential to interfere with the replication of the standard virus. These *defector* genomes therefore include defective-interfering (DI) particles, which are dependent on standard virus for completion of their infectious cycle (Moreno, Tejero et al. 2012). The influence of such genomes cannot be ruled out as mutations leading to their creation may have occurred within the *un-sequenced* region (containing non-structural proteins). Production of mutant and defector viruses is favoured by increased mutation frequency at low MOI. However, positive stranded RNA viruses, including FMDV, have been shown to be more tolerant of defector genomes, replicating at enhanced mutation rates at low MOI, compared to negative strand viruses such as vesicular stomatitis (Moreno, Tejero et al. 2012). The presence of a high proportion of defector genomes is often given as an explanation for the high virus particle to PFU ratio (Holland, Spindler et al. 1982; Domingo, Martinez-Salas et al. 1985). However, a study by Belsham *et al.* (1988) demonstrated that, FMDV RNA molecules, microinjected directly into the cytoplasm of BHK cells, had an infectivity close to 1 PFU per molecule. The same study speculated that high virus particle number to PFU ratios may reflect some inefficiency within a component of viral RNA delivery to the cytoplasm rather than due to a large proportion of defective viral genomes (Belsham and Bostock 1988).

The exact cause for the decrease in viral replication observed during these experimental passages is not known. Future work may include a repeat of this experiment, with population L only initially, using the same MOI but quantifying viral titre by plaque assay at each passage to ascertain whether the same decrease in viral replication was observed. If a greater number of passages could be achieved, it may also be interesting to note the progression of mutations associated with the cellular receptor switch with passage and if/when this resulted in any changes in plaque morphology.

The same experiment was attempted using a variant of the infectious copy studied here. This chimera virus contained an approximately 5 kb long sequence for the surface-exposed capsid proteins from the FMDV field-strain, O/UKG/34/2001, as described in (Botner, Kakker et al. 2011). However, no increase in RNA copy number was observed following passage 1, which is why this study was terminated. Nevertheless, future work could involve further investigation of the evolutionary progression of phenotypically distinct viruses, subjected to different environmental pressures *in vitro*, at the ultra-deep level. Sequencing the full FMDV genome by NGS, with or without enhanced mutagenesis, may also clarify the potential impact of defector genomes.

Although all practical measures were taken to keep conditions between the two populations studied as constant as possible, this process was not exhaustive. In particular, it is speculated that surface tension and so called 'slosh dynamics' would have varied between culture vessels due to variations in surface area relative to depth. Such variations in fluid dynamics may have implications for virus spread over the cell monolayer. Therefore, care would need to be taken especially when using this experimental design to investigate fine-scale viral evolutionary dynamics at the inter and intra-cellular scale.

This pilot study has demonstrated the importance of population size to the evolutionary dynamics of FMDV, at a previously un-obtained depth of resolution using NGS. Specifically, the study shows that the variance in the level of site-specific polymorphism depends directly on the bottleneck size – with higher variances related to narrower bottlenecks. Furthermore, the successful application of NGS, in conjunction with a known biological *marker*, may provide ‘proof of concept’ for its application for similar investigations at scales previously not thought possible in terms of background sequence noise.

Chapter 7

Discussion

7.1 Overview of thesis

The genetic relatedness between full genome consensus sequences provides valuable insights into the evolutionary dynamics of FMDV at the epidemic scale, as described for the 1967 UK FMD outbreak, within Chapter 2. However, during an epidemic, virus replicates within multiple animals, where it is also replicating and evolving within multiple tissues and cells. Each scale of evolution, from a single cell to multiple animals across the globe, involves evolutionary processes that shape the viral diversity generated below the level of the consensus. The use of next-generation sequencing (NGS) for the dissection of the finer scales of viral population diversity has been evaluated for FMDV within this thesis. Substantially increased coverage provided by NGS has enabled improved resolution and characterisation of viral populations below the level of the consensus, as described in Chapter 3.

Collaboration with the Institute of Biodiversity, Animal Health and Comparative Medicine at the University of Glasgow provided the specialist bioinformatic and statistical capabilities required for the analysis NGS datasets. As part of this collaboration, a new systematic approach was developed to process data produced by NGS and distinguish genuine mutations from artefacts. Additionally, NGS data produced during this PhD was used within evolutionary models to estimate parameters such as the genome-wide mutation rate of FMDV.

The bulk of the diversity identified by NGS is provided by low frequency mutations which poses a challenge in terms of distinguishing genuine mutations from artefacts. As one of the most significant challenges facing the use of this technology, experiments were undertaken to quantify the occurrence of these artefacts (Chapter 4). Analysis of the mutation spectra generated from a clonal control study established a mutation frequency threshold of 0.5% above which there can be a high degree of confidence that a mutation is real in the sense that it is present in the sampled virus population. This threshold, together with an optimized sample processing protocol, was used for the more extensive investigation of within and between host viral population dynamics during transmission (Chapter 5). Variations in the polymorphic structure of FMDV populations extracted from biological samples revealed evidence for two bottleneck sizes occurring within and between hosts. However, following the pattern of mutation frequency at single nucleotide (nt) positions was also revealing about forces of evolution at play.

A limited number of sites typically demonstrated higher frequency mutations (>1%), including sites of biological significance, both *in vivo* and *in vitro*. These sites provided the opportunity to determined fine-scale fluctuations in mutation frequency during *in vitro* passage (Chapter 6), thus providing the necessary resolution to further demonstrate the impact of bottleneck size on viral populations.

7.2 Site specific polymorphisms

Cell-culture adapted viruses of the same topotype were used during both *in vitro* passage series in Chapter 6 (O_1K B64) and the *in vivo* transmission chain discussed in Chapter 3 (O_1OUK 2007). The *in vitro* cell-culture system used was a proxy for the host environment, as discussed in Chapter 6. However, direct comparison between this *in vivo* and *in vitro* system should be attempted cautiously since the defined area of infection, rounds of replication, number of infected cells and interchange of virus particles are not necessarily equivalent. In spite of these differences, it is interesting to compare the frequency of site-specific mutations observed within the 'large' population *in vitro* passage series and those *in vivo.* Where seven amino acid motifs associated with the cellular receptor switch from HS to integrin had an average mutation frequency of 3.2% by passage three *in vitro*, only two were present in calf 1 (A1) two days post inoculation *in vivo*. Both of these mutations were present as a minority within $VP3^{56}$ (C→T change at the first codon position resulting in an Arginine to Cysteine amino acid change and a G → A change at the second codon position resulting in an Arginine to Histidine amino acid change). Following the progression of these known sites through the *in vivo* transmission chain, where the G→A change within $VP3^{56}$ fell in frequency until it was no longer observed above background sequence error, the C→T change within $VP3^{56}$ increased in frequency and became fixed in the population. However, although the G→A change within $VP3^{56}$ was below consensus level in calf 2 (A2) when sampled six days post first contact, a probang sample taken from A2 32 days later revealed this mutation at consensus level (Juleff, Valdazo-Gonzalez et al. 2013), resulting in a Histidine at $VP3^{56}$ rather than a Cysteine.

Although fitness was not directly tested, it may be reasonable to infer that replicating viruses exhibiting mutations at sites associated with the switch in cellular receptor usage, both *in vitro* and *in vivo*, were moving towards increased fitness. Additional evidence for positive selection acting on the viral swarm subjected to ‘large’ population passages *in vitro* was provided by the ratio of transitions to transversions. Transversions were relatively common among the observed mutations *in vitro*. However, this observation was not mirrored *in vivo* (κ defined as 2Ts/Tv, of 7.1 compared to > 60 respectively), where additional evidence for positive selection was absent. The frequency distribution over time of site-specific mutation, as measured by NGS, therefore provides additional evidence of the impact of both selective pressure and bottleneck size on the fixation of mutations in the consensus sequence. Furthermore, where minimal variation in entropy levels were measured, by virtue of the mutation spectra *in vitro*, frequency distribution of site-specific mutations provided the additional resolution necessary to observe the impact of bottleneck size.

7.2.3 Implications

NGS allows both the measure of mutation *spectra* (all mutations) and tracking of site-specific mutation frequency over time and through transmission events. The combined application of both of these features has enabled the effective evaluation of bottleneck size and its impact on viral population diversity and fixation of mutations into the consensus sequence. NGS analysis confirmed that within-host viral populations are highly diverse and demonstrated often greater variations in population heterogeneity compared to those measured between different hosts. According to the findings of this thesis, a faithful representation of within-host diversity is typically transmitted to the next host, whereas more acute impacts of bottleneck events are more frequently demonstrated within a single host. The analysis of FMDV deep sequence data across epidemiologically significant scales has resulted in a more sophisticated understanding of the consensus sequence, in terms of its use for transmission tree reconstruction, and a better calibration of how viral genetic differences accumulate with transmission. Furthermore, the ability to distinguish between the population structure of multiple samples taken from a single host may provide the means to reconstruct both intra- and inter-host transmission routes in the future.

7.4 Future work

The volume of data generated during this PhD was only possible by use of NGS. However, the viral diversity observed, within-host and through transmission, could be further validated via targeted molecular cloning techniques. Such techniques could also validate future haplotype frequency estimations using the Bayesian statistical tool, ShoRAH (Zagordi, Geyrhofer et al. 2010). Although labour-intensive, endpoint dilution of template would avoid PCR introduced amplification errors and would therefore also be employed during future validation experiments. The dataset compiled represents a novel opportunity to evaluate different filtering/alignment/SNP calling algorithm and software. Running this data set through multiple pipelines would not only provide a controlled evaluation of those pipelines but would also potentially validate the analysis conducted within this PhD.

The observation of variations in FMDV population diversity within-host requires further investigation. Applying NGS analysis to tissue-specific sampling may yield additional information regarding the influence of tissue-specific amplification of virus to the generation of viral diversity within-host. Future application of NGS to strand-specific amplification strategies may also play an important role in improving the estimates of viral mutation rate.

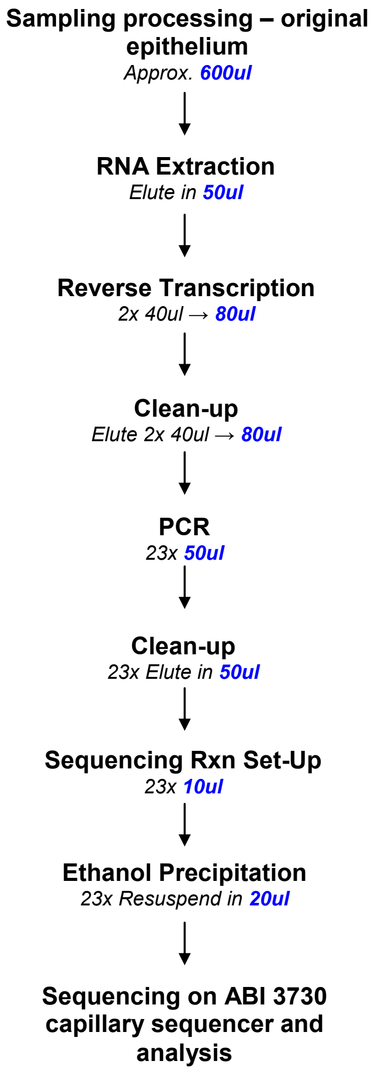
Although NGS provides the means to characterise intra-sample viral swarms at previously unobtainable depth, questions remain about the impact of introduced bias and error during RT-PCR and sequencing, as well as variations in the efficiency of the processes themselves. Ideally, an RNA template control of known sequence would be processed in parallel to assess and quantify the accumulation of such artefacts. The External RNA Control Consortium (ERCC) ‘synthesizes’ RNA by in vitro transcription of synthetic DNA sequences, as well as DNA derived from *Bacillus subtilis* and the deep-sea vent microbe *Methanocaldococcus jannaschii.* These standardised control RNAs are being developed to be used in microarray, qPCR and sequencing applications (Baker, Bauer et al. 2005; Devonshire, Elaswarapu et al. 2010). A study by Jiang et al (2011) examined the use of spike-in ERCC RNA controls during RNA-seq data generation on the Illumina GAII and concluded that their inclusion allowed the measurement of systematic biases in quantification, such as underrepresentation of short transcripts and read coverage heterogeneity (Jiang, Schlesinger et al. 2011). Such a standardised set of RNA controls (spike-in and/or external) could also be used for the direct measurement of sequencing error rates, coverage biases and other variables that affect the accurate representation of viral populations using NGS. The ERCC are addressing fundamental questions about RNA control sequence uniqueness (for spike-in), secondary structure and length. The future may entail the availability of entirely chemically synthesized viral RNA/DNA genomes, which take into account target genome sequence length (with improved synthesis fidelity), secondary structure and nucleotide composition. At such a time, aside from obvious biosecurity and ethical concerns, one could imagine effectively using such synthetic genomes, not only as an authentic control but potentially, as template material upon which to conduct experimental evolution studies themselves.

