Micro-evolution of Foot-and-mouth disease virus

Eleanor Myfanwy Cottam

Thesis submitted for the degree of Doctor of Philosophy

University of Glasgow

Department of Environmental and Evolutionary Biology

October 2007

Abstract

Foot-and-mouth disease virus (FMDV) causes an acute vesicular disease of domesticated and wild ruminants, and pigs. The virus is highly variable with 7 serotypes and numerous subtypes. A method was developed to sequence the complete genome of the Pan Asia O serotype FMDV that caused the UK 2001 epidemic. The sequence data was used to quantify the genetic diversity of FMDV that arose over the course of this epidemic. 197 nucleotide substitutions were observed at 191 different sites across the genome, with a ratio of non-synonymous to synonymous change of 0.09. It was estimated that the date at which FMDV first infected livestock in the UK was the 7th February 2001 (95% CI Jan 20th – Feb 19th) using coalescent methods, which is in close agreement to that generated on the basis of lesion aging at the first infected premises. The rate of nucleotide substitution during the epidemic was estimated to be 2.26 x 10^{-5} per site per day (95% CI 1.75-2.80 x 10^{-5}). This was a sufficiently high rate that detailed histories of transmission pathways could be reliably reconstructed and motivated a comprehensive study of genetic changes that arose between infected premises in the Darlington area. This study highlighted the potential of tracing virus transmission between farms using genetic data. A maximum likelihood methodology was proposed to combine epidemiological data (describing the infection profile of individual farms) with the genetic data. Integration of genetic and epidemiological data reduced the number of transmission trees (describing infection transmission within this cluster) that were consistent with the genetic data from 41,472, to 1,728, of which 9 represented 95% of the total likelihood calculated. An average of 4.3 (S.D. 2.1) variant nucleotides within the complete genome, were observed between consecutive farm infections. Difficulties in identifying direct transmission events in this analysis arose mainly as a result of limited data on the extent of genetic variation on a single premises (complicated by variation in animal types, and farm sizes), and the manner in which mutations become fixed within the consensus sequence (i.e. upon replication and selection, and transmission through a 'bottleneck'). This encouraged an analysis of the minimum mutant frequency of Pan Asia O FMDV from a single lesion sampled from naturally infected hosts which led to an estimate of 3.06 x 10^{-4} mutations per nucleotide sequenced (this parameter has previously only been measured in cell culture). In addition an experiment was initiated to measure the spontaneous mutation rate of O_1BFS FMDV. The relevance of this analysis of the evolution of FMDV during the UK 2001 epidemic was demonstrated during the outbreak of O_1BFS FMDV in Surrey in 2007, where complete genome sequencing was used in real time to trace virus spread. Genetic analysis of complete viral genomes generated in real-time revealed a chain of transmission events, predicting undisclosed infected premises, and connecting a second cluster of outbreaks in September to the initial outbreaks in August. Thus, this thesis has pioneered the use of complete genome sequencing for fine scale molecular epidemiology.

Declaration

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

Eleanor M Cottam BSc

Institute for Animal Health

Pirbright Laboratory

Ash Road, Surrey

GU24 0NF

Nid wy'n gofyn bywyd moethus, Aur y byd na'i berlau mân
Gofyn rwyf am calon hapus, Calon onest, calon lân.

Calon lân yn llawn daioni, Tecach yw na'r lili dlos
Does ond calon lân all ganu, Canu'r dydd a chanu'r nos.

Pe dymunwn olud bydol, Chwim adenydd iddo sydd
Golud calon lân rinweddol, Yn dwyn bythol elw fydd.

Hwyr a bore fy nymuniad, Esgyn ar adenydd cân
Ar i Dduw, er mwyn fy Ngheidwad, Roddi imi galon lân.

Acknowledgments

First and foremost, I am indebted to my three supervisors Dan, Don, and David for the support, encouragement and scientific foundation they have given to me and this project. They have been an inspiration, persuading me to question, explore and enjoy my PhD. Thank you for believing in me, and in the project. In the lab I would not have succeeded without the help of Nigel and Geoff, and sound advice from Claudia, Veronica, Nick and Scott. Thank you to you and all others in the lab; Jacq, Juliet, Eoin, Caroline, Laura, Jemma, Katja, and Kate, that have made my PhD fun! In studying the 2001 UK outbreak, meeting Sam Mansley was not only a marvellous stroke of luck, but a great pleasure and I am grateful to the insight he has given me to the true events of that devastating epidemic. I am also thankful for the advice and assistance that Andrew King, and John Gloster have given me, and to Lara Harrup for help with the maps and Anthony Wilson for statistics advice. I must also say thank you to Gael, who has helped take the project forward and assisted me greatly in Chapter 4 of this thesis. Outside of work there are many people who have supported me socially and emotionally; here's a big thank you to the Institute bar and all who frequent it for keeping me sane; Paul, Neil, Sparks, Sambit, John, Luke, Josh, Emma, Becks, and to Alice for making sure I have fun. Also thank you to my family, Mummy, Daddy, Sophia and Andrew, Gwen and Will, and Thomas, without whom I would have given up a long time ago. And finally I thank Andrew for helping me with all things PCR, and always being there for me to shout at, cry at and hug!

Table of Contents

List of Figures

List of Tables

Glossary

Capsid	Protein coat encapsulating virus nucleic acid
Consensus nucleotide sequence	The representation of the predominant nucleotide at each position within a variant population
Convergent evolution	The independent evolution of similar traits in two or more unrelated or distantly related lineages
Error threshold	Maximum mutation rate that is tolerable for a given genome size.
Genetic Drift	Stochastic changes in nucleotide frequencies in a finite population due to the random sampling of genetic material at reproduction/transmission
Microevolution	The occurrence of small-scale genetic changes in a population, over a few generations, also known as genetic change at or below the species level
Molecular clock	The theory that the evolution of genetic or protein sequences proceeds at a constant rate
Muller's Ratchet	The successive build up of deleterious mutations in finite asexual populations
Mutation frequency of a population	The number of mutations from the consensus per number of nucleotides sequenced
Mutation rate	The rate at which mutations are introduced into a nucleotide sequence
Natural selection	The process by which a favourable phenotype that is heritable becomes more common in successive generations of a population
Non-synonymous substitution	A nucleotide change that results in an amino acid change in the translated protein
Poly 'A' tract	Long stretch of 'A' residues at the 3' end of the FMDV genome
Poly 'C' tract	Long stretch of 'C' residues within the 5'UTR of the FMDV genome that separates the s- and l-fragments
Population Bottleneck	A severe reduction in population size that causes a loss of genetic variation. Under strong bottlenecks, genetic drift can be more important in determining the fixation of mutations than natural selection

Positive selection	Active selection of a nucleotide sequence through natural selection of the phenotype
Purifying selection	Removal of deleterious mutations from a population
Synonymous substitution	A nucleotide change does not alter the amino acid that is translated
Transition	A substitution involving a change from a purine to a purine or a pyrimidine to a pyrimidine
Transversion	A substitution from a purine to a pyrimidine or vice versa