With the intended commercialisation of ‘Strand sequencing’ by Oxford Nanopore Technologies by the end of this year, it would be extremely short sighted not to discuss direct RNA sequencing within this section. Briefly, strand sequencing is a technique that passes intact DNA or RNA polymers through a protein nanopore set in an electrically resistant membrane bilayer, sequencing in real time as the DNA/RNA translocates this pore. Error rate is typically estimated to be high for current single molecule platforms, which are susceptible to quenching effects between adjacent dye molecules as well as the effects of ‘Dark nucleotides/ probes’ (a nucleotide or probe that does not contain a fluorescent label) (Metzker 2010). However, developers at Oxford Nanopore technologies are apparently already working on improving accuracy. Reads of several kb in length will be invaluable for haplotyping, the study of epistasis and structural variant analysis. Among other applications, this technology will enable more informative, achievable and cost effective exploration of RNA virus fitness landscapes. By resolving the issue of linkage between mutations and allowing the reconstruction of haplotypes, these ‘Third’ generation sequencing platforms are closing the gap between viral populations that exist at the finest scales in nature and what we are capable of sequencing.

7.5 Conclusions

When applied to the investigation of rapidly evolving RNA viruses, the power of NGS lies with its depth of resolution into the viral swarm. As demonstrated within this thesis and previous studies, this level of resolution is imperative for the dissection of such viral populations present within a single host. While viral diversity at any given scale can be considered as a function of that observed at the scale immediately below, this function diminishes with distance between scales. Therefore, while ultra-deep sequencing by NGS has provided clarification of the use of the consensus sequence, it would be unnecessary and counter-productive to use such depth of sequence coverage for routine tracing measures above the host-to-host scale. Nevertheless, the high-throughput nature of the technology, without the need for additional depth of sequence coverage, can be utilized for processing a greater number of samples more rapidly. Future research in the field of RNA virus evolution, which utilizes these advancing sequencing technologies, should rather focus on refining our understanding of virus- host interaction and pathogenesis. This thesis therefore provides a stepping stone in a rapidly evolving field of research and also demonstrates the invaluable partnership between ‘wet’ and ‘dry’ science. The dynamic between model driven and experiment driven research is potentially very powerful, especially when combined with continued advances in viral genomic sequencing. A more sophisticated, tailored understanding of viral diversity at its finest scales will lead to more well informed and accurate models of viral evolution. This in turn could hold the key to the better understanding of viral pathogenesis and, therefore development of effective and sustainable disease treatment and control measures.

Appendix 1



Sampling processing – original epithelium
Approx. ***600ul***

RNA Extraction
Elute in ***50ul***

Reverse Transcription
*2x 40ul → **80ul***

Clean-up
*Elute 2x 40ul → **80ul***

PCR
*23x **50ul***

Clean-up
*23x Elute in **50ul***

Sequencing Rxn Set-Up
*23x **10ul***

Ethanol Precipitation
*23x Resuspend in **20ul***

Sequencing on ABI 3730 capillary sequencer and analysis

Figure 1

Full genome consensus sequencing overview

Protocol for (complete genome consensus sequencing (CGCS)) and (next-generation sequencing (NGS)) template production

qPCR quantification of RNA template (NGS), using the SuperScript III One-Step RT-PCR System with Platinum Taq High Fidelity (Invitrogen)

	X1
2x reaction mix	12.5µl
Nuclease free water	1.5 µl
Forward primer (3D Callahan) 10mM	2 µl
Reverse primer (3D Callahan) 10mM	2 µl
Probe (3D Callahan) 5mM	1.5 µl
High fidelity enzyme mix	0.5 µl

20 µl of above master mix added to 5 µl of RNA template before quantification on Stratagene Mx3005P machine (Agilent Technologies, UK) with following thermal cycling conditions:

Step	Temp ^{o}C	time
1	60	30 minutes
2	95	10 minutes
3	95	15 seconds
4	60	1 min 6 seconds
5	Go to step 3 x 50 times	

Reverse transcription using 2 x 15µl RNA (CGCS) and 1x 15 µl RNA (NGS)

Set up reverse transcription reaction/s as follows;

In a 0.2ml eppendorf add 15µl RNA, 3µl oligo-dT primer (10µM), and 3µl 10mM dNTP (CGCS as detailed in Chapter 2), 3µl FMDV specific primer 1 (10µM) and 3µl FMDV specific primer 2 and 3µl 10mM dNTP (NGS as detailed in Chapter 3, 4 and 5).

Heat to 70°C for 3 minutes

Place immediately on ice for 3 minutes

Add RNA\primer\dNTP from above to 17 µl (CGCS) and 14 µl (NGS) freshly made RT mix in 0.2ml eppendorfs;

RT mix

	X1
5 x FS buffer	8 µl
0.1mM DTT	2 µl
RNase OUT	2 µl
Nuclease free water	5 µl

Add 2 µl of Superscript III reverse transcriptase (Invitrogen) to each reaction

Run on a thermocyler on the following programme

Temperature °C	Time
45	60 min (CGCS) 90 min (NGS)
85	5 min
4	∞

cDNA clean-up to give pooled total of 80µl (CGCS) or single 40µl (NGS)

Use the GFX clean-up columns (GE Healthcare) (CGCS) QIAquick PCR Purification Kit (QIAGEN) (NGS) both as per the manufactures instructions

23 PCR reactions (CGCS) and 4 PCR reactions (NGS) using 3µl cDNA per reaction

Master PCR mix made up as follows:

	X 1	X 50
10x PCR Buffer (*Invitrogen kit*)	5 µl	250 µl
MgSO4 (*Invitrogen kit*)	2 µl	100 µl
dNTP (10mM)	1 µl	50 µl
Platinum Hi fidelity Taq (*Invitrogen kit*)	0.25	12.5 µl
Nuclease free water	37	1850 µl
Forward primer (10µM)[1, 2]	1µl	
Reverse primer (10µM) [1, 2]	1µl	
cDNA	3µl	
Mastermix (Table 1)	45µl	
Total	**50µl**	

[1] As detailed in Chapter 2 (with M13 tagged primers)

[2] As described in Chapter 3, 4 and 5

Add 45 µl of master mix to each reaction

Add 3 µl of cDNA

PCR thermal programme:

Step	Temp ^{o}C	time
1	94	5 minutes
2	94	30 seconds
3	55	30 seconds
4	72	1 minute (CGCS) 4 minute (NGS)
5	Go to step 2 x 39 times	
6	72	7 minutes
7	4	∞

PCR reaction clean-ups to give 50µl DNA

Clean up the PCR reactions as described for the cDNA reactions, however elute in 50µl.

Ethanol Precipitation (method)

Turn on centrifuge to be at 4^{o}C

Make up stop solution using 170µl of 100mM EDTA, 170µl 3MNaOAc, and 85µl of glycogen.

Add 5µl of freshly prepared stop solution to each reaction

Add 60µl ice cold 95% ethanol to each well

Seal plate with adhesive foil and mix thoroughly by vortexing

Centrifuge at maximum speed (>1100 x g) for 30 minutes at 4oC

Invert plate and pour out supernatant over sink – three gentle shakes

Gently add 200µl ice cold 70% ethanol

Centrifuge at maximum speed for 15 minutes at 4oC

Invert plate over sink and pour off supernatant

Gently add 200µl ice cold 70% ethanol

Centrifuge at maximum speed for 15 minutes at 4oC

Invert plate over sink and pour off supernatant keep plate upside down and gently blot on tissue paper

Vacuum dry for - until dry!

Re-suspend pellets in 40µl Sample Loading Solution

Leave for 5 minutes then vortex.

Either put in freezer or add a drop of mineral oil and run on sequencing machine

Sequencing reaction set-up using the BigDye Terminator V3.1 Cycle Sequencing Kit by Applied Biosystems for sequencing on the ABI 3730 genetic analyser (CGCS).

	X1
5x sequencing buffer	1.88µl
BigDye	0.25 µl
Primer (1mM)*	1.5 µl
Nuclease free water	5.37 µl
Template	1 µl
Total	**10 µl**

* After PCR in performed with M13 tagged primers, sequencing is performed with both Uni-F1 and Uni-R1 primers and 23 times untagged

Sequencing thermal cycle

Step	Temperature ºC	Time
1	96	1 min
2	96	10 seconds
3	50	5 seconds
4	60	4 min
5	Go to 2 25 times	
6	4	Hold

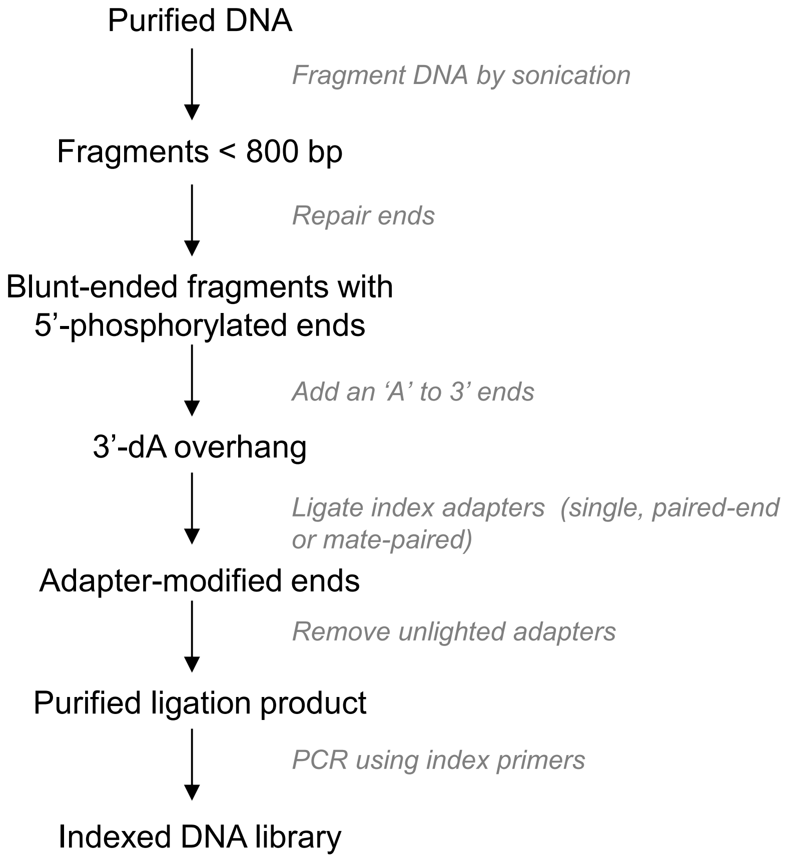

Purified DNA
Fragment DNA by sonication
Fragments < 800 bp
Repair ends
Blunt-ended fragments with 5’-phosphorylated ends
Add an ‘A’ to 3’ ends
3’-dA overhang
Ligate index adapters (single, paired-end or mate-paired)
Adapter-modified ends
Remove unlighted adapters
Purified ligation product
PCR using index primers
Indexed DNA library

Figure 1

Summary of the of the Illumina GA library preparation workflow

Appendix 2

Basic statistics of Illumina yield

The reads obtained with the Illumina Genome Analyzer were collected in *.fastq* files. The first run consisted of a total of 7,190,884 reads of 57-nucleotides (nt) in length. The last 7 nts of each read defined the sequence tag, and were used to assign individual reads to each sample. Reads containing at least one unresolved nt (387,809, 5.55% of the total), and reads having a corrupted tag (207,749, 2.89% of the total) were removed from the analysis. The 6,595,326 remaining, 50nt-long reads, were assigned to the three samples: 1,723,151 (26.1%) had the first tag (corresponding to the Inoculum), 2,751,260 (41.7%) had the second tag (lesion on the Front Left Foot, or FLF) and 2,112,932 (32.0%) had the third tag (lesion on the Back Right Foot, or BRF).

The second run yielded 10,116,147 79-nt long reads, with the last 9 nts containing the sequence tag. 26,428 (0.27%) reads contained at least one unresolved nt and 288,230 (2.85%) reads had a corrupted tag and were removed from the analysis. Among the remaining 9,801,489 70-nt reads, 3,775,685 (38.5%) belonged to the Inoculum, 2,542,913 (25.9%) to FLF and 3,482,891 (35.5%) to BRF.

Data filtering

The quality scores associated with each nucleotide were lower on the first run and decreased towards the end of reads (Figure 1). In order to make direct comparisons between the two runs, we trimmed reads from the second run to 50nt. Typically, quality scores decreased along a read, as the reliability of the sequencing process decreased with the number of cycles of the Sequencing Platform. As Figure A1 shows, a trade-off is present between the number of reads kept and their quality. For both runs, we discarded reads with average error per nt below θ = 0.2%, %, resulting in a flatter error profile along the read.

With this choice of the threshold, 66% of the reads were retained from the first run (a total of 4,361,101 reads: 1,060,906 for the Inoculum, 1,736,381 for FLF and 1,328,588 for BRF), and 95% of reads from the second run (a total of 9,277,876 reads: 3,567,541 for the Inoculum, 2,412,897 for FLF and 3,303,438 for BRF). The better performance of the second dataset has to be attributed to an upgrade of the Illumina platform.