Abbreviations

°C	degrees Celsius
A/T/G/C	adenine/thymine/guanine/cytosine
BFS	British field sample
BHK	baby hamster kidney
BTY	bovine thyroid
cDNA	complementary DNA
CI	confidence interval
CO_2	carbon dioxide
cre	*cis*-acting replication element
d.f.	degrees of freedom
DEFRA	Department for environment, farming and rural affairs
DMSO	dimethyl sulfoxide
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
ELISA	enzyme linked immunosorbent assay
ER	endoplasmic reticulum
FCS	foetal calf serum
FMD	foot-and-mouth disease
FMDV	foot-and-mouth disease virus
g	unit of acceleration
g	gram
HCV	hepatitis C virus
HIV	human immunodeficiency virus
HS	heparan sulphate
IAH	Institute for Animal Health
IgG	immunoglobulin G
IP	infected premises
IRES	internal ribosome entry site
kb	kilobase
km	kilometre
KS	Kolmorogov Smirnov
LB	Luria-Bertani broth
M	molar
MAb	monoclonal antibody
MAFF	Ministry of agriculture, fisheries and food
MCMC	Markov Chain Monte Carlo
$MgCl_2$	magnesium chloride
$MgSO_4$	magnesium sulphate
ml	millilitre
mM	millimolar
mm	millimetre

moi	Multiplicity of infection
nm	nanometre
NMB	national movement ban
nt	nucleotide
OD	optical density
PCR	polymerase chain reaction
Pen/Strep	Penicillin/Streptomycin
PFU	plaque forming units
pH	potential of hydrogen
pmol	picomole
RACE	Rapid amplification of cDNA ends
RNA	ribonucleic acid
rpm	revolutions per minute
RT	reverse transcriptase
RT-PCR	reverse transcriptase - polymerase chain reaction
S	Svedberg unit
SARS	severe acute respiratory syndrome
SAT	South African Territories
SD	standard deviation
siRNA	silencing RNA
TCID	tissue culture infectious dose
TPB	Tryptose Phosphate Broth
UK	United Kingdom
UTR	untranslated region
UV	ultra violet
v	volts
VP1-4	viral protein 1-4
WRL	World Reference Laboratory
X-gal	5-bromo-4-chloro-3-indolyl- beta-D-galactopyranoside
$\alpha_v\beta$	alpha v beta
µ	micro
X^2	chi-squared

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 its causative agent

1.1 Introduction

1.1.1 Summary

This chapter provides an introduction to foot-and-mouth disease (FMD) and the virus that causes it. The virus genetic and protein structure is discussed followed by a detailed description of the replication cycle of the virus. Section 1.2 provides background information regarding the UK 2001 FMD outbreak, which is studied extensively during this thesis. Sections 1.3 and 1.4 discuss the evolution of the virus, and finally the rationale and objectives of the whole project are considered in section 1.5.

1.1.2 Foot-and-mouth disease

FMD is an extremely contagious disease affecting cattle, pigs, sheep, goats, and other cloven-hoofed animals. The disease causes typically only low mortality in adult animals, but frequently a high mortality in young animals due to myocarditis. Although the majority of animals recover, the disease leaves them debilitated causing severe losses in meat and milk production. This, combined with the disruption of international livestock trade and the cost of disease eradication, results in severe economic consequences of FMDV outbreaks.

FMD is endemic in parts of Asia, Africa and South America, with sporadic outbreaks occurring in the FMD free areas within these continents. The causative agent is foot-and-mouth disease virus (FMDV); a member of the virus family *Picornaviridae*, genus *Aphthovirus*. Transmission of the virus can be direct from animal contact or indirect via fomites carried on animate/inanimate vectors such as humans or vehicles, or airborne in aerosols.

1.1.3 Clinical disease

The disease is characterised by fever and blister-like lesions or vesicles occurring on the tongue, dental pad, gums, hard palate, lips, nostrils, muzzle, coronary bands, teats,

udder, snout of pigs, and inter-digital spaces. Clinically, these lesions result in excessive salivating and lameness. In sheep and goats, the lesions are usually less pronounced. FMD is clinically indistinguishable from vesicular stomatitis, swine vesicular disease and vesicular exanthema of pigs, which are notifiable diseases for this reason. The pathogenesis and diagnosis of FMD has been reviewed by Alexandersen *et al* 2003c.

1.1.4 Foot-and-mouth disease virus

The causative agent of FMD, first discovered by Loeffler and Frosch in 1898, is a positive strand RNA virus. The FMDV genome is approximately 8500 nucleotides in length (Forss et al. 1984), contained within a non-enveloped icosahedral capsid about 30nm in diameter. The RNA on its own is infectious if inside a cell. The virus is sensitive to acidic conditions (<pH6.0) (Randrup 1954), temperatures above 50°C and UV light, but persistance in contaminated fodder and the environment for up to 1 month is possible in temperate climates.

There are seven serotypes of the virus, namely A, C, O, Asia 1, and South African Territories (SAT) 1-3. The serotypes are determined by the antigenic structure of the virus capsid. Sixty copies of four structural proteins come together to make up the virus capsid (Wild et al. 1969); VP1, VP2, VP3 on the surface and VP4 internally in contact with the RNA. VP1, 2 and 3 proteins have a similar eight-stranded beta-barrel topology, although there is no sequence homology between the genes, whereas VP4 has a distinct extended conformation (Acharya et al. 1989; Domingo et al. 2002). There are 5 known neutralising antigenic sites located within the capsid structure.

Field strains of FMDV attach to cells using integrin receptors $\alpha_v\beta_6$, $\alpha_v\beta_3$ and $\alpha_v\beta_8$ (Jackson et al. 2003; Jackson et al. 2000). The virus enters the cell via receptor-mediated endocytosis, and uncoating occurs due to acidification within the endosomal pathway.

1.1.5 Structure of the Foot-and-mouth disease virus genome

The positive sense RNA genome of FMDV is covalently attached to a protein (VPg) at its 5' end. The genome consists of a protein coding section flanked by two untranslated regions (UTR) (Figure 1.1). The RNA is highly structured throughout the genome with the 5' and 3' UTRs containing specific structures. The 5' UTR consists of the s fragment, a poly 'C' tract, several pseudoknots, a *cis*-acting replication element (*cre*), and an Internal Ribosome Entry Site (IRES).

The s fragment is approximately 360 nucleotides in length and predicted to form a hairpin structure (Witwer et al. 2001), it's precise function is unknown as yet, although it is possible that it prevents exonuclease digestion of the virus genome in infected cells (Mason et al. 2003a). Although the importance and role of the poly 'C' tract is uncertain, if shortened to 2 residues, it has been shown to elongate again in cell culture (Rieder et al. 1993).

The precise role of pseudoknots is again unknown. The *cre* element is involved in genome replication as described later and the IRES is involved in cap independent translation of the genome. The 3' UTR is shorter than the 5' UTR, and contains 2 stem-loops and a poly 'A' tract. The s fragment and the IRES have recently been shown to interact specifically with the 3' UTR which could play a role in translation and or replication of the genome (Serrano et al. 2006).

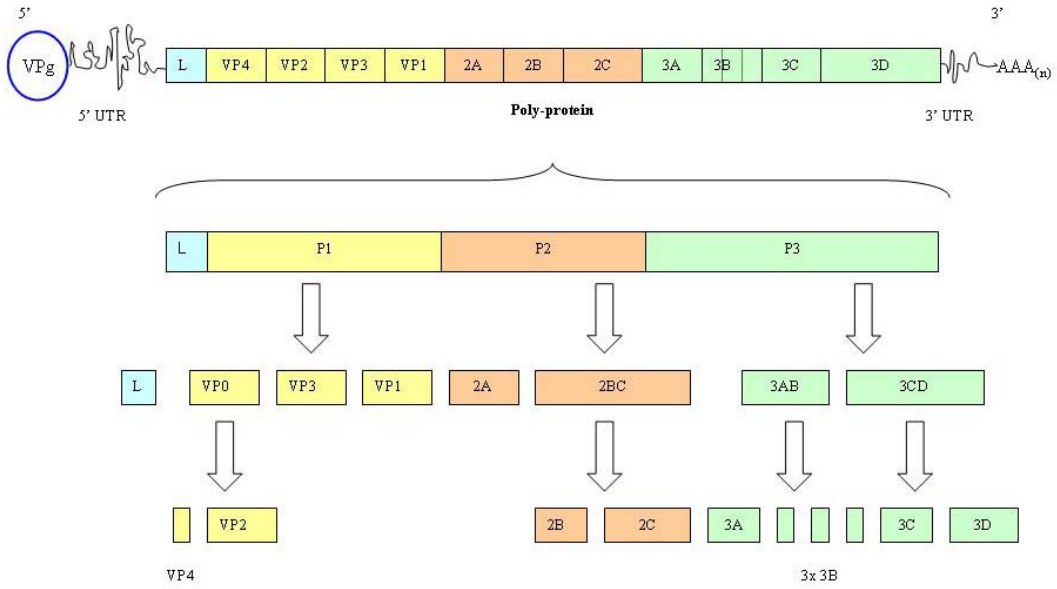


Figure 1.1

Schematic of the FMDV genome showing the translated polyprotein and final cleaved proteins. The schematic represents the layout of the FMDV RNA genome which is covalently attached to the protein VPg at the 5' end. Flanking a single open reading frame are two untranslated regions (UTR). P1 represents the capsid coding region for viral proteins (VP) 1 to 4. The remainder encodes non-structural proteins. Block arrows represent cleavage sites that produce the final mature proteins. (Racaniello 1996)

1.1.6 Foot-and-mouth disease virus replication

FMDV replication is thought to mirror that of the picornavirus Poliovirus, taking place solely within the cytoplasm of the cell (Follett et al. 1975). The molecular aspects of FMDV have been reviewed recently (Mason et al. 2003a). On entry into a cell the genome is immediately translated. The VPg protein is cleaved by a cellular enzyme and the IRES directs translation in a cap independent manner to produce a polyprotein. This polyprotein then undergoes cleavage, resulting in various structural and non-structural proteins which assist in the replication of the genome.