Reads alignment and trimming

The vast majority of the filtered 50nt-reads aligned unambiguously with less than 5 mismatches to the reference inoculum genome, previously established using conventional Sanger sequencing (Cottam, Wadsworth et al. 2008b) (Genbank accession no. EU448369), (run 1: 92.5% for Inoculum, 98.9% for FLF, and 97.8% for BRF, run 2: 95.8% for Inoculum, 98.4% for FLF and 96.2% for BRF). The remaining reads were either ambiguously aligned reads or contained a large number (>4) of mismatches to the reference sequence, and were discarded from the analysis. For each sample, an almost equal number of reads were derived from positive and negative strands of the viral cDNA.

Further filtering of the data was undertaken after alignment of the reads. Within the aligned reads, mismatches occurred with similar frequency at each of the 50nt of the reads, except from the edges, where a higher number of mismatches was observed (Figure 2). Presumably, these peaks were due to a small number of sequences with insertions or deletions close to the ends of the reads: for subsequent analysis we trimmed away the first and last 5 nts of each aligned read, leaving only the 40 central nucleotides where the mismatch curve was flat.

Data handling

All data handling was performed with parsing scripts, written in C language, acting on plain text files.

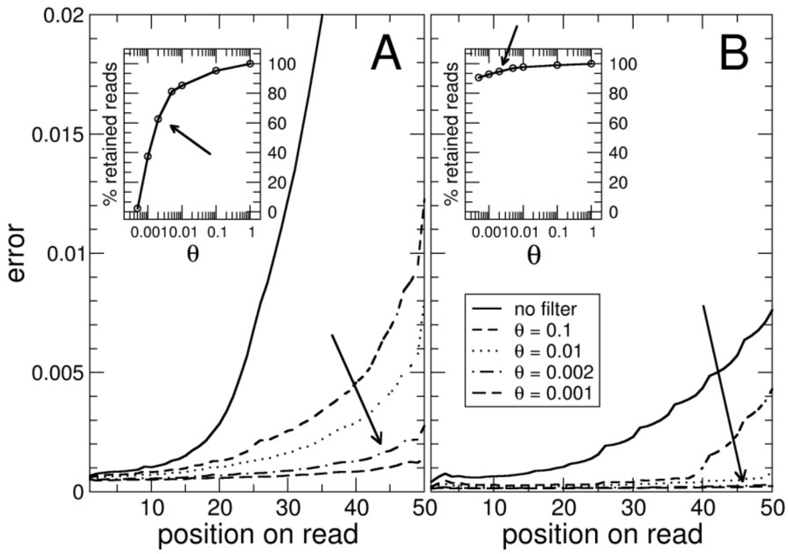


Figure 1

Average error on reads, computed with base qualities. Panel A: first dataset; panel B: second dataset. The average error increases greatly towards the end of the reads (solid lines). The second dataset was less noisy. Different filtering strategies were tested: only the reads whose average error was below a threshold θ were accepted. More stringent thresholds decrease the errors on the reads (small dashed, dotted, dot-dashed and dashed lines). The insets show the fraction of reads retained after the filtering process (using a threshold θ =0.2%) and retaining 66% of the reads in the first dataset and 95% of the second dataset.

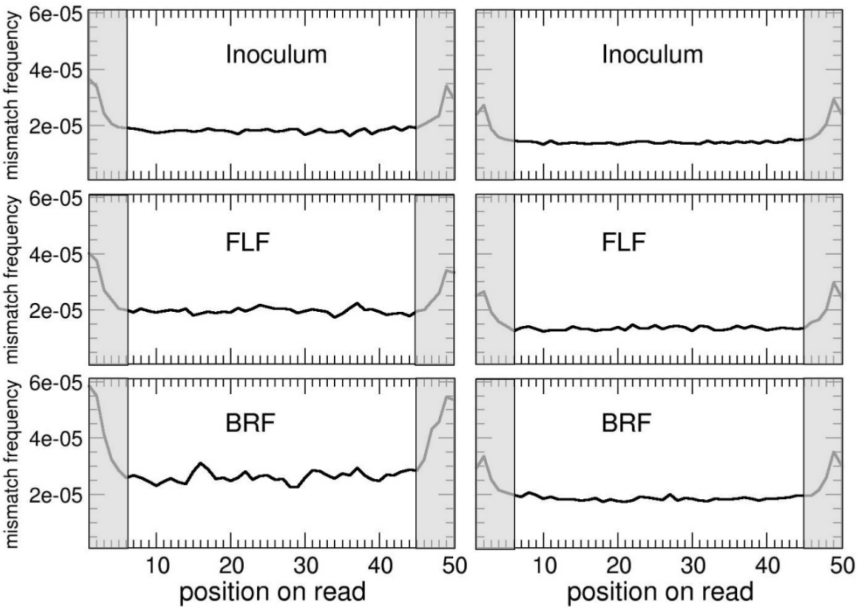


Figure 2:

Distribution of mismatches to the reference genome on the reads after alignment. Left column: first dataset, right column: second dataset. The curves are largely flat, indicating an even distribution of mismatches over the reads, apart for a mild increase towards the edges of the reads, possibly due to reads containing insertions and deletions. We kept only data coming from the flat region of the curve, *i.e.* nucleotides from 5 to 45 in each aligned read.

Appendix 3

Alternative oligonucleotide primers tested for the amplification of the FMDV genomes studied.

	PCR	Primer[1] (OBFS)	Location on genome (region)	Amplicon size (nt)	Primer Sequence (5' to 3')	T_m^{oC}	GC(%)
Optimised-1	1i	516+F	499-520[2] (5'UTR)	4411	CCTTCGCTCGGAAGTAAAACGA	57	50
		4926 R	4908-4926[2] (2C)		AAGTCCTTGCCGTCAGGGT	59	58
	2i	3876 F	3876-3893[2] (VP1)	4303	AAATTGTGGCACCGGTGA	55	50
		8159 R	8142-8161[2] (3'UTR)		ATTAAGGAAGCGGGAAAAGC	53	45
Optimised-3	1iii	516+F	499-520[2] (5'UTR)	4003	CCTTCGCTCGGAAGTAAAACGA	57	50
		4501 R	4481-4501[2] (2C)		GCGATCCAAGCCTTAATCCAG	56	52
	2iii	4035 F	4035-4055[2] (2B)	4009	AACCGGTTAGTGTCCGCATTT	57	48
		8043 R	8019-8043[2] (3D)		GCAGGTAAAGTGATCTGTAGCTTGG	58	48
Optimised-4	2iiii	4035 F	4035-4055[2] (2B)	4120	AACCGGTTAGTGTCCGCATTT	57	48
		8154 R	8135-8154[2] (3'UTR)		AAGCGGGAAAAGCCCTTTCG	59	55

[1] The last letter indicates a Forward (F) or Reverse (R) primer

2 Numbering according to GenBank sequence EU448369

Appendix 4

Table S1 Details on illumina data: number of raw and filtered reads and coverage for both replicates of each sample.

Sample	#read (1)	#filtered read (1)	Coverage (1)	#read (2)	# filtered read (2)	Coverage (2)
A2-2DPFC-PB	2790963	2128716	15351x	2101069	1709078	12325x
A2-3DPFC-SR	2807388	2168346	15637x	2626534	2162173	15592x
A2-4DPFC-PB	2092694	1609323	11605x	2797204	2311235	16667x
A2-4DPFC-SR	2720760	2076448	14974x	2073010	1722013	12418x
A2-5DPFC-SR	3336753	2542656	18336x	2974976	2463627	17766x
A2-6DPFC-BRF	2561453	1962495	14152x	2650213	2214366	15967x
A2-6DPFC-FLF	2704138	2085893	15042x	2657830	2207046	15916x
A2-6DPFC-FRF	2550724	1958710	14125x	2626899	2188872	15785x
A2-6DPFC-PB	2249190	1691592	12199x	2607965	2142547	15451x
A3-1DPFC-PB	2752115	2140025	15432x	2326107	1930781	13924x
A3-3DPFC-PB	2458092	1870211	13487x	2059365	1705933	12302x
A3-3DPFC-SR	2691979	2075898	16330x	2926522	2411961	18974x
A3-4DPFC-SR	4746119	2761230	21721x	5450750	3778399	29723x
A3-5DPFC-BLF	5311265	3079516	24225x	6000979	4094216	32208x
A3-5DPFC-PB	4353393	2469838	19429x	5598961	3724627	29300x
A3-5DPFC-SR	5231498	3049485	23989x	5611303	3891223	30611x
A5-5DPFC-PB	5444899	3238943	25479x	5622686	3931106	30924x
A5-7DPFC-PB	5420473	3249013	25559x	4858806	3410646	26830x

Table S2 NGS data analysis pipeline

Stage 1	Raw Reads	(demultiplexed)
Stage 2	Filtering	Removing all reads with average quality score < 30 (corresponding to probability of error of 0.1%)
Stage 3	Trimming	Removing last nucleotides of each read (3-5 according to quality scores)
Stage 4	Alignment	Aligning the reads to the O1BFS1860 FMDV genome with a simple custom-made scoring routine. Reads with 5 or more mismatches were discarded.
Stage 5	Trimming	Trimming the first and last 5 nucleotides of the aligned reads to remove indels
Stage 6	Masking	Removing from analysis every nucleotide with quality score < 30 (corresponding to probability of error of 0.1%)
Stage 7	Consensus genomes	Determination of consensus genomes by counting the most abundant nt in the reads at every genomic position
Stage 8	Validation	Statistical validation of observed polymorphisms, based on a binomial null distribution. Polymorphisms at frequencies <0.5% were discarded because potentially due to amplification artefacts
Stage 9	Analysis	Generation of quantity of interest: mutation spectra, population distances, Shannon entropy, dN/dS

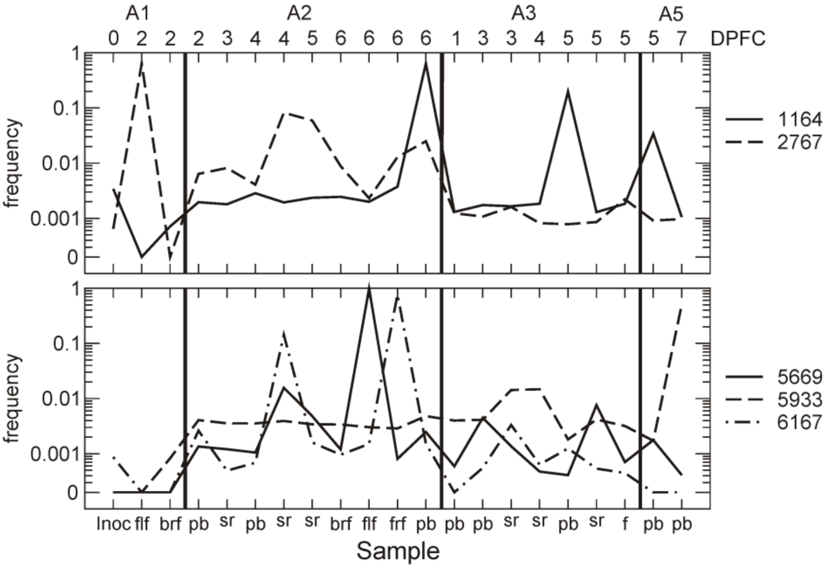


Figure S1

Frequencies across samples of the four remaining mutations reaching consensus in one sample only (for the nine mutations described in the main text, see Figure 4), together with site 2767, previously found mutated in the inoculated calf A1. Top panel: Mutations prevalently present in the probangs. Bottom panel: Mutations present at high frequency in a single sample (6167 is present in a second sample at about 10% frequency).

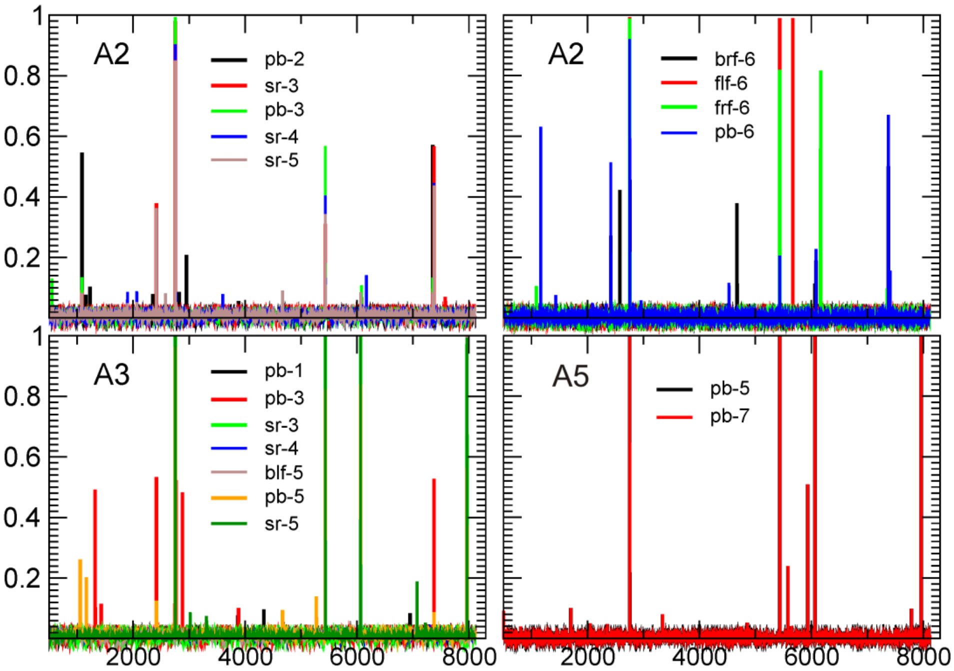


Figure S2

Frequencies of mutations across the genome, computed with respect to the initial inoculum.