1.1.6.a The Proteins

The viral capsid proteins are encoded by the P1 region, and regions P2 and P3 encode the non structural proteins involved in protein processing (2A, 3C) and genome replication (2B, 2C, 3AB, 3B, 3CD, and 3D) (Figure 1.1). The L protein is a protease; one of its cleavage targets is host cell initiation factor eIF4G whose cleavage leads to shut off of host cap dependent translation (Medina et al. 1993; Racaniello 1996). Protein 2A mediates auto cleavage at its C terminus, and is also thought to have a role in viral RNA replication, although its exact function is not known. It has been determined that 2B and 2C work together to inhibit the secretary pathway blocking transport from the endoplasmic reticulum to the golgi apparatus, 2C determining the site of the block (Moffat et al. 2007). 2B and 3C have been found to be the most conserved genes inter-serotypically (Carrillo et al. 2005). 3A co-localises with 3D in the ER membrane associated replication complexes assisting in viral RNA synthesis by 3D and stimulating cleavage of the 3CD precursor. The strategy of translation of a polyprotein followed by cleavage allows for protein expression to be controlled by the extent and rate of proteolytic processing.

1.1.6.b The genome

Replication of the genome occurs via a negative-stranded RNA intermediate (Racaniello 1996). The VPg protein is used to prime the replication reactions following addition of two uridylate (U) residues to its tyrosine residue. The replication of the genome is carried out by the 3D protein (RNA Polymerase) with the help of the *cis*-acting replication element (*cre*) (Bedard & Semler 2004; Goodfellow et al. 2003). The *cre* element serves as a template for 3D mediated uridylation of VPg which serves as a primer for RNA replication (Mason et al. 2002). The 3' poly 'A' tract is the initiation site for the synthesis of negative sense RNA. A double-stranded RNA intermediate termed RF (replicative form) is formed at this stage. Newly synthesised negative strand templates provide the templates for multiple initiation events resulting in positive strand genome RNA. VPg is

found attached to the 5' ends of both the positive and negative strands (Bedard & Semler 2004). The RNA synthesis is asymmetric in that the synthesis of positive strands can be 30-50 times more than that of negative strands (Novak & Kirkegaard 1991). The RNA dependent polymerase of FMDV is reported to have low fidelity and is predicted to give rise to mutant progeny with a mutation rate of 1 per 10^3 to 10^5 nucleotides copied (Holland et al. 1982; Sobrino et al. 1983).

1.1.6.c Virus assembly

The 60-subunit capsid particle self-assembles, beginning with 5S protomers of VP0, VP3, and VP1, which join to produce a 14S pentamer, which subsequently creates the 80S empty capsid. Only positive strand RNA and not negative strand viral RNA or viral mRNA, is incorporated producing the 150S provirion. The final maturation step involves cleavage of VP0 to VP2 and VP4 resulting in the 160S virion which is then released (Mason et al. 1994; Racaniello 1996).

1.2 UK 2001 Outbreaks of Foot-and-mouth disease

1.2.1 Overview

FMD was confirmed present in the UK for 214 days in 2001, from 20th February until 30th September. The main affected areas were Dumfries and Galloway, Cumbria, Northumberland, Yorkshire, Lancashire, Anglesey, Powys, Hereford, and Devon (Gibbens et al. 2001; Gibbens & Wilesmith 2002). There were also a limited number of outbreaks in Ireland, France and the Netherlands. The virus responsible for the epidemic was characterised as the Pan Asia O strain, with the closest sequenced relative being from South Africa in 2000 (Mason et al. 2003b). During the epidemic 2030 premises were declared infected (Figure 1.2) and culled. A further 8131 premises were culled as dangerous contacts or as a result of 3km or contiguous culls; the cumulative toll of animals killed being at least 6.5 million (Scudamore 2002). The economic cost of the outbreak to agriculture and the food chain in the UK was estimated to exceed £3 billion, with similar costs to the tourist industry (Thompson et al. 2002).

1.2.2 The main events

A single primary outbreak triggered the epidemic, identified as Burnside Farm, Heddon-on-the-Wall, Northumberland (Scudamore 2002). It is likely that clinical signs of virus were present on this farm for at least a week prior to the first laboratory confirmation of disease (Anderson 2002; Scudamore 2002). Transport of pigs to Cheale's abattoir in Essex resulted in the outbreaks identified in the Southeast of England on 19th February 2001, confirmed positive for FMDV the following day. Presumed airborne spread from Burnside Farm infected Prestwick Hall Farm, Ponteland, 7km away (Gloster & Alexandersen 2004) and sheep from this farm were taken to market before the first case had been identified (Alexandersen et al. 2003a). These sheep subsequently moved through Hexham and Longtown markets to 96 locations resulting in outbreak clusters throughout the UK.