References

Abdul-Hamid, N. F., M. Firat-Sarac, A. D. Radford, N. J. Knowles and D. P. King (2011). "Comparative sequence analysis of representative foot-and-mouth disease virus genomes from Southeast Asia." Virus Genes **43**(1): 41-45.

Abdul-Hamid, N. F., N. M. Hussein, J. Wadsworth, A. D. Radford, N. J. Knowles and D. P. King (2011). "Phylogeography of foot-and-mouth disease virus types O and A in Malaysia and surrounding countries." Infect Genet Evol **11**(2): 320-328.

Acharya, R., E. Fry, D. Stuart, G. Fox, D. Rowlands and F. Brown (1989). "The three-dimensional structure of foot-and-mouth disease virus at 2.9 A resolution." Nature **337**(6209): 709-716.

Adessi, C., G. Matton, G. Ayala, G. Turcatti, J. J. Mermod, P. Mayer and E. Kawashima (2000). "Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms." Nucleic Acids Res **28**(20): E87.

Agol, V. I. (2002). Picornavirus Genome: an Overview. Molecular Biology of Picornaviruses. B. L. S. a. E. Wimmer. Washington, D. C., ASM Press**:** 134-135.

Agol, V. I., G. A. Belov, E. A. Cherkasova, G. V. Gavrilin, M. S. Kolesnikova, L. I. Romanova and E. A. Tolskaya (2001). "Some problems of molecular biology of poliovirus infection relevant to pathogenesis, viral spread and evolution." Dev Biol (Basel) **105**: 43-50.

Airaksinen, A., N. Pariente, L. Menendez-Arias and E. Domingo (2003). "Curing of foot-and-mouth disease virus from persistently infected cells by ribavirin involves enhanced mutagenesis." Virology **311**(2): 339-349.

Alexandersen, S., M. E. Bloom and J. Wolfinbarger (1988). "Evidence of restricted viral replication in adult mink infected with Aleutian disease of mink parvovirus." J Virol **62**(5): 1495-1507.

Alexandersen, S. and N. Mowat (2005). "Foot-and-mouth disease: host range and pathogenesis." Curr Top Microbiol Immunol **288**: 9-42.

Alexandersen, S., M. B. Oleksiewicz and A. I. Donaldson (2001). "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.

Alexandersen, S., M. Quan, C. Murphy, J. Knight and Z. Zhang (2003). "Studies of quantitative parameters of virus excretion and transmission in pigs and cattle experimentally infected with foot-and-mouth disease virus." J Comp Pathol **129**(4): 268-282.

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

Alexandersen, S., Z. Zhang, A. I. Donaldson and A. J. Garland (2003). "The pathogenesis and diagnosis of foot-and-mouth disease." J Comp Pathol **129**(1): 1-36.

Alexandersen, S., Z. Zhang, S. M. Reid, G. H. Hutchings and A. I. Donaldson (2002b). "Quantities of infectious virus and viral RNA recovered from sheep and cattle experimentally infected with foot-and-mouth disease virus O UK 2001." J Gen Virol **83**(Pt 8): 1915-1923.

Ali, A., H. Li, W. L. Schneider, D. J. Sherman, S. Gray, D. Smith and M. J. Roossinck (2006). "Analysis of genetic bottlenecks during horizontal transmission of Cucumber mosaic virus." J Virol **80**(17): 8345-8350.

Andino, R., G. E. Rieckhof and D. Baltimore (1990). "A functional ribonucleoprotein complex forms around the 5' end of poliovirus RNA." Cell **63**(2): 369-380.

Anon (April 1969). Report of the Committee of Inquiry on Foot-and-Mouth Disease 1968. Part One, HMSO.

Ansorge, W. J. (2009). "Next-generation DNA sequencing techniques." N Biotechnol **25**(4): 195-203.

Arezi, B. and H. H. Hogrefe (2007). "Escherichia coli DNA polymerase III epsilon subunit increases Moloney murine leukemia virus reverse transcriptase fidelity and accuracy of RT-PCR procedures." Anal Biochem **360**(1): 84-91.

Arias, A., E. Lazaro, C. Escarmis and E. Domingo (2001). "Molecular intermediates of fitness gain of an RNA virus: characterization of a mutant spectrum by biological and molecular cloning." J Gen Virol **82**(Pt 5): 1049-1060.

Arias, A., C. M. Ruiz-Jarabo, C. Escarmis and E. Domingo (2004). "Fitness increase of memory genomes in a viral quasispecies." J Mol Biol **339**(2): 405-412.

Arias, A., J. J. Arnold, M. Sierra, E. D. Smidansky, E. Domingo and C. E. Cameron (2008). "Determinants of RNA-dependent RNA polymerase (in)fidelity revealed by kinetic analysis of the polymerase encoded by a foot-and-mouth disease virus mutant with reduced sensitivity to ribavirin." J Virol **82**(24): 12346-12355.

Arzt, J., N. Juleff, Z. Zhang and L. L. Rodriguez (2011). "The Pathogenesis of Foot-and-Mouth Disease I: Viral Pathways in Cattle." Transboundary and Emerging Diseases **58**(4): 291-304.

Arzt, J., J. M. Pacheco and L. L. Rodriguez (2010). "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.

Bachrach, H. L. (1968). "Foot-and-mouth disease." Annu Rev Microbiol **22**: 201-244.

Bachrach, H. L., R. Trautman and S. S. Breese, Jr. (1964). "Chemical Physical Properties of Virtually Pure Foot-and-Mouth Disease Virus." Am J Vet Res **25**: 333-342.

Baker, S. C., S. R. Bauer, R. P. Beyer, J. D. Brenton, B. Bromley, J. Burrill, . . . R. Zadro (2005). "The External RNA Controls Consortium: a progress report." Nat Methods **2**(10): 731-734.

Baranowski, E., C. M. Ruiz-Jarabo, N. Pariente, N. Verdaguer and E. Domingo (2003). "Evolution of cell recognition by viruses: a source of biological novelty with medical implications." Adv Virus Res **62**: 19-111.

Baranowski, E., C. M. Ruiz-Jarabo, N. Sevilla, D. Andreu, E. Beck and E. Domingo (2000). "Cell recognition by foot-and-mouth disease virus that lacks the RGD integrin-binding motif: flexibility in aphthovirus receptor usage." J Virol **74**(4): 1641-1647.

Barnett, P. V., E. J. Ouldridge, D. J. Rowlands, F. Brown and N. R. Parry (1989). "Neutralizing epitopes of type O foot-and-mouth disease virus. I. Identification and characterization of three functionally independent, conformational sites." J Gen Virol **70 (Pt 6)**: 1483-1491.

Bartley, L. M., C. A. Donnelly and R. M. Anderson (2002). "Review of foot-and-mouth disease virus survival in animal excretions and on fomites." Veterinary Record **151**(22): 667-669.

Bastos, A. D., D. T. Haydon, O. Sangare, C. I. Boshoff, J. L. Edrich and G. R. Thomson (2003). "The implications of virus diversity within the SAT 2 serotype for control of foot-and-mouth disease in sub-Saharan Africa." J Gen Virol **84**(Pt 6): 1595-1606.

Batschelet, E., E. Domingo and C. Weissmann (1976). "The proportion of revertant and mutant phage in a growing population, as a function of mutation and growth rate." Gene **1**(1): 27-32.

Beerenwinkel, N. and O. Zagordi (2011). "Ultra-deep sequencing for the analysis of viral populations." Curr Opin Virol **1**(5): 413-418.

Belsham, G. J. (2005). Translation and replication of FMDV RNA. Foot-and-Mouth Disease Virus. B. W. J. Mahy. Germany, Springer**:** 43-70.

Belsham, G. J. and C. J. Bostock (1988). "Studies on the infectivity of foot-and-mouth disease virus RNA using microinjection." J Gen Virol **69 (Pt 2)**: 265-274.

Bentley, D. R., S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, . . . A. J. Smith (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.

Berinstein, A., M. Roivainen, T. Hovi, P. W. Mason and B. Baxt (1995). "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.

Boeras, D. I., P. T. Hraber, M. Hurlston, T. Evans-Strickfaden, T. Bhattacharya, E. E. Giorgi, . . . E. Hunter (2011). "Role of donor genital tract HIV-1 diversity in the transmission bottleneck." Proc Natl Acad Sci U S A **108**(46): E1156-1163.

Borca, M. V., J. M. Pacheco, L. G. Holinka, C. Carrillo, E. Hartwig, D. Garriga, . . . M. E. Piccone (2012). "Role of arginine-56 within the structural protein VP3 of foot-and-mouth disease virus (FMDV) O1 Campos in virus virulence." Virology **422**(1): 37-45.

Bornberg-Bauer, E. and H. S. Chan (1999). "Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space." Proc Natl Acad Sci U S A **96**(19): 10689-10694.

Borrego, B., I. S. Novella, E. Giralt, D. Andreu and E. Domingo (1993). "Distinct repertoire of antigenic variants of foot-and-mouth disease virus in the presence or absence of immune selection." J Virol **67**(10): 6071-6079.

Botner, A., N. K. Kakker, C. Barbezange, S. Berryman, T. Jackson and G. J. Belsham (2011). "Capsid proteins from field strains of foot-and-mouth disease virus confer a pathogenic phenotype in cattle on an attenuated, cell-culture-adapted virus." J Gen Virol **92**(Pt 5): 1141-1151.

Brackney, D. E., J. E. Beane and G. D. Ebel (2009). "RNAi targeting of West Nile virus in mosquito midguts promotes virus diversification." PLoS Pathog **5**(7): e1000502.

Branton, D., D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, . . . J. A. Schloss (2008). "The potential and challenges of nanopore sequencing." Nat Biotechnol **26**(10): 1146-1153.

Brehm, K. E., N. P. Ferris, M. Lenk, R. Riebe and B. Haas (2009). "Highly sensitive fetal goat tongue cell line for detection and isolation of foot-and-mouth disease virus." J Clin Microbiol **47**(10): 3156-3160.

Brown, C. C., H. J. Olander and R. F. Meyer (1995). "Pathogenesis of foot-and-mouth disease in swine, studied by in-situ hybridization." J Comp Pathol **113**(1): 51-58.

Bull, R. A., J. S. Eden, F. Luciani, K. McElroy, W. D. Rawlinson and P. A. White (2012). "Contribution of intra- and interhost dynamics to norovirus evolution." J Virol **86**(6): 3219-3229.

Bull, R. A., F. Luciani, K. McElroy, S. Gaudieri, S. T. Pham, A. Chopra, . . . A. R. Lloyd (2011). "Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection." PLoS Pathog **7**(9): e1002243.

Burch, C. L. and L. Chao (2004). "Epistasis and its relationship to canalization in the RNA virus phi 6." Genetics **167**(2): 559-567.

Burrows, R., J. A. Mann, A. J. Garland, A. Greig and D. Goodridge (1981). "The pathogenesis of natural and simulated natural foot-and-mouth disease infection in cattle." J Comp Pathol **91**(4): 599-609.

Burrows, R., J. A. Mann, A. Greig, W. G. Chapman and D. Goodridge (1971). "The growth and persistence of foot-and-mouth disease virus in the bovine mammary gland." J Hyg (Lond) **69**(2): 307-321.

Callahan, J. D., F. Brown, F. A. Osorio, J. H. Sur, E. Kramer, G. W. Long, . . . W. M. Nelson (2002). "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**(11): 1636-1642.

Carrillo, C., Z. Lu, M. V. Borca, A. Vagnozzi, G. F. Kutish and D. L. Rock (2007b). "Genetic and phenotypic variation of foot-and-mouth disease virus during serial passages in a natural host." Journal of Virology **81**(20): 11341-11351.

Carrillo, C., E. R. Tulman, G. Delhon, Z. Lu, A. Carreno, A. Vagnozzi, . . . D. L. Rock (2005). "Comparative genomics of foot-and-mouth disease virus." J Virol **79**(10): 6487-6504.

Cases-Gonzalez, C., M. Arribas, E. Domingo and E. Lazaro (2008). "Beneficial effects of population bottlenecks in an RNA virus evolving at increased error rate." J Mol Biol **384**(5): 1120-1129.

Chao, L., C. U. Rang and L. E. Wong (2002). "Distribution of spontaneous mutants and inferences about the replication mode of the RNA bacteriophage phi6." J Virol **76**(7): 3276-3281.