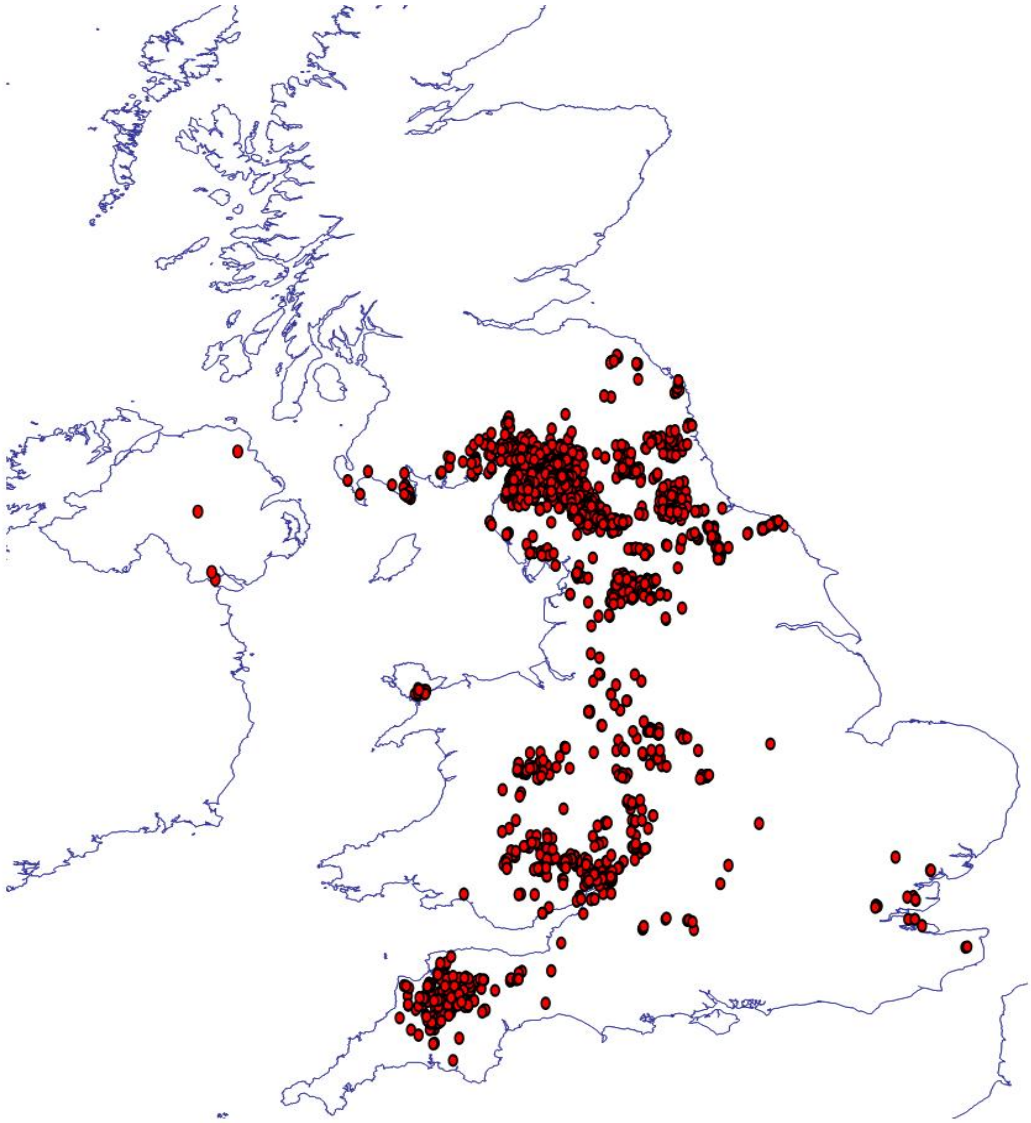


Figure 1.2

The figure shows a map of the United Kingdom with the location of all 2030 designated FMDV infected premises in 2001 depicted by a red dot.

On the 20th February 2001, movement restrictions were placed around farms in Stroud, Gloucestershire, Great Horwood, Buckinghamshire and Isle of Wight, following contact tracing from Cheale's abattoir in Essex where disease was detected. Two days later, an 8km restriction was placed around Burnside farm in Northumberland, while it was investigated by Defra (Department for environment, food and rural affairs, formerly known as MAFF). Also on the 21st of February a ban on export of UK livestock, meat products and milk was instigated. At the same time, the Countryside Alliance banned

hunting throughout the UK to help prevent the spread of disease. Foot-and-mouth disease was confirmed the next day (23rd February) on Burnside farm, Heddon-on-the-Wall. At this point, the agriculture minister ordered a 7 day ban on all livestock movements excluding horses, milk and poultry, throughout the UK. Even with a national movement ban (NMB) in place by the end of February it was clear that the virus was disseminated throughout the UK with over 30 confirmed outbreaks. Footpaths in all infected areas were closed on Tuesday 27th February, adding to losses to the tourist industries. On the 15th March, it was decided in Cumbria, Dumfries and Galloway, to not only cull animals on infected premises but also all those in a 3km ring surrounding each infected premise. Later on in March, this policy was extended to some other areas, but also other control measures such as a cull of contiguous premises within 48 hours, and a 24 hour report to slaughter policy were implemented to various degrees throughout the UK (Anderson 2002).

Despite these rigorous control measures, the epidemic continued until September. It is thought that the absence of obvious clinical signs in sheep infected with virus hampered the attempts to control the epidemic by allowing virus to replicate in an area for a length of time before being detected. The outbreak demonstrated the highly infectious nature of the virus, as it spread between farms with no animal movements and in the majority of cases there is little evidence to suggest that it spread via the airborne route. This suggests that once the movement standstill had been put in place the majority of transmissions in the outbreak were caused by contaminated vehicles or persons.

Many studies since 2001 have attempted to characterise virus spread during the outbreak with a focus on geographical distance, temporal infection characteristics, and airborne fomites (Gloster et al. 2003; Honhold et al. 2004; Keeling et al. 2001; Taylor et al. 2004). However, the source of infection for many farms remains elusive and further research is required to improve our understanding of how the virus spread so successfully.

1.3 Foot-and-mouth disease virus evolution

1.3.1 Overview of RNA virus evolution

The genetic evolution of RNA viruses is governed by the interplay of forces such as mutation, recombination, natural selection, and genetic drift.

Studies of RNA virus evolution primarily consider the dynamics of mutation, selection and genetic drift in relation to the two important RNA virus properties of 1) explosive replication to produce vast population sizes and 2) high mutation rates acting on a short genome.

The fast generation of large populations (~100,000 viral copies in 10 hours) suggests that natural selection should be able to work efficiently on a complete array of all possible mutants; RNA viruses should evolve in a deterministic manner with the fate of a given mutation being determined by its initial frequency, the mutation rate and the probability of selection of that particular mutant. However, while individual populations may be large the *effective* population is often potentially much smaller. The effective population is influenced by the probability of population bottlenecks caused by migration within a host to different replication sites, transmission between hosts and between populations of hosts, and could also be influenced by variation in replication potential among virus variants. It has been predicted that in populations where the effective population is small, that genetic drift is critical in determining the frequency and fate of mutations (Rouzine et al. 2001) and hence evolution progresses in a much less predictable fashion.

It has been proposed that RNA viruses evolve to maximise their mutation rate such that beneficial mutations can arise at a rate sufficient for the requirement to adapt, yet deleterious mutations are kept at a level at which they can be excised from the

population by the forces of selection (Domingo et al. 1978; Holland et al. 1982). The high mutation rates of RNA viruses are thought to be partly responsible for their limited genome sizes as a rate too high would prevent reproduction of the fittest genome and cause a reduction in viral population size, possibly resulting in eventual extinction (Holmes 2003). Coronaviruses are an exception to the rule because they have unusually large genomes of 16 to 31Kb. Recent studies suggest however that these viruses encode specific enzymes that increase their replication fidelity (Eckerle et al. 2007). However, high mutation rates of RNA viruses may not occur solely to improve virus adaptability, because the vast majority of mutations are deleterious. Another possibility could be that the low fidelity of RNA virus polymerases is a by-product of selection for fast replication. This can also explain the limited genome size, as shorter genomes would be favoured by natural selection due to their faster replication. A trade off between replication efficiency and replication fidelity has been documented (Back et al. 1996). It has yet to be proved that increases in the mutant spectrum within a population due to low fidelity replication result in an increased rate of adaptation (Keulen et al. 1999; Lee et al. 1997).

Recombination has been shown to be an important evolutionary mechanism for some RNA viruses. An extreme example is HIV where the genomic recombination rate exceeds the genomic mutation rate (Rhodes et al. 2003). Recombination may play more of a role for some viruses than others and its importance in FMDV evolution is uncertain although recombination between FMDV isolates has been shown to occur (Heath et al. 2006; Jackson et al. 2007); Haydon, 2004 #56}. Indeed, very recently a recombined Asia 1 and O Pan Asia FMD virus was reported (Li et al. 2007).

1.3.2 Constraints and limitations on FMDV evolution and mutation

FMDV may have evolved to preserve the ability to adapt quickly, by maintaining an error prone RNA dependent RNA polymerase for its replication. There are other virus properties that could also promote viable mutation. Firstly, within the FMDV genome there are no overlapping open reading frames, which are often found in other viruses such as Paramyxoviruses. The lack of overlapping open reading frames reduces the constraint on viable mutation as for each possible point mutation only one protein is affected. Secondly, in other positive strand RNA viruses (e.g. Poliovirus, Cardiovirus, Human Rhinitis Virus 2) the *cre* is encoded within the open reading frame forcing a requirement that any mutation within this region be viable for both the protein and the *cre* element. The FMDV *cre* element is encoded within the 5'UTR relaxing the constraint on viable mutation.

Although the above properties of FMDV can be thought to promote successful virus mutation and evolution it is important to realise the numerous constraints that restrict viable virus mutation. Many studies have investigated an alleged error threshold, which is a limit on acceptable mutation rates (Domingo et al. 2005; Holmes 2003; Pariente et al. 2003). Beyond this limit, 'fit' genomes are unable to precisely reproduce themselves resulting in the accumulation of deleterious mutations and population extinction. Another limit on the rate of mutation could be epistasis where for a particular mutation to be viable it is necessary for a compensatory change to occur elsewhere in the genome to counteract any deleterious effect, thus requiring multiple point mutations and not 1 for a viable progeny virus.