Charleston, B., B. M. Bankowski, S. Gubbins, M. E. Chase-Topping, D. Schley, R. Howey, . . . M. E. Woolhouse (2011). "Relationship between clinical signs and transmission of an infectious disease and the implications for control." Science **332**(6030): 726-729.

Charpentier, N., M. Davila, E. Domingo and C. Escarmis (1996). "Long-term, large-population passage of aphthovirus can generate and amplify defective noninterfering particles deleted in the leader protease gene." Virology **223**(1): 10-18.

Chou, L. S., C. S. Liu, B. Boese, X. Zhang and R. Mao (2010). "DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model." Clin Chem **56**(1): 62-72.

Clarke, B. E., A. L. Brown, K. M. Currey, S. E. Newton, D. J. Rowlands and A. R. Carroll (1987). "Potential secondary and tertiary structure in the genomic RNA of foot and mouth disease virus." Nucleic Acids Res **15**(17): 7067-7079.

Clement, M., D. Posada and K. A. Crandall (2000). "TCS: a computer program to estimate gene genealogies." Mol Ecol **9**(10): 1657-1659.

Cock, P. J., C. J. Fields, N. Goto, M. L. Heuer and P. M. Rice (2010). "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." Nucleic Acids Res **38**(6): 1767-1771.

Coffin, J. M. (1995). "HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy." Science **267**(5197): 483-489.

Cordey, S., T. Junier, D. Gerlach, F. Gobbini, L. Farinelli, E. M. Zdobnov, . . . L. Kaiser (2010). "Rhinovirus genome evolution during experimental human infection." PLoS One **5**(5): e10588.

Cottam, E. M., D. T. Haydon, D. J. Paton, J. Gloster, J. W. Wilesmith, N. P. Ferris, . . . D. P. King (2006). "Molecular epidemiology of the foot-and-mouth disease virus outbreak in the United Kingdom in 2001." J Virol **80**(22): 11274-11282.

Cottam, E. M., D. P. King, A. Wilson, D. J. Paton and D. T. Haydon (2009). "Analysis of Foot-and-mouth disease virus nucleotide sequence variation within naturally infected epithelium." Virus Research **140**(1-2): 199-204.

Cottam, E. M., G. Thebaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, . . . D. T. Haydon (2008). "Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus." Proc Biol Sci **275**(1637): 887-895.

Cottam, E. M., J. Wadsworth, A. E. Shaw, R. J. Rowlands, L. Goatley, S. Maan, . . . N. J. Knowles (2008). "Transmission pathways of foot-and-mouth disease virus in the United Kingdom in 2007." Plos Pathogens **4**(4): -.

Cottral, G. E. and H. L. Bachrach (1968). "Food-and-mouth disease viremia." Proc Annu Meet U S Anim Health Assoc **72**: 383-399.

Crawford, N. M. and D. Baltimore (1983). "Genome-linked protein VPg of poliovirus is present as free VPg and VPg-pUpU in poliovirus-infected cells." Proc Natl Acad Sci U S A **80**(24): 7452-7455.

Crowther, J. R., S. Farias, W. C. Carpenter and A. R. Samuel (1993). "Identification of a fifth neutralizable site on type O foot-and-mouth disease virus following characterization of single and quintuple monoclonal antibody escape mutants." J Gen Virol **74 (Pt 8)**: 1547-1553.

Curry, J., C. McHale and M. T. Smith (2002). "Low efficiency of the Moloney murine leukemia virus reverse transcriptase during reverse transcription of rare t(8;21) fusion gene transcripts." Biotechniques **32**(4): 768, 770, 772, 754-765.

Date, T., T. Kato, J. Kato, H. Takahashi, K. Morikawa, D. Akazawa, . . . T. Wakita (2012). "Novel cell culture-adapted genotype 2a hepatitis C virus infectious clone." J Virol **86**(19): 10805-10820.

Davie, J. (1964). "A Complement Fixation Technique for the Quantitative Measurement of Antigenic Differences between Strains of the Virus of Foot-and-Mouth Disease." J Hyg (Lond) **62**: 401-411.

Devaney, M. A., V. N. Vakharia, R. E. Lloyd, E. Ehrenfeld and M. J. Grubman (1988). "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.

Devonshire, A. S., R. Elaswarapu and C. A. Foy (2010). "Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements." BMC Genomics **11**: 662.

Diaconis, P., S. Goel and S. Holmes (2008). "Horseshoes in Multidimensional Scaling and Local Kernel Methods." Annals of Applied Statistics **2**(3): 777-807.

Diez, J., M. G. Mateu and E. Domingo (1989). "Selection of antigenic variants of foot-and-mouth disease virus in the absence of antibodies, as revealed by an in situ assay." J Gen Virol **70 (Pt 12)**: 3281-3289.

Dohm, J. C., C. Lottaz, T. Borodina and H. Himmelbauer (2008). "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing." Nucleic Acids Res **36**(16): e105.

Domingo-Calap, P., V. Sentandreu, M. A. Bracho, F. Gonzalez-Candelas, A. Moya and R. Sanjuan (2009). "Unequal distribution of RT-PCR artifacts along the E1-E2 region of Hepatitis C virus." J Virol Methods **161**(1): 136-140.

Domingo, E., M. Davila and J. Ortin (1980). "Nucleotide sequence heterogeneity of the RNA from a natural population of foot-and-mouth-disease virus." Gene **11**(3-4): 333-346.

Domingo, E., C. Escarmis, E. Baranowski, C. M. Ruiz-Jarabo, E. Carrillo, J. I. Nunez and F. Sobrino (2003). "Evolution of foot-and-mouth disease virus." Virus Res **91**(1): 47-63.

Domingo, E., C. Escarmis, E. Lazaro and S. C. Manrubia (2005). "Quasispecies dynamics and RNA virus extinction." Virus Res **107**(2): 129-139.

Domingo, E., C. Escarmis, M. A. Martinez, E. Martinez-Salas and M. G. Mateu (1992). "Foot-and-mouth disease virus populations are quasispecies." Curr Top Microbiol Immunol **176**: 33-47.

Domingo, E., C. Escarmis, N. Sevilla, A. Moya, S. F. Elena, J. Quer, . . . J. J. Holland (1996). "Basic concepts in RNA virus evolution." Faseb Journal **10**(8): 859-864.

Domingo, E. and J. J. Holland (1997). "RNA virus mutations and fitness for survival." Annu Rev Microbiol **51**: 151-178.

Domingo, E., V. Martin, C. Perales, A. Grande-Perez, J. Garcia-Arriaza and A. Arias (2006). "Viruses as quasispecies: biological implications." Curr Top Microbiol Immunol **299**: 51-82.

Domingo, E., E. Martinez-Salas, F. Sobrino, J. C. de la Torre, A. Portela, J. Ortin, . . . et al. (1985). "The quasispecies (extremely heterogeneous) nature of viral RNA genome populations: biological relevance--a review." Gene **40**(1): 1-8.

Domingo, E., N. Pariente, A. Airaksinen, C. Gonzalez-Lopez, S. Sierra, M. Herrera, . . . C. Escarmis (2005). "Foot-and-mouth disease virus evolution: exploring pathways towards virus extinction." Curr Top Microbiol Immunol **288**: 149-173.

Domingo, E., D. Sabo, T. Taniguchi and C. Weissmann (1978). "Nucleotide sequence heterogeneity of an RNA phage population." Cell **13**(4): 735-744.

Donnelly, M. L., G. Luke, A. Mehrotra, X. Li, L. E. Hughes, D. Gani and M. D. Ryan (2001). "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.

Drake, J. W. (1993). "Rates of spontaneous mutation among RNA viruses." Proc Natl Acad Sci U S A **90**(9): 4171-4175.

Drake, J. W. and J. J. Holland (1999). "Mutation rates among RNA viruses." Proc Natl Acad Sci U S A **96**(24): 13910-13913.

Dressman, D., H. Yan, G. Traverso, K. W. Kinzler and B. Vogelstein (2003). "Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations." Proc Natl Acad Sci U S A **100**(15): 8817-8822.

Drummond, A. J. and A. Rambaut (2007). "BEAST: Bayesian evolutionary analysis by sampling trees." BMC Evol Biol **7**: 214.

Duarte, E., D. Clarke, A. Moya, E. Domingo and J. Holland (1992). "Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet." Proc Natl Acad Sci U S A **89**(13): 6015-6019.

Duffy, S., L. A. Shackelton and E. C. Holmes (2008). "Rates of evolutionary change in viruses: patterns and determinants." Nat Rev Genet **9**(4): 267-276.

Dulbecco, R. and M. Vogt (1954). "Plaque formation and isolation of pure lines with poliomyelitis viruses." J Exp Med **99**(2): 167-182.

Eckerle, L. D., M. M. Becker, R. A. Halpin, K. Li, E. Venter, X. Lu, . . . M. R. Denison (2010). "Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing." PLoS Pathog **6**(5): e1000896.

Eigen, M. (1971a). "Molecular self-organization and the early stages of evolution." Experientia **27**(11): 149-212.

Eigen, M. (1971b). "Selforganization of Matter and Evolution of Biological Macromolecules." Naturwissenschaften **58**(10): 465-&.

Eigen, M. and P. Schuster (1978). "Hypercycle - Principle of Natural Self-Organization .B. Abstract Hypercycle." Naturwissenschaften **65**(1): 7-41.

Elena, S. F., F. Gonzalez-Candelas and A. Moya (1992). "Does the VP1 gene of foot-and-mouth disease virus behave as a molecular clock?" J Mol Evol **35**(3): 223-229.

Ellard, F. M., J. Drew, W. E. Blakemore, D. I. Stuart and A. M. King (1999). "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.

Elsik, C. G., R. L. Tellam, K. C. Worley, R. A. Gibbs, D. M. Muzny, G. M. Weinstock, . . . F. Q. Zhao (2009). "The genome sequence of taurine cattle: a window to ruminant biology and evolution." Science **324**(5926): 522-528.

Eriksson, N., L. Pachter, Y. Mitsuya, S. Y. Rhee, C. Wang, B. Gharizadeh, . . . N. Beerenwinkel (2008). "Viral population estimation using pyrosequencing." PLoS Comput Biol **4**(4): e1000074.

Erlich, Y., P. P. Mitra, M. delaBastide, W. R. McCombie and G. J. Hannon (2008). "Alta-Cyclic: a self-optimizing base caller for next-generation sequencing." Nat Methods **5**(8): 679-682.

Escarmis, C., M. Davila, N. Charpentier, A. Bracho, A. Moya and E. Domingo (1996). "Genetic lesions associated with Muller's ratchet in an RNA virus." J Mol Biol **264**(2): 255-267.

Escarmis, C., M. Davila and E. Domingo (1999). "Multiple molecular pathways for fitness recovery of an RNA virus debilitated by operation of Muller's ratchet." J Mol Biol **285**(2): 495-505.

Escarmis, C., J. Dopazo, M. Davila, E. L. Palma and E. Domingo (1995). "Large deletions in the 5'-untranslated region of foot-and-mouth disease virus of serotype C." Virus Res **35**(2): 155-167.

Escarmis, C., G. Gomez-Mariano, M. Davila, E. Lazaro and E. Domingo (2002). "Resistance to extinction of low fitness virus subjected to plaque-to-plaque transfers: diversification by mutation clustering." J Mol Biol **315**(4): 647-661.

Escarmis, C., E. Lazaro, A. Arias and E. Domingo (2008). "Repeated bottleneck transfers can lead to non-cytocidal forms of a cytopathic virus: implications for viral extinction." J Mol Biol **376**(2): 367-379.

Escarmis, C., M. Toja, M. Medina and E. Domingo (1992). "Modifications of the 5' untranslated region of foot-and-mouth disease virus after prolonged persistence in cell culture." Virus Res **26**(2): 113-125.

Evans, M., N. A. J. Hastings and J. B. Peacock (2000). Statistical distributions. New York, Wiley.

Ewing, B., L. Hillier, M. C. Wendl and P. Green (1998). "Base-calling of automated sequencer traces using phred. I. Accuracy assessment." Genome Res **8**(3): 175-185.

Ferris, N. P. and M. Dawson (1988). "Routine application of enzyme-linked immunosorbent assay in comparison with complement fixation for the diagnosis of foot-and-mouth and swine vesicular diseases." Vet Microbiol **16**(3): 201-209.

Fischer, W., V. V. Ganusov, E. E. Giorgi, P. T. Hraber, B. F. Keele, T. Leitner, . . . B. T. Korber (2010). "Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing." PLoS One **5**(8): e12303.

Fishman, S. L. and A. D. Branch (2009). "The quasispecies nature and biological implications of the hepatitis C virus." Infect Genet Evol **9**(6): 1158-1167.

Forss, S., K. Strebel, E. Beck and H. Schaller (1984). "Nucleotide sequence and genome organization of foot-and-mouth disease virus." Nucleic Acids Res **12**(16): 6587-6601.

Forster, R., C. Adami and C. O. Wilke (2006). "Selection for mutational robustness in finite populations." J Theor Biol **243**(2): 181-190.