When a non-synonymous mutation occurs the resulting amino acid change may negatively affect protein function and this may be expected to constrain possible replacement of amino acids. Synonymous changes that don't affect the translated product may still result in non viable virus progeny, since RNA structure has been

increasingly found to influence virus pathogenesis and survival (Simmonds et al. 2004). It is also possible that the specific secondary structure of the RNA could encourage or discourage the polymerase to introduce an error in certain areas of the genome. For example it might be advantageous to have a higher rate of change in the capsid protein genes than in the IRES structure.

Finally, an important constraint on FMDV evolution is the continually changing environment. The virus may benefit from the ability to escape the host immune response, and to transmit to and successfully infect a new host animal (of a number of species). Potentially virus could evolve to reach a high fitness that is specific to a particular environment, which could be detrimental upon spread to a new host. The fast spread of FMD may prevent this occurring, by restricting the time available for the virus to adapt to a specific host.

1.3.3 Quasispecies

Many RNA viruses are often described as quasispecies (Domingo 2002; Moya et al. 2004). For a quasispecies population the target of natural selection is not the fastest growing replicator but a broad spectrum of mutants that are produced by erroneous copying of the most quickly replicating virus. Natural selection acts on the entire quasispecies because mutation rates close to the error threshold mean that individual viral genomes are linked by a mutational coupling – all the possible mutational links between viral genomes are established – so that the whole population evolves as a single unit.

Much research has been undertaken to prove or disprove the theory that some RNA viruses conform to a quasispecies population structure. Although high levels of genetic variation have been demonstrated for many RNA viruses this is not sufficient to prove the existence of quasispecies. A quasispecies dynamic can only be proved by

demonstration that natural selection acts on viral populations as a unit. In support of the quasispecies theory, a high fitness Vesicular stomatitis virus variant was suppressed by one of lower fitness, which could have resulted from the latter having a high probability of being generated by mutation from closely related variants (de la Torre & Holland 1990). This observation, however, could be explained equally by genetic drift where the frequency of a viral variant is dependent upon its initial frequency within the population. More recent studies have investigated the possible quasispecies structure of poliovirus through an altered polymerase with increased fidelity. It was shown that increased fidelity of the virus polymerase can lead to reduced pathogenicity (Pfeiffer & Kirkegaard 2005), and also that high fidelity replication can alter the virus pathogenicity in preventing spread to the brain (Vignuzzi et al. 2006). However, again this does not necessarily prove the existence of quasispecies, only the effect that an error prone polymerase can have.

Many studies *in vitro* have determined virus population properties that are consistent with the quasispecies theory such as an error threshold and differences in mutation spectra affecting fitness. However, the relevance of the quasispecies dynamic, if it does exist, in nature is not known. The effective population size of FMDV is likely to be small due to bottleneck transmission events between animals and between farms, and perhaps also within an animal. A recent study on Poliovirus demonstrated the effect of bottleneck events on the virus population structure within a single host (Pfeiffer & Kirkegaard 2006). Therefore the term quasispecies to describe FMDV populations in nature should be used with caution until it has been categorically shown that natural selection acts on the virus population as a unit within a natural setting.

1.3.4 Studies on Foot-and-mouth disease virus evolution

The micro-evolution of FMDV has been extensively investigated in cell culture, largely focusing upon serotype C, revealing high levels of genetic variability (Arias et al. 2001;

Ruiz-Jarabo et al. 2004; Sobrino et al. 1983) and adaptability (Garcia-Arriaza et al. 2004). Viral populations have been shown to evolve to resist antibody induced selection (Borrego et al. 1993; Diez et al. 1989; Holguin et al. 1997; Sevilla et al. 1996), and to exhibit what has been referred to as genomic 'memory' (Arias et al. 2004; Ruiz-Jarabo et al. 2000). Genomic memory refers to the maintenance of a virus variant that was previously dominant, in low numbers within a population. They have also been shown to adapt to different multiplicities of infection (moi) (Sevilla et al. 1998) incurring the effects of Muller's ratchet at low moi through the irreversible accumulation of deleterious mutations (Escarmis et al. 1996), although the virus population can subsequently recover (Escarmis et al. 1999; Lazaro et al. 2003). The virus population has also been shown to maintain defective RNAs (Charpentier et al. 1996), and to persist and co-evolve alongside cell lines (de la Torre et al. 1988). All serotypes of FMDV have been shown to exhibit substantial levels of genetic diversity in the field, particularly within the nucleotide sequence encoding the capsid proteins, over broad temporal and spatial scales (Carrillo et al. 2005). This is typified by a continuous accumulation of silent changes over time and a more restricted set of amino-acid substitutions (Martinez et al. 1992).

1.3.5 Foot-and-mouth disease virus mutation rates and frequencies

The mutation rate of FMDV is an extremely important parameter because it impacts upon the speed and degree to which the virus can change to adapt to new environments and also to escape the immune system and protective effect of vaccines. It could also assist in determining the relevance of the quasispecies concept as discussed in section 1.3.3. The rate of change observed in the field has been measured over long time periods for different serotypes and found to lie within 0.0004 – 0.045 substitutions/nucleotide/year (Bastos et al. 2003; Haydon et al. 2001; Sobrino et al. 1986). This represents nucleotide substitutions that have become fixed within the consensus of the virus population and not the spontaneous mutation rate of FMDV. The rate of nucleotide change has also been studied *in vivo* by sequencing the VP1 coding region from persistently infected cattle to

give a rate of 0.913-7.41 x 10^{-2} substitutions/site/year (Gebauer et al. 1988). A study of persistent infection of buffalo with SAT 2 and SAT 1 found substitution rates 1.5 x 10^{-2} and 1.6 x 10^{-2} substitutions/nucleotide/year respectively (Vosloo et al. 1996). Another study found no change within the consensus nucleotide sequence of the VP1 region in acute infection in swine, although monoclonal antibody escape mutants were detected (Carrillo et al. 1990).

The actual spontaneous mutation rate of the FMDV RNA dependent polymerase is often reported as being very high (in the order of 1 x 10^{-3} to 1 x 10^{-5} misincorporations per nucleotide per genome replication). This is based upon a review of other RNA viruses in 1982 (Holland et al. 1982), which noted high mutation rates in Vesicular Stomatitis Virus and some negative strand viruses (Portner et al. 1980). Although the genetic heterogeneity that is introduced upon virus propagation in cell culture has been noted (Sobrino et al. 1983), the actual spontaneous mutation rate of FMDV has never been measured.

Many *in vitro* studies to look at mutation rates have assessed the within population mutation frequency of FMDV, of mainly serotype C but also a single study on serotype O (Gu et al. 2006), by molecular cloning and sequencing of PCR products from VP1 and occasionally the IRES or the 3D region. All have found a mutation frequency of mutants of the order of 10^{-4} per nucleotide sequenced (Airaksinen et al. 2003; Arias et al. 2001; Gu et al. 2006; Pariente et al. 2001; Sierra et al. 2000). However a mutation frequency does not directly translate into a mutation rate, as the frequency of mutants is dependent upon their initial frequency within the population and the forces of selection.

Poliovirus RNA polymerase mutation rate has been studied more thoroughly. An early estimate of the polymerase fidelity assessed by copying of homopolymeric RNA templates gave a mutation rate of 7 x 10^{-4} - 5.4 x 10^{-3} (Ward et al. 1988), however this

was unlikely to be an accurate representation of polymerase fidelity within a real cell and working upon its usual RNA template. A more recent *in vitro* study came to a similar conclusion of a mutation rate of 4.5 x 10^{-4} mutations per nucleotide copied (Wells et al. 2001). However the enzyme analysed was produced in *E.coli* and thus may not be completely representative of that produced in mammalian cells during normal virus replication, and again the enzyme was assessed in an artificial environment acting upon an RNA template that was not Poliovirus genome RNA. Another study assessing mutant frequency following a single replication cycle determined a mutation rate of 4.1 x 10^{-3} ± 0.6 x 10^{-3} (Ward & Flanegan 1992), which is very high and would result in a mutation being present in all virus progeny. Other studies of the Poliovirus RNA polymerase mutation rate have found much lower rates. For example, in assessing the rate of reversion of an amber mutant (requiring a single point mutation within the 3D polymerase coding region) a mutation rate of 2.5 x 10^{-6} was observed (Sedivy et al. 1987). In addition a study sequencing Poliovirus VP1 from 105 individual virus plaques found no mutations, suggesting a mutation rate of less than 2.1 x 10^{-6} mutations per nucleotide per infectious cycle (Parvin et al. 1986). A recent paper on the fidelity of the RNA polymerase of Poliovirus also suggests that previous results may be over estimates of the mutation rate as it found the polymerase fidelity to be in the range of 1.2 x 10^{4} to 1 x 10^{6} for transition mutations, and 3.2 x 10^{5} to 4.3 x 10^{7} for transversion mutations (Freistadt et al. 2007).