Freistadt, M. S., J. A. Vaccaro and K. E. Eberle (2007). "Biochemical characterization of the fidelity of poliovirus RNA-dependent RNA polymerase." Virology Journal **4**: 44.

Froissart, R., D. Roze, M. Uzest, L. Galibert, S. Blanc and Y. Michalakis (2005). "Recombination every day: abundant recombination in a virus during a single multi-cellular host infection." PLoS Biol **3**(3): e89.

Fry, E. E., S. M. Lea, T. Jackson, J. W. I. Newman, F. M. Ellard, W. E. Blakemore, . . . D. I. Stuart (1999). "The structure and function of a foot-and-mouth disease virus-oligosaccharide receptor complex." Embo Journal **18**(3): 543-554.

Fry, E. E., J. W. Newman, S. Curry, S. Najjam, T. Jackson, W. Blakemore, . . . D. I. Stuart (2005). "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.

Gago, S., S. F. Elena, R. Flores and R. Sanjuan (2009). "Extremely high mutation rate of a hammerhead viroid." Science **323**(5919): 1308.

Garcia-Arriaza, J., S. C. Manrubia, M. Toja, E. Domingo and C. Escarmis (2004). "Evolutionary transition toward defective RNAs that are infectious by complementation." J Virol **78**(21): 11678-11685.

Gelman, A. (2004). Bayesian data analysis. Boca Raton, Fla., Chapman & Hall/CRC.

Glenn, T. C. (2011). "Field guide to next-generation DNA sequencers." Mol Ecol Resour **11**(5): 759-769.

Golde, W. T., T. de Los Santos, L. Robinson, M. J. Grubman, N. Sevilla, A. Summerfield and B. Charleston (2011). "Evidence of activation and suppression during the early immune response to foot-and-mouth disease virus." Transbound Emerg Dis **58**(4): 283-290.

Graves, J. H., J. W. McVicar, P. Sutmoller, R. Trautman and G. G. Wagner (1971). "Latent viral infection in transmission of foot-ana-mouth disease by contact between infected and susceptible cattle." J Infect Dis **124**(3): 270-276.

Grubman, M. J. and B. Baxt (2004). "Foot-and-mouth disease." Clin Microbiol Rev **17**(2): 465-493.

Gu, C. J., C. Y. Zheng, Q. Zhang, L. L. Shi, Y. Li and S. F. Qu (2006). "An antiviral mechanism investigated with ribavirin as an RNA virus mutagen for foot-and-mouth disease virus." J Biochem Mol Biol **39**(1): 9-15.

Haas, B., R. Ahl, R. Bohm and D. Strauch (1995). "Inactivation of viruses in liquid manure." Rev Sci Tech **14**(2): 435-445.

Hall, T. A. (1999). "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT." Nucleic Acids Symp Ser **41**: 95-98.

Harismendy, O. and K. Frazer (2009). "Method for improving sequence coverage uniformity of targeted genomic intervals amplified by LR-PCR using Illumina GA sequencing-by-synthesis technology." Biotechniques **46**(3): 229-231.

Harismendy, O., P. C. Ng, R. L. Strausberg, X. Wang, T. B. Stockwell, K. Y. Beeson, . . . K. A. Frazer (2009). "Evaluation of next generation sequencing platforms for population targeted sequencing studies." Genome Biol **10**(3): R32.

Haydon, D. T., A. D. Bastos and P. Awadalla (2004). "Low linkage disequilibrium indicative of recombination in foot-and-mouth disease virus gene sequence alignments." J Gen Virol **85**(Pt 5): 1095-1100.

Haydon, D. T., A. D. Bastos, N. J. Knowles and A. R. Samuel (2001). "Evidence for positive selection in foot-and-mouth disease virus capsid genes from field isolates." Genetics **157**(1): 7-15.

Haydon, D. T., A. R. Samuel and N. J. Knowles (2001). "The generation and persistence of genetic variation in foot-and-mouth disease virus." Prev Vet Med **51**(1-2): 111-124.

Heath, L., E. van der Walt, A. Varsani and D. P. Martin (2006). "Recombination patterns in aphthoviruses mirror those found in other picornaviruses." J Virol **80**(23): 11827-11832.

Henderson, W. M. (1948). "Further consideration of some of the factors concerned in intracutaneous injection of cattle." J Pathol Bacteriol **60**(1): 137-139.

Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Fewell, D. Barnett, . . . E. R. Mardis (2008). "Whole-genome sequencing and variant discovery in C. elegans." Nat Methods **5**(2): 183-188.

Hofacker, I. L., P. F. Stadler and R. R. Stocsits (2004). "Conserved RNA secondary structures in viral genomes: a survey." Bioinformatics **20**(10): 1495-1499.

Hoffmann, B., M. Scheuch, D. Hoper, R. Jungblut, M. Holsteg, H. Schirrmeier, . . . M. Beer (2012). "Novel orthobunyavirus in Cattle, Europe, 2011." Emerging Infectious Diseases **18**(3): 469-472.

Hoffmann, C., N. Minkah, J. Leipzig, G. Wang, M. Q. Arens, P. Tebas and F. D. Bushman (2007). "DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations." Nucleic Acids Res **35**(13): e91.

Holguin, A., J. Hernandez, M. A. Martinez, M. G. Mateu and E. Domingo (1997). "Differential restrictions on antigenic variation among antigenic sites of foot-and-mouth disease virus in the absence of antibody selection." J Gen Virol **78 (Pt 3)**: 601-609.

Holland, J., K. Spindler, F. Horodyski, E. Grabau, S. Nichol and S. VandePol (1982). "Rapid evolution of RNA genomes." Science **215**(4540): 1577-1585.

Holmes, E. C. (2003). "Error thresholds and the constraints to RNA virus evolution." Trends Microbiol **11**(12): 543-546.

Holmes, E. C. (2010a). "Does hepatitis C virus really form quasispecies?" Infect Genet Evol **10**(4): 431-432.

Holmes, E. C. (2010b). "The RNA Virus Quasispecies: Fact or Fiction?" J Mol Biol.

Holmes, E. C. and A. Moya (2002b). "Is the quasispecies concept relevant to RNA viruses?" Journal of Virology **76**(1): 460-465.

Holste, D., I. Grosse and H. Herzel (2001). "Statistical analysis of the DNA sequence of human chromosome 22." Phys Rev E Stat Nonlin Soft Matter Phys **64**(4 Pt 1): 041917.

Hong, X., H. Doddapaneni, J. M. Comeron, M. J. Rodesch, H. A. Halvensleben, C. Y. Nien, . . . J. R. Manak (2012). "Microarray-Based Capture of Novel Expressed Cell Type-Specific Transfrags (CoNECT) to Annotate Tissue-Specific Transcription in Drosophila melanogaster." G3 (Bethesda) **2**(8): 873-882.

Hope, D. A., S. E. Diamond and K. Kirkegaard (1997). "Genetic dissection of interaction between poliovirus 3D polymerase and viral protein 3AB." J Virol **71**(12): 9490-9498.

Ibanez, A., B. Clotet and M. A. Martinez (2000). "Human immunodeficiency virus type 1 population bottleneck during indinavir therapy causes a genetic drift in the env quasispecies." J Gen Virol **81**(Pt 1): 85-95.

Ishii, K. and M. Fukui (2001). "Optimization of annealing temperature to reduce bias caused by a primer mismatch in multitemplate PCR." Appl Environ Microbiol **67**(8): 3753-3755.

Jackson, A. L., H. O'Neill, F. Maree, B. Blignaut, C. Carrillo, L. Rodriguez and D. T. Haydon (2007). "Mosaic structure of foot-and-mouth disease virus genomes." J Gen Virol **88**(Pt 2): 487-492.

Jackson, T., S. Clark, S. Berryman, A. Burman, S. Cambier, D. Mu, . . . A. M. King (2004). "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.

Jackson, T., F. M. Ellard, R. A. Ghazaleh, S. M. Brookes, W. E. Blakemore, A. H. Corteyn, . . . A. M. King (1996b). "Efficient infection of cells in culture by type O foot-and-mouth disease virus requires binding to cell surface heparan sulfate." Journal of Virology **70**(8): 5282-5287.

Jackson, T., A. P. Mould, D. Sheppard and A. M. King (2002). "Integrin alphavbeta1 is a receptor for foot-and-mouth disease virus." J Virol **76**(3): 935-941.

Jenkins, G. M., M. Worobey, C. H. Woelk and E. C. Holmes (2001). "Evidence for the non-quasispecies evolution of RNA viruses [corrected]." Mol Biol Evol **18**(6): 987-994.

Jiang, L., F. Schlesinger, C. A. Davis, Y. Zhang, R. Li, M. Salit, . . . B. Oliver (2011). "Synthetic spike-in standards for RNA-seq experiments." Genome Res **21**(9): 1543-1551.

Johns, H. L., S. Berryman, P. Monaghan, G. J. Belsham and T. Jackson (2009). "A dominant-negative mutant of rab5 inhibits infection of cells by foot-and-mouth disease virus: implications for virus entry." J Virol **83**(12): 6247-6256.

Jonsson, N., M. Gullberg and A. M. Lindberg (2009). "Real-time polymerase chain reaction as a rapid and efficient alternative to estimation of picornavirus titers by tissue culture infectious dose 50% or plaque forming units." Microbiol Immunol **53**(3): 149-154.

Jridi, C., J. F. Martin, V. Marie-Jeanne, G. Labonne and S. Blanc (2006). "Distinct viral populations differentiate and evolve independently in a single perennial host plant." J Virol **80**(5): 2349-2357.

Juleff, N., Valdazo-Gonzalez, B., Wadsworth, J., Wright, C. F., Charleston, B., Paton, D. J., King, D. P., Knowles, N. J. (2013). "Accumulation of Nucleotide Substitutions Occuring During Experimental Transmission of Foot-and-Mouth Disease Virus." J Gen Virol **93**: 000-000.

Kampmann, M. L., S. L. Fordyce, M. C. Avila-Arcos, M. Rasmussen, E. Willerslev, L. P. Nielsen and M. T. Gilbert (2011). "A simple method for the parallel deep sequencing of full influenza A genomes." J Virol Methods **178**(1-2): 243-248.

Kasambula, L., G. J. Belsham, H. R. Siegismund, V. B. Muwanika, A. R. Ademun-Okurut and C. Masembe (2011). "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.

Khatib, R., J. L. Chason, B. K. Silberberg and A. M. Lerner (1980). "Age-dependent pathogenicity of group B coxsackieviruses in Swiss-Webster mice: infectivity for myocardium and pancreas." J Infect Dis **141**(3): 394-403.

Kinnunen, L., T. Poyry and T. Hovi (1991). "Generation of Virus Genetic Lineages during an Outbreak of Poliomyelitis." J Gen Virol **72**: 2483-2489.

Kitson, J. D., D. McCahon and G. J. Belsham (1990). "Sequence analysis of monoclonal antibody resistant mutants of type O foot and mouth disease virus: evidence for the involvement of the three surface exposed capsid proteins in four antigenic sites." Virology **179**(1): 26-34.

Knowles, N. J. and A. R. Samuel (2003). "Molecular epidemiology of foot-and-mouth disease virus." Virus Res **91**(1): 65-80.

Knowles, N. J., A. R. Samuel, P. R. Davies, R. J. Midgley and J. F. Valarcher (2005). "Pandemic strain of foot-and-mouth disease virus serotype O." Emerging Infectious Diseases **11**(12): 1887-1893.

Konig, G. A., E. M. Cottam, S. Upadhyaya, J. Gloster, L. M. Mansley, D. T. Haydon and D. P. King (2009). "Sequence data and evidence of possible airborne spread in the 2001 foot-and-mouth disease epidemic in the UK." Veterinary Record **165**(14): 410-411.

Krakauer, D. C. and N. L. Komarova (2003). "Levels of selection in positive-strand virus dynamics." J Evol Biol **16**(1): 64-73.

Kuge, S., N. Kawamura and A. Nomoto (1989). "Strong inclination toward transition mutation in nucleotide substitutions by poliovirus replicase." J Mol Biol **207**(1): 175-182.

Le, T., J. Chiarella, B. B. Simen, B. Hanczaruk, M. Egholm, M. L. Landry, . . . M. J. Kozal (2009). "Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use." PLoS One **4**(6): e6079.

Lee, K. N., J. K. Oem, J. H. Park, S. M. Kim, S. Y. Lee, S. Tserendorj, . . . H. Kim (2009). "Evidence of recombination in a new isolate of foot-and-mouth disease virus serotype Asia 1." Virus Res **139**(1): 117-121.

Lewis-Rogers, N., D. A. McClellan and K. A. Crandall (2008). "The evolution of foot-and-mouth disease virus: impacts of recombination and selection." Infect Genet Evol **8**(6): 786-798.

Li, D., Y. J. Shang, Z. X. Liu, X. T. Liu and X. P. Cai (2007). "Molecular relationships between type Asia 1 new strain from China and type O Panasia strains of foot-and-mouth-disease virus." Virus Genes **35**(2): 273-279.

Li, H. and M. J. Roossinck (2004). "Genetic bottlenecks reduce population variation in an experimental RNA virus population." J Virol **78**(19): 10582-10587.

Li, J. J., G. L. Pei, H. X. Pang, A. Bilderbeck, S. S. Chen and S. H. Tao (2006). "A new method for RAPD primers selection based on primer bias in nucleotide sequence data." J Biotechnol **126**(4): 415-423.

Liao, C. L. and M. M. Lai (1992). "RNA recombination in a coronavirus: recombination between viral genomic RNA and transfected RNA fragments." J Virol **66**(10): 6117-6124.

Liu, Z., S. S. Venkatesh and C. C. Maley (2008). "Sequence space coverage, entropy of genomes and the potential to detect non-human DNA in human samples." BMC Genomics **9**: 509.

Lole, K. S., R. C. Bollinger, R. S. Paranjape, D. Gadkari, S. S. Kulkarni, N. G. Novak, . . . S. C. Ray (1999). "Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination." J Virol **73**(1): 152-160.

Luria, S. E. (1951). "The frequency distribution of spontaneous bacteriophage mutants as evidence for the exponential rate of phage reproduction." Cold Spring Harb Symp Quant Biol **16**: 463-470.

Malet, I., M. Belnard, H. Agut and A. Cahour (2003). "From RNA to quasispecies: a DNA polymerase with proofreading activity is highly recommended for accurate assessment of viral diversity." J Virol Methods **109**(2): 161-170.

Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends Genet **24**(3): 133-141.

Maree, F. F., B. Blignaut, T. A. de Beer, N. Visser and E. A. Rieder (2010). "Mapping of amino acid residues responsible for adhesion of cell culture-adapted foot-and-mouth disease SAT type viruses." Virus Res **153**(1): 82-91.

Margeridon-Thermet, S., N. S. Shulman, A. Ahmed, R. Shahriar, T. Liu, C. Wang, . . . R. W. Shafer (2009). "Ultra-deep pyrosequencing of hepatitis B virus quasispecies from nucleoside and nucleotide reverse-transcriptase inhibitor (NRTI)-treated patients and NRTI-naive patients." J Infect Dis **199**(9): 1275-1285.

Marguerat, S., B. T. Wilhelm and J. Bahler (2008). "Next-generation sequencing: applications beyond genomes." Biochem Soc Trans **36**(Pt 5): 1091-1096.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, . . . J. M. Rothberg (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Martinez, M. A., C. Carrillo, F. Gonzalez-Candelas, A. Moya, E. Domingo and F. Sobrino (1991). "Fitness alteration of foot-and-mouth disease virus mutants: measurement of adaptability of viral quasispecies." J Virol **65**(7): 3954-3957.

Mason, P. W., M. J. Grubman and B. Baxt (2003). "Molecular basis of pathogenesis of FMDV." Virus Res **91**(1): 9-32.

Mathews, D. H. (2006). "RNA secondary structure analysis using RNAstructure." Curr Protoc Bioinformatics **Chapter 12**: Unit 12 16.

Mathieu-Daude, F., J. Welsh, T. Vogt and M. McClelland (1996). "DNA rehybridization during PCR: the 'Cot effect' and its consequences." Nucleic Acids Res **24**(11): 2080-2086.

McColl, K. A., H. A. Westbury, R. P. Kitching and V. M. Lewis (1995). "The persistence of foot-and-mouth disease virus on wool." Aust Vet J **72**(8): 286-292.

McVicar, J. W. and P. Sutmoller (1974). "Neutralizing activity in the serum and oesophageal-pharyngeal fluid of cattle after exposure to foot-and-mouth disease virus and subsequent re-exposure." Arch Gesamte Virusforsch **44**(2): 173-176.

McVicar, J. W. and P. Sutmoller (1976). "Growth of foot-and-mouth disease virus in the upper respiratory tract of non-immunized, vaccinated, and recovered cattle after intranasal inoculation." J Hyg (Lond) **76**(3): 467-481.

Medina, M., E. Domingo, J. K. Brangwyn and G. J. Belsham (1993). "The two species of the foot-and-mouth disease virus leader protein, expressed individually, exhibit the same activities." Virology **194**(1): 355-359.

Melchers, W. J., J. M. Bakkers, H. J. Bruins Slot, J. M. Galama, V. I. Agol and E. V. Pilipenko (2000). "Cross-talk between orientation-dependent recognition determinants of a complex control RNA element, the enterovirus oriR." RNA **6**(7): 976-987.

Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

Milan, M., A. Coppe, R. Reinhardt, L. M. Cancela, R. B. Leite, C. Saavedra, . . . L. Bargelloni (2011). "Transcriptome sequencing and microarray development for the Manila clam, Ruditapes philippinarum: genomic tools for environmental monitoring." BMC Genomics **12**: 234.

Mitsuya, Y., V. Varghese, C. Wang, T. F. Liu, S. P. Holmes, P. Jayakumar, . . . R. W. Shafer (2008). "Minority human immunodeficiency virus type 1 variants in antiretroviral-naive persons with reverse transcriptase codon 215 revertant mutations." J Virol **82**(21): 10747-10755.

Moffat, K., G. Howell, C. Knox, G. J. Belsham, P. Monaghan, M. D. Ryan and T. Wileman (2005). "Effects of foot-and-mouth disease virus nonstructural proteins on the structure and function of the early secretory pathway: 2BC but not 3A blocks endoplasmic reticulum-to-Golgi transport." J Virol **79**(7): 4382-4395.

Moffat, K., C. Knox, G. Howell, S. J. Clark, H. Yang, G. J. Belsham, . . . T. Wileman (2007). "Inhibition of the secretory pathway by foot-and-mouth disease virus 2BC protein is reproduced by coexpression of 2B with 2C, and the site of inhibition is determined by the subcellular location of 2C." J Virol **81**(3): 1129-1139.

Monaghan, P., S. Gold, J. Simpson, Z. Zhang, P. H. Weinreb, S. M. Violette, . . . T. Jackson (2005). "The alpha(v)beta6 integrin receptor for Foot-and-mouth disease virus is expressed constitutively on the epithelial cells targeted in cattle." J Gen Virol **86**(Pt 10): 2769-2780.

Moreno, H., H. Tejero, J. C. de la Torre, E. Domingo and V. Martin (2012). "Mutagenesis-mediated virus extinction: virus-dependent effect of viral load on sensitivity to lethal defection." PLoS ONE **7**(3): e32550.

Mullan, B., E. Kenny-Walsh, J. K. Collins, F. Shanahan and L. J. Fanning (2001). "Inferred hepatitis C virus quasispecies diversity is influenced by choice of DNA polymerase in reverse transcriptase-polymerase chain reactions." Anal Biochem **289**(2): 137-146.

Murcia, P. R., G. J. Baillie, J. Daly, D. Elton, C. Jervis, J. A. Mumford, . . . J. L. N. Wood (2010). "Intra- and Interhost Evolutionary Dynamics of Equine Influenza Virus." Journal of Virology **84**(14): 6943-6954.

Murcia, P. R., J. Hughes, P. Battista, L. Lloyd, G. J. Baillie, R. H. Ramirez-Gonzalez, . . . J. L. Wood (2012). "Evolution of an Eurasian avian-like influenza virus in naive and vaccinated pigs." PLoS Pathog 8(5): e1002730.

Murphy, C., J. B. Bashiruddin, M. Quan, Z. Zhang and S. Alexandersen (2010). "Foot-and-mouth disease viral loads in pigs in the early, acute stage of disease." Veterinary Record **166**(1): 10-14.

Murray, K. E. and D. J. Barton (2003). "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.

Murray, K. E., A. W. Roberts and D. J. Barton (2001). "Poly(rC) binding proteins mediate poliovirus mRNA stability." RNA **7**(8): 1126-1141.

Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, . . . S. Kanaya (2011). "Sequence-specific error profile of Illumina sequencers." Nucleic Acids Res **39**(13): e90.

Nayak, A., I. G. Goodfellow, K. E. Woolaway, J. Birtley, S. Curry and G. J. Belsham (2006). "Role of RNA structure and RNA binding activity of foot-and-mouth disease virus 3C protein in VPg uridylylation and virus replication." J Virol **80**(19): 9865-9875.

Neff, S. and B. Baxt (2001). "The ability of integrin alpha(v)beta(3) To function as a receptor for foot-and-mouth disease virus is not dependent on the presence of complete subunit cytoplasmic domains." J Virol **75**(1): 527-532.

Neff, S., P. W. Mason and B. Baxt (2000). "High-efficiency utilization of the bovine integrin alpha(v)beta(3) as a receptor for foot-and-mouth disease virus is dependent on the bovine beta(3) subunit." J Virol **74**(16): 7298-7306.

Nei, M. and T. Gojobori (1986). "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions." Mol Biol Evol **3**(5): 418-426.

Newton, S. E., A. R. Carroll, R. O. Campbell, B. E. Clarke and D. J. Rowlands (1985). "The sequence of foot-and-mouth disease virus RNA to the 5' side of the poly(C) tract." Gene **40**(2-3): 331-336.

Novak, J. E. and K. Kirkegaard (1991). "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.

Novella, I. S., E. A. Duarte, S. F. Elena, A. Moya, E. Domingo and J. J. Holland (1995). "Exponential increases of RNA virus fitness during large population transmissions." Proc Natl Acad Sci U S A **92**(13): 5841-5844.

Novella, I. S., S. F. Elena, A. Moya, E. Domingo and J. J. Holland (1995). "Size of genetic bottlenecks leading to virus fitness loss is determined by mean initial population fitness." J Virol **69**(5): 2869-2872.

Novella, I. S., R. N. Dutta and C. O. Wilke (2008). "A linear relationship between fitness and the logarithm of the critical bottleneck size in vesicular stomatitis virus populations." J Virol **82**(24): 12589-12590.

O'Donnell, V., M. Larocco and B. Baxt (2008). "Heparan sulfate-binding foot-and-mouth disease virus enters cells via caveola-mediated endocytosis." J Virol **82**(18): 9075-9085.

O'Donnell, V., J. M. Pacheco, D. Gregg and B. Baxt (2009). "Analysis of foot-and-mouth disease virus integrin receptor expression in tissues from naive and infected cattle." J Comp Pathol **141**(2-3): 98-112.

O'Donnell, V. K., J. M. Pacheco, T. M. Henry and P. W. Mason (2001). "Subcellular distribution of the foot-and-mouth disease virus 3A protein in cells infected with viruses encoding wild-type and bovine-attenuated forms of 3A." Virology **287**(1): 151-162.

Oleksiewicz, M. B., A. I. Donaldson and S. Alexandersen (2001). "Development of a novel real-time RT-PCR assay for quantitation of foot-and-mouth disease virus in diverse porcine tissues." J Virol Methods **92**(1): 23-35.

Ossowski, S., K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann and D. Weigel (2008). "Sequencing of natural strains of Arabidopsis thaliana with short reads." Genome Res **18**(12): 2024-2033.

Pariente, N., A. Airaksinen and E. Domingo (2003). "Mutagenesis versus inhibition in the efficiency of extinction of foot-and-mouth disease virus." J Virol **77**(12): 7131-7138.

Pariente, N., S. Sierra, P. R. Lowenstein and E. Domingo (2001). "Efficient virus extinction by combinations of a mutagen and antiviral inhibitors." J Virol **75**(20): 9723-9730.

Parker, J. (1971). "Presence and inactivation of foot-and-mouth disease virus in animal faeces." Veterinary Record **88**(25): 659-662.

Parvin, J. D., A. Moscona, W. T. Pan, J. M. Leider and P. Palese (1986). "Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1." J Virol **59**(2): 377-383.

Pena, L., M. P. Moraes, M. Koster, T. Burrage, J. M. Pacheco, F. D. Segundo and M. J. Grubman (2008). "Delivery of a foot-and-mouth disease virus empty capsid subunit antigen with nonstructural protein 2B improves protection of swine." Vaccine **26**(45): 5689-5699.

Peng, H. Z., P. G. Isaacson, T. C. Diss and L. X. Pan (1994). "Multiple PCR analyses on trace amounts of DNA extracted from fresh and paraffin wax embedded tissues after random hexamer primer PCR amplification." J Clin Pathol **47**(7): 605-608.

Pfeiffer, J. K. and K. Kirkegaard (2005). "Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice." PLoS Pathog **1**(2): e11.

Pfeiffer, J. K. and K. Kirkegaard (2006). "Bottleneck-mediated quasispecies restriction during spread of an RNA virus from inoculation site to brain." Proc Natl Acad Sci U S A **103**(14): 5520-5525.

Quan, M., C. M. Murphy, Z. Zhang and S. Alexandersen (2004). "Determinants of early foot-and-mouth disease virus dynamics in pigs." J Comp Pathol **131**(4): 294-307.

Radford, A. D., D. Chapman, L. Dixon, J. Chantrey, A. C. Darby and N. Hall (2012). "Application of next-generation sequencing technologies in virology." J Gen Virol **93**(Pt 9): 1853-1868.