1.4 Molecular evolution of RNA viruses

1.4.1Analytical tools for the study of virus evolution

Currently the genetic evolution and relationships of viruses are studied by analysing their genetic sequence data by phylogenetic methods. Phylogenetic trees are constructed and used to deduce the genetic relatedness of the viruses. There are different methods for constructing phylogenetic trees; the first approach developed was Maximum

Parsimony methodology, but more recently Maximum Likelihood (Felsenstein 1981) and Bayesian methods (Huelsenbeck et al. 2001) are the preferred technique for tree construction. Maximum parsimony determines the most parsimonious tree requiring the least evolutionary steps. This method is simple, and as such makes very few assumptions about the evolutionary process. However certain features of genetic evolution of organisms present problems when using this method of tree construction. Firstly inaccuracies can occur as a result of the existence of homoplasy. Homoplasy describes processes such as convergent evolution where a single mutation can occur twice on independent branches of a tree, and hence does not necessarily determine that two sequences sharing a mutation were derived from a common ancestor which also contained this mutation. Another hurdle to overcome is back mutation, where a mutation reverts to its original genotype. This would cause the specific sequence to look more ancestral than is necessarily the case. A further draw back to the method of Maximum Parsimony is that it takes no account of the rate at which mutations arise and the varying probabilities of different mutations occurring (i.e. transversions vs transitions). For these reasons the parametric method of Maximum Likelihood provides an improved alternative, because the most probable tree that suits a specific evolutionary model is determined. A parametric method is potentially much more powerful providing that the model employed is a reasonable approximation of the evolutionary processes that gave rise to the observed genetic data. The evolutionary model may include a large number of parameters accounting for differences in the probabilities of various character states, differences in the occurrence of particular substitutions/mutations, and differences in the probabilities of change among characters. With the sophisticated models now developed (eg the Hasegawa-Kishino-Yano (HKY) model (Hasegawa et al. 1985) and the General Time Reversible (GTR) model (Lanave et al. 1984)) an improved idea of phylogeny is achieved, although fitting an incorrect model can give incorrect results. The suitability of models can be tested using a program such as ModelTest (Posada & Crandall 1998). Maximum Likelihood estimation of tree phylogeny is generally preferable to Maximum

Parsimony because it is statistically consistent with a better statistical foundation and it allows complex modeling of evolutionary processes. To infer statistical confidence in either Maximum Parsimony or Maximum Likelihood constructed phylogenies bootstrap analyses are performed, which can be time consuming. A further method to infer phylogenies is that of Bayesian inference which generates a posterior distribution for a parameter, based on the prior for that parameter and the likelihood of the data, (represented by the sequence alignment). In other words where as Maximum Likelihood analysis investigates the probability of the observed data given a specific evolutionary model, Bayesian inference looks at the probability that a model is correct given the observed dataset. With the availability of MCMC (Markov chain Monte Carlo) methods (Yang & Rannala 1997) Bayesian inference can be a preferred choice for tree estimation because it can be faster than Maximum Likelihood and no bootstrapping is required as the posterior probabilities determine the statistical confidence in the tree.

Although in the majority of incidences Maximum likelihood or Bayesian inference is preferable for tree construction, in certain situations Maximum Parsimony can be a viable alternative. When studying closely related sequences over a short time period the likelihood of back mutation is relatively low and hence Maximum Parsimony tree construction is likely to give an accurate estimation of tree phylogeny. In this thesis the phylogenetic analysis is often performed with the aim of tracing specific virus history and in these cases the method of statistical parsimony was used. The distances depicted by parsimony trees represent the actual number of differences between sequences, whereas for a Maximum Likelihood tree the probability of change is shown. Within this thesis closely related sequences were investigated, with a focus on the accumulation of changes, and in this case a simpler representation of the raw data as depicted by Parsimony is desirable. The TCS statistical parsimony program (Clement et al. 2000) used within this thesis can position sequences internally on a branch which assists in depicting directly ancestral sequences. Although the statistical parsimony trees drawn

have not been bootstrapped, the fact that the data comprises the complete genome sequences of the sampled viruses indicates that the tree is as accurate and as representative as it can be, and is not sensitive to the choice of a single arbitrary locus. There is no further genetic data retrievable.

Bayesian and MCMC techniques have enabled more computationally demanding analyses of data, the software program BEAST (Drummond & Rambaut 2007) harnesses these techniques and allows the relatively straightforward implementation of evolutionary models that contain parameters relating to viral demography. By implementing a relaxed molecular clock model (Drummond et al. 2006) BEAST explicitly models the rate of molecular evolution on each branch of a tree, when given time stamped data to analyse. This enabled the informed determination of the rate of evolution of virus genetic sequences obtained during this PhD project, by incorporating sophisticated substitution, site rate, and branch rate models.

1.4.2 Tracing

The movement of FMDV worldwide can be successfully traced using genetic sequence data. Spread of FMDV was successfully traced through Asia, Africa and into Europe (Knowles & Samuel 2003). Using genetic data a possible introduction of virus into Africa from Saudi Arabia was identified (Bronsvoort et al. 2004), additionally, the close relationship between the virus in Taiwan in 1997 and isolates from Hong Kong and Russia was identified through genetic analysis (Tsai et al. 2000). However, this is broad scale tracing between countries and for other viruses higher resolution tracing has been demonstrated (Table 1.1). For example, it has been possible to distinguish between rhinovirus epidemics, seasons, and co-circulating strains using genetic data based on a short 420nt stretch (Savolainen et al. 2002). An investigation of the Classical swine fever outbreak in the Netherlands in 1997-1998 showed the existence of two sub-populations within one outbreak (Widjojoatmodjo et al. 1999), and previous analysis of outbreaks in Italy enabled the identification of separate introductions of virus (Lowings et al. 1994).

Tracing the origin and evolution of avian influenza viruses among bird populations is also common (Banks et al. 2001). Also, forensic genetics has been important in determining Norovirus transmission between and within hospitals (Lopman et al. 2006). Direct human-to-human transmission has also been traced using nucleotide sequence data for Hepatitis C virus (HCV), human immunodeficiency virus (HIV), and SARS corona virus infections.

FMDV is an RNA virus that is reported to have a high mutation rate in accord with many of the viruses for which transmission tracing has been successful. For this reason it is highly likely that sequence data could be used to trace the virus movement between hosts. However, it is unknown to what extent this will be possible as selection events may purge many changes from the consensus virus, and also bottleneck events may confuse interpretation of the results. A previous study investigated tracing the virus within an epidemic using VP1 sequences (Christensen et al. 2005) and found that there was not enough variation to distinguish between farms. However VP1 represents less than 8% of the genome, and thus increased amounts of sequence data might improve the resolution at which the virus can be traced.