Regoes, R. R., S. Crotty, R. Antia and M. M. Tanaka (2005). "Optimal replication of poliovirus within cells." Am Nat **165**(3): 364-373.

Reysenbach, A. L., L. J. Giver, G. S. Wickham and N. R. Pace (1992). "Differential amplification of rRNA genes by polymerase chain reaction." Appl Environ Microbiol **58**(10): 3417-3418.

Rhodes, T., H. Wargo and W. S. Hu (2003). "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.

Rieder, E., T. Bunch, F. Brown and P. W. Mason (1993). "Genetically engineered foot-and-mouth disease viruses with poly(C) tracts of two nucleotides are virulent in mice." J Virol **67**(9): 5139-5145.

Rouzine, I. M., A. Rodrigo and J. M. Coffin (2001). "Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology." Microbiol Mol Biol Rev **65**(1): 151-185.

Rozera, G., I. Abbate, A. Bruselles, C. Vlassi, G. D'Offizi, P. Narciso, . . . M. R. Capobianchi (2009). "Massively parallel pyrosequencing highlights minority variants in the HIV-1 env quasispecies deriving from lymphomonocyte sub-populations." Retrovirology **6**: 15.

Ruiz-Jarabo, C. M., A. Arias, E. Baranowski, C. Escarmis and E. Domingo (2000). "Memory in viral quasispecies." J Virol **74**(8): 3543-3547.

Ruiz-Jarabo, C. M., N. Pariente, E. Baranowski, M. Davila, G. Gomez-Mariano and E. Domingo (2004). "Expansion of host-cell tropism of foot-and-mouth disease virus despite replication in a constant environment." J Gen Virol **85**(Pt 8): 2289-2297.

Sa-Carvalho, D., E. Rieder, B. Baxt, R. Rodarte, A. Tanuri and P. W. Mason (1997). "Tissue culture adaptation of foot-and-mouth disease virus selects viruses that bind to heparin and are attenuated in cattle." Journal of Virology **71**(7): 5115-5123.

Samuel, A. R. and N. J. Knowles (2001). "Foot-and-mouth disease type O viruses exhibit genetically and geographically distinct evolutionary lineages (topotypes)." J Gen Virol **82**(Pt 3): 609-621.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Sardanyes, J., R. V. Sole and S. F. Elena (2009). "Replication mode and landscape topology differentially affect RNA virus mutational load and robustness." J Virol **83**(23): 12579-12589.

Schaaper, R. M. (1993). "Base selection, proofreading, and mismatch repair during DNA replication in Escherichia coli." J Biol Chem **268**(32): 23762-23765.

Schrag, S. J., P. A. Rota and W. J. Bellini (1999). "Spontaneous mutation rate of measles virus: direct estimation based on mutations conferring monoclonal antibody resistance." Journal of Virology **73**(1): 51-54.

Sedivy, J. M., J. P. Capone, U. L. RajBhandary and P. A. Sharp (1987). "An inducible mammalian amber suppressor: propagation of a poliovirus mutant." Cell **50**(3): 379-389.

Serrano, P., M. R. Pulido, M. Saiz and E. Martinez-Salas (2006). "The 3' end of the foot-and-mouth disease virus genome establishes two distinct long-range RNA-RNA interactions with the 5' end region." J Gen Virol **87**(Pt 10): 3013-3022.

Sevilla, N., C. M. Ruiz-Jarabo, G. Gomez-Mariano, E. Baranowski and E. Domingo (1998). "An RNA virus can adapt to the multiplicity of infection." J Gen Virol **79 (Pt 12)**: 2971-2980.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-1145.

Shendure, J., G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, . . . G. M. Church (2005). "Accurate multiplex polony sequencing of an evolved bacterial genome." Science **309**(5741): 1728-1732.

Sierra, S., M. Davila, P. R. Lowenstein and E. Domingo (2000). "Response of foot-and-mouth disease virus to increased mutagenesis: influence of viral load and fitness in loss of infectivity." J Virol **74**(18): 8316-8323.

Simen, B. B., J. F. Simons, K. H. Hullsiek, R. M. Novak, R. D. Macarthur, J. D. Baxter, . . . M. J. Kozal (2009). "Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naive patients significantly impact treatment outcomes." J Infect Dis **199**(5): 693-701.

Simmonds, P., P. Balfe, J. F. Peutherer, C. A. Ludlam, J. O. Bishop and A. J. Brown (1990). "Human immunodeficiency virus-infected individuals contain provirus in small numbers of peripheral mononuclear cells and at low copy numbers." J Virol **64**(2): 864-872.

Simmonds, P., A. Tuplin and D. J. Evans (2004). "Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence." RNA **10**(9): 1337-1351.

Sipos, R., A. J. Szekely, M. Palatinszky, S. Revesz, K. Marialigeti and M. Nikolausz (2007). "Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis." FEMS Microbiol Ecol **60**(2): 341-350.

Smith, D. B., J. McAllister, C. Casino and P. Simmonds (1997). "Virus 'quasispecies': making a mountain out of a molehill?" J Gen Virol **78 (Pt 7)**: 1511-1519.

Sobrino, F., M. Davila, J. Ortin and E. Domingo (1983). "Multiple genetic variants arise in the course of replication of foot-and-mouth disease virus in cell culture." Virology **128**(2): 310-318.

Sobrino, F., E. L. Palma, E. Beck, M. Davila, J. C. de la Torre, P. Negro, . . . E. Domingo (1986). "Fixation of mutations in the viral genome during an outbreak of foot-and-mouth disease: heterogeneity and rate variations." Gene **50**(1-3): 149-159.

Solmone, M., D. Vincenti, M. C. Prosperi, A. Bruselles, G. Ippolito and M. R. Capobianchi (2009). "Use of massively parallel ultradeep pyrosequencing to characterize the genetic diversity of hepatitis B virus in drug-resistant and drug-naive patients and to detect minor variants in reverse transcriptase and hepatitis B S antigen." Journal of Virology **83**(4): 1718-1726.

Stangegaard, M., I. H. Dufva and M. Dufva (2006). "Reverse transcription using random pentadecamer primers increases yield and quality of resulting cDNA." Biotechniques **40**(5): 649-657.

Sutmoller, P. and J. W. McVicar (1972). "Three variants of foot-and-mouth disease virus type O: exposure of cattle." Am J Vet Res **33**(8): 1641-1647.

Sutmoller, P. and J. W. McVicar (1976). "Pathogenesis of foot-and-mouth disease: the lung as an additional portal of entry of the virus." J Hyg (Lond) **77**(2): 235-243.

Sztuba-Solinska, J., A. Dzianott and J. J. Bujarski (2011). "Recombination of 5' subgenomic RNA3a with genomic RNA3 of Brome mosaic bromovirus in vitro and in vivo." Virology **410**(1): 129-141.

Sztuba-Solinska, J., A. Urbanowicz, M. Figlerowicz and J. J. Bujarski (2011). "RNA-RNA recombination in plant virus replication and evolution." Annu Rev Phytopathol **49**: 415-443.

Tamura, K. and M. Nei (1993). "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.

Tamura, K., D. Peterson, N. Peterson, G. Stecher, M. Nei and S. Kumar (2011). "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods." Mol Biol Evol **28**(10): 2731-2739.

Thebaud, G., J. Chadoeuf, M. J. Morelli, J. W. McCauley and D. T. Haydon (2010). "The relationship between mutation frequency and replication strategy in positive-sense single-stranded RNA viruses." Proc Biol Sci **277**(1682): 809-817.

Tiley, L., A. M. King and G. J. Belsham (2003). "The foot-and-mouth disease virus cis-acting replication element (cre) can be complemented in trans within infected cells." J Virol **77**(3): 2243-2246.

Tsibris, A. M., B. Korber, R. Arnaout, C. Russ, C. C. Lo, T. Leitner, . . . D. R. Kuritzkes (2009). "Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo." PLoS One **4**(5): e5683.

Valdazo-Gonzalez, B., N. J. Knowles, J. Wadsworth, D. P. King, J. M. Hammond, F. Ozyoruk, . . . G. K. Georgiev (2011). "Foot-and-mouth disease in Bulgaria." Veterinary Record **168**(9): 247.

Van Nieuwerburgh, F., R. C. Thompson, J. Ledesma, D. Deforce, T. Gaasterland, P. Ordoukhanian and S. R. Head (2012). "Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination." Nucleic Acids Res **40**(3): e24.

Van Nimwegen, E., J. P. Crutchfield and M. Huynen (1999). "Neutral evolution of mutational robustness." Proc Natl Acad Sci U S A **96**(17): 9716-9720.

Varghese, V., R. Shahriar, S. Y. Rhee, T. Liu, B. B. Simen, M. Egholm, . . . R. W. Shafer (2009). "Minority variants associated with transmitted and acquired HIV-1 nonnucleoside reverse transcriptase inhibitor resistance: implications for the use of second-generation nonnucleoside reverse transcriptase inhibitors." J Acquir Immune Defic Syndr **52**(3): 309-315.

Verdaguer, N., I. Fita, E. Domingo and M. G. Mateu (1997). "Efficient neutralization of foot-and-mouth disease virus by monovalent antibody binding." J Virol **71**(12): 9813-9816.

Victoria, J. G., C. Wang, M. S. Jones, C. Jaing, K. McLoughlin, S. Gardner and E. L. Delwart (2010). "Viral nucleic acids in live-attenuated vaccines: detection of minority variants and an adventitious virus." Journal of Virology **84**(12): 6033-6040.

Vignuzzi, M., J. K. Stone, J. J. Arnold, C. E. Cameron and R. Andino (2006). "Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population." Nature **439**(7074): 344-348.

Villaverde, A., M. A. Martinez, F. Sobrino, J. Dopazo, A. Moya and E. Domingo (1991). "Fixation of mutations at the VP1 gene of foot-and-mouth disease virus. Can quasispecies define a transient molecular clock?" Gene **103**(2): 147-153.

Wang, C., Y. Mitsuya, B. Gharizadeh, M. Ronaghi and R. W. Shafer (2007). "Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance." Genome Res **17**(8): 1195-1201.

Wang, G. P., S. A. Sherrill-Mix, K. M. Chang, C. Quince and F. D. Bushman (2010). "Hepatitis C virus transmission bottlenecks analyzed by deep sequencing." Journal of Virology **84**(12): 6218-6228.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Ward, C. D. and J. B. Flanegan (1992). "Determination of the poliovirus RNA polymerase error frequency at eight sites in the viral genome." J Virol **66**(6): 3784-3793.

Ward, C. D., M. A. Stokes and J. B. Flanegan (1988). "Direct measurement of the poliovirus RNA polymerase error frequency in vitro." J Virol **62**(2): 558-562.

Welch, B. L. (1947). "The generalisation of student's problems when several different population variances are involved." Biometrika **34**(1-2): 28-35.

Wells, V. R., S. J. Plotch and J. J. DeStefano (2001). "Determination of the mutation rate of poliovirus RNA-dependent RNA polymerase." Virus Res **74**(1-2): 119-132.

Whitton, J. L., C. T. Cornell and R. Feuer (2005). "Host and virus determinants of picornavirus pathogenesis and tropism." Nat Rev Microbiol **3**(10): 765-776.

Wilke, C. O. (2001). "Adaptive evolution on neutral networks." Bull Math Biol **63**(4): 715-730.

Wilke, C. O. and C. Adami (2003). "Evolution of mutational robustness." Mutat Res **522**(1-2): 3-11.

Witwer, C., S. Rauscher, I. L. Hofacker and P. F. Stadler (2001). "Conserved RNA secondary structures in Picornaviridae genomes." Nucleic Acids Res **29**(24): 5079-5089.

Wong, K. K., L. C. Stillwell, C. A. Dockery and J. D. Saffer (1996). "Use of tagged random hexamer amplification (TRHA) to clone and sequence minute quantities of DNA--application to a 180 kb plasmid isolated from Sphingomonas F199." Nucleic Acids Res **24**(19): 3778-3783.

Yin, J. L., N. A. Shackel, A. Zekry, P. H. McGuinness, C. Richards, K. V. Putten, . . . G. A. Bishop (2001). "Real-time reverse transcriptase-polymerase chain reaction (RT-PCR) for measurement of cytokine and growth factor mRNA expression with fluorogenic probes or SYBR Green I." Immunol Cell Biol **79**(3): 213-221.

Yuste, E., S. Sanchez-Palomino, C. Casado, E. Domingo and C. Lopez-Galindez (1999). "Drastic fitness loss in human immunodeficiency virus type 1 upon serial bottleneck events." J Virol **73**(4): 2745-2751.

Zagordi, O., L. Geyrhofer, V. Roth and N. Beerenwinkel (2010). "Deep sequencing of a genetically heterogeneous sample: local haplotype reconstruction and read error correction." J Comput Biol **17**(3): 417-428.

Zou, N., S. Ditty, B. Li and S. C. Lo (2003). "Random priming PCR strategy to amplify and clone trace amounts of DNA." Biotechniques **35**(4): 758-760, 762-755.