Table 1.1 Examples of sequence based transmission tracing of RNA viruses

Virus	Nucleic acid type	Disease infectious period	Fragment Length (nt)	% of genome (genome size in nt)	Sequence origin	Tracing Resolution	Reference
Ebola	- ss RNA	Short	2049	10.80% (19000)	genbank	Followed spread from village to village	(Walsh et al. 2005)
HIV	+ ss RNA	Long	966	10.50% (9200)	Blood	Human to human	(Albert et al. 1994)
	+ ss RNA	Long	1130	12.30% (9200)	Blood	Human to human	(Salzberger et al. 2000)
SARS Coronavirus	+ ss RNA	Short	29800	100% (29800)	Primary tissue and cell culture	Human to human	(Vega et al. 2004)
	+ ss RNA	Short	29800	100% (29800)	genbank	Human to human	(Zhao et al. 2004)
Norovirus	+ ss RNA	Short	123	1.64% (7500)	Faecal samples	Linked human outbreak with oysters	(Dowell et al. 1995)
	+ ss RNA	Short	280 + 80	4.80% (7500)	Faecal samples	Distinguished transmission between and within hospitals	(Lopman et al. 2006)
Hepatitis A	+ ss RNA	Short	-	-	Serum	Linked an outbreak in homosexuals with one in a family	(Stene-Johansen et al. 2002)
Hepatitis C	+ ss RNA	Long	472	4.90% (9600)	Serum	Distinguished between 6 patients	(Gonzalez-Candelas et al. 2003)
	+ ss RNA	Long	188	1.90% (9600)	Bone marrow smear sample	Determined source of outbreak and patient-to-patient transmission chain	(Spada et al. 2004)
	+ ss RNA	Long	337 + 472	8.43% (9600)	Serum	Determined transmission events between patients	(Bracho et al. 2005)
Rhinovirus	+ ss RNA	Short	420	5.25% (8000)	Cell culture 1 or 2 passages	Distinguished between epidemics, seasons and co-circulating strains	(Savolainen et al. 2002)
Classical Swine Fever virus	+ ss RNA	Short	321 + 850	9.37% (12500)	Epithelium and cell culture	Detection of 2 sub-populations	(Widjojoatmodjo et al. 1999)

1.5 Objectives of PhD thesis

The procedure by which mutations are fixed into the consensus of FMDV populations is highly complicated stemming from an intricate process of mutation and selection. However, this overall process can be broken down into defined events to assist our understanding of FMDV evolution. These events include initial spontaneous mutation rate of the virus at cell level, the maintenance and characteristics of a virus mutant spectrum, the pressure of immune selection, the effect of transmission within a host, the effect of transmission between hosts and populations of hosts, and the accumulation of genetic mutations over time in an epidemic. The objective of this PhD thesis is to characterise further the evolution of FMDV both in the field and *in vitro* by investigating nucleotide variation, with a broader aim of applying the data in an epidemiological context.

Chapter 2 describes initial development of a method to obtain complete consensus genome sequence from original epithelium. The objective was to design a method that was robust, reliable and simple such that it could be used to examine a large number of virus samples. In Chapter 3 this method was used to analyse the extent of nucleotide variation that was apparent within an epidemic, by studying infected epithelium samples from the UK 2001 FMDV epidemic. Genetic evolution was studied across the entire genome investigating the type and frequency of changes that occurred. Also investigated were the relationships of virus genetic evolution during the epidemic to time and geography of virus spread and transmission. This included assessing whether Pan Asia O FMDV evolved during the epidemic in a manner such that mutations were inherited along a transmission chain in a regular fashion. The potential of using the genetic data for epidemiological tracing studies, including dating the start of the epidemic was also considered.

Following the new observation that the virus evolution during the outbreak conformed to a molecular clock with nucleotide changes being inherited along an infection transmission tree (Cottam et al. 2006) the potential of using complete genome sequence data for transmission tracing was assessed in Chapter 4. Epidemiological data concerning infection profiles of individual farms were also incorporated improving the tracing potential of the genetic data by allowing statistical evaluation of the likely transmission trees estimated.

With the objective of increasing our understanding of the virus evolution, and the fixation of mutations into the consensus sequence, the mutant frequency within virus populations from naturally infected animals was assessed in Chapter 5. This type of study can provide information regarding the type and distribution of mutations within a virus population and is imperative to understanding how nucleotide changes become fixed into the consensus sequence of a virus population. Chapter 6 describes a study to gain an understanding of the spontaneous mutation rate of the virus, a process underpinning the nucleotide variation seen in the field.

The findings of this thesis are concluded and discussed in Chapter 7. Chapter 8 reports on the recent use of complete genome sequencing in real-time during the UK 2007 outbreaks of O_1BFS FMDV.

Chapter 2

Development and optimisation of protocols for complete genome amplification and sequencing of Foot-and-mouth disease virus: Pan Asia O serotype

The work in this chapter is expected to be published;

Cottam, E.M., King D.P., edited by Caugant, D.A. Molecular Epidemiology of Microorganisms: Viral genome sequencing, Chapter 21, *Methods in Molecular Biology, Human Press*

I would like to thank the Wellcome Trust Sanger Institute for their collaboration and contribution to the pilot study detailed within this chapter.

2.1 Summary

A method was developed for complete genome sequencing of foot-and-mouth disease virus from the UK 2001 epidemic from the archive of epithelium samples held at the World Reference Laboratory, Institute for Animal Health, Pirbright. This method reliably produces sequence data spanning the entire genome of the virus, directly from epithelium samples that have been stored in 50% glycerol for 5 years or more. In addition a successful pilot study was undertaken in collaboration with the Wellcome Trust Sanger Institute to sequence large proportions of the genome from a single RT-PCR product. A RT-PCR that amplifies the whole poly-protein coding region, and also one that represents 80% of the genome have been designed and used to amplify virus RNA extracted from epithelium samples. The sequencing strategy designed for use at the IAH produces identical results to the sequencing method used at the Wellcome Trust Sanger Institute, supported by the independent analyses of four virus infected epithelium samples.

2.2 Introduction

2.2.1 Overview

The overall objective of this project was to develop and optimise a method for sequencing a significant number of complete genomes of the Pan Asia O serotype of FMDV from the UK 2001 outbreak held within the library of epithelium samples at the World Reference Laboratory (WRL) Pirbright. Epithelium samples were submitted to the WRL, IAH Pirbright from 1224 FMDV positive infected premises, and there was epithelium remaining from 92% of these. However, epithelium was not stored from infected premises that were designated as laboratory negative for FMDV, or from farms whose animals were culled as dangerous contacts and not subject to definitive laboratory diagnosis.

The FMDV genome is a highly structured positive sense single stranded RNA of approximately 8500 nucleotides. Within the 5' UTR is a long stretch of cytosine residues (poly 'C' tract) which prevent direct PCR through this region. At the 3' end of the genome is a long poly adenosine tail (poly 'A' tract). The genome structure is described in more detail in Chapter 1; section 1.1.5.

2.2.2 Literature review of methods to sequence the complete genome

Since 1999 there have been many projects to sequence the complete genome of FMDV, the majority on a small scale of 1 or 2 genomes (Toja et al. 1999) and more recently in 2005 a study to sequence 103 genomes across all serotypes (Carrillo et al. 2005). Only one study (Oem et al. 2004) did not amplify the virus by *in vitro* propagation prior to sequencing. Some studies incorporated a separate reverse transcription step prior to PCR (Carrillo et al. 2005; Mason et al. 2003b; Nobiron et al. 2005), and some carried out a one step RT-PCR (Kanno et al. 2002; Oem et al. 2004). Carrillo *et al* 2005 amplified

the full length cDNA non-specifically prior to PCR to increase its concentration. All studies amplified the genome in overlapping PCR fragments ranging from 5 fragments (Nobiron et al. 2005) to as many as 21 fragments (Carrillo et al. 2005). Nobiron *et al* 2005 cloned the fragments before sequencing, whereas the other studies sequenced the PCR products directly. When the PCR products were cloned prior to sequencing, specific PCRs were later performed on the original material to confirm any nucleotide sequence changes found.

Sequence from PCR products contain the predetermined primer sequence at either end, and in some studies PCR RACE (Rapid amplification of cDNA ends) has been performed in order to obtain the viral sequence for the extreme ends of the genome (Kanno et al. 2002; Nobiron et al. 2005). In other studies, it was argued that if an oligo dT primer was used to amplify the 3' end, only the 5' ends of the s- and l-fragments would be predetermined and these sites are known to be highly conserved across serotypes (Mason et al. 2003b). Studies on other RNA viruses such as Dengue (Barrero & Mistchenko 2004) and Ebola (Sanchez & Rollin 2005) have used similar methods including overlapping PCR reactions, cloning and sequencing.

2.2.3 Study objective

The aim of this study was to develop a method to analyse the complete genome of foot-and-mouth disease virus (FMDV) with high sensitivity from raw epithelial clinical samples. Nucleotide sequencing of virus direct from epithelium samples would increase confidence that any nucleotide changes seen were present within the infected animal and were not a consequence of virus propagation in cell culture. However, the quantity and quality of viral RNA recovered from stored epithelium is likely to be lower or less consistent than that following propagation in cell culture, which may hinder attempts to amplify the genome by PCR.

The final method developed needed to be sensitive in order that virus could be sequenced directly from the epithelium without propagation. It also needed to be accurate, involving the least amount of *in vitro* amplification possible to minimise the introduction of copying errors, particularly where the starting concentration was low as would be likely for some samples. A method with high redundancy is also desirable in order to provide high confidence in the results. Finally, if a large number of FMDV genomes were to be sequenced the assay must be simple and robust.

It was attempted to amplify the genome in multiple overlapping PCR fragments, and also in two fragments (one of the short s-fragment and one long PCR fragment of the remainder of the genome from the poly 'C' tract to the poly 'A' tract). The amplification of the majority of the genome in one fragment provided an opportunity to collaborate externally with the Wellcome Trust Sanger Institute to explore an alternative sequencing strategy. However, the bio-security regulations at IAH require that only 80% of the FMDV genome (6560nt) may be transported off site as a PCR product at any one time, and thus amplification of a fragment of this size was prioritised.

2.3 Development of methods

2.3.1 General methods used

2.3.1.a Original sample suspension and RNA extraction

Approximately 1.5g of vesicular epithelium (originally stored in 0.04 M25 phosphate buffer [disodium hydrogen phosphate, potassium dihydrogen phosphate, pH 7.5] and 50% [vol/vol] glycerol) was ground by using a pestle and mortar in a class II safety cabinet and resuspended to a 10% suspension with 0.04 M phosphate buffer. The solution was then centrifuged for 10 minutes at 3,500*g* at room temperature, and the

supernatant was removed and stored at -70°C until testing. RNA extraction was performed using TRIzol reagent (Invitrogen) as described previously (Reid et al. 1998).

2.3.1.b Reverse transcription

Reverse transcription reactions were set up according to manufacturers instructions unless otherwise stated. Briefly, RNA, Primer, and dNTPs were incubated at 70^{o}C for 3 minutes, followed by incubation on ice for 3 minutes. Fresh reverse transcriptase mix including 10x RT buffer, 25mM $MgCl_2$, 0.1M DTT, RNase OUT and Superscript III reverse transcriptase (Invitrogen) were then added to the reaction, which was then incubated at 42^{o}C for 4 hours.

2.3.1.c PCR

The PCR reactions were set up (unless otherwise stated) using Platinum High Fidelity Taq (Invitrogen) following the manufacturer's recommendations per reaction. Each comprised 5µl 10x Hi-Fi buffer, 2µl 50mM $MgSO_4$, 1µl 10mM dNTPs, 1µl 10mM forward primer, 1µl 10mM reverse primer, 0.2µl Taq (5 units/µl), plus nuclease free water and cDNA to make up a reaction volume of 50µl. The PCR reactions were subject to a basic amplification procedure in 0.2ml tubes on thermal cyclers (MJ Research PTC-100). The results of PCR reactions were visualised using a UV Camera after running 2-10µl of PCR product on a 1.2% Agarose gel with 0.002% Ethidium bromide at 90v for 40 minutes. The quantitative DNA ladder HyperLadder 1 (Bioline) was run alongside the products for comparative quantification of product size and quantity.

2.3.1.d Sequencing reactions

The sequencing reactions were set up using the master mix supplied with the Beckman DTCS Quick start sequencing kit following the manufacturer's protocol. Briefly 8µl of Beckman Master-mix was added to 8µl of DNA of a suitable concentration (dependant upon the PCR product size) and 4µl of primer. 4pmol of primer was used per reaction as opposed to the 3.2pmol suggested. The reactions were run on a programme of 30

cycles of 96°C for 20 seconds, 50°C for 20 seconds, and 60°C for 4 minutes, followed by holding at 4°C.

2.3.2 Strategy for complete genome amplification with overlapping PCR products

2.3.2.a From epithelium to DNA

The strategy designed in this study for the amplification of the complete FMDV genome was similar to those used previously; including virus suspension from original material, RNA extraction, reverse transcription, PCR amplification and degenerate sequencing. The genome was amplified in a small number of overlapping PCR fragments <3kb. Finally the sequencing strategy was constrained to use no more than one 96 well sequencing plate for each FMDV genome.

Following sample suspension preparation from epithelium, the RNA was extracted manually using TRIzol reagent as the RNA yield from manual extraction is higher than robotic extraction (S Reid, A Shaw personal communication). Also to avoid cross-contamination, each sample was processed individually and consequently the number of samples for RNA extraction was not large enough at any one time to warrant using the robotic system for RNA extraction.

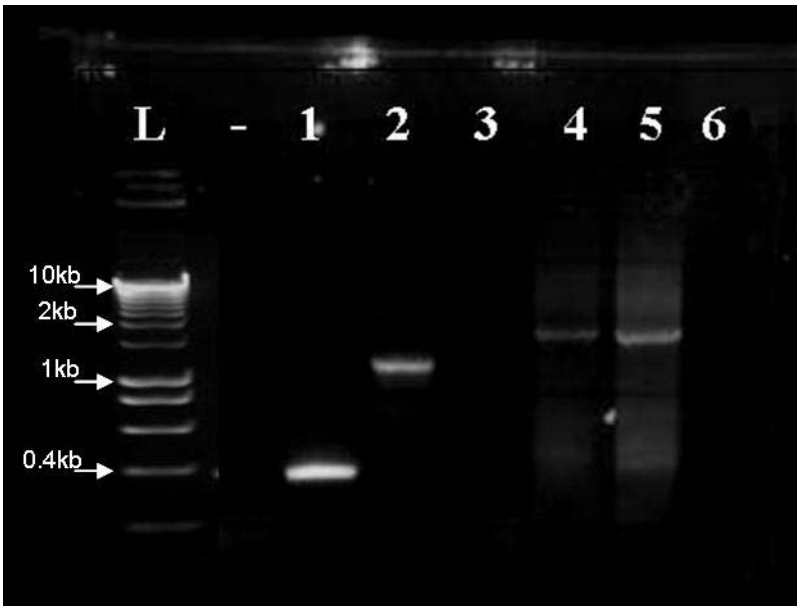


Figure 2.1

The gel depicts the products obtained from initial PCR amplification of the FMDV genome using the six primer sets (1-6) outlined in Table 2.1. Only sets 1, 2, 4, and 5 gave a visible product.

In order to amplify large size PCR fragments, it is essential that the cDNA is of good quality, especially if the RNA has been stored for a long period of time as its quality may have declined. Superscript III reverse transcriptase (Invitrogen) was used to reverse transcribe the RNA. It was decided to use an oligo dT primer in the RT reaction to improve specificity and concentration of full length cDNA. However, during replication of FMDV a negative sense copy of the genome is formed and so random primers which could bind to the negative strand as well as the positive strand were included to potentially improve cDNA concentration. The RT enzyme, Superscript III, had the ability to transcribe through the poly 'C' tract that is present within the FMDV genome. This characteristic was shown in initial experiments by PCR amplification of the S fragment (upstream of the poly 'C' tract) from RT products made using only the oligo dT primer (data not shown). The reverse transcription was run for 4 hours to improve the quality of the cDNA for the purpose of long PCR (personal communication with V Fowler). The

cDNA was then cleaned up prior to PCR, removing the oligo dT primer to prevent interference with any subsequent PCR reaction.

2.3.2.b Amplification of complete genome in overlapping PCR products

Initially, it was decided to amplify the genome in six overlapping fragments. The published genetic sequence for FMDV UKG 35/2001 (AJ539141) and other Pan Asia O sequences were aligned (AJ539136-AJ539140) using ClustalW implemented in Bioedit. This alignment was then used to design suitable primers (Table 2.1). For initial testing, the following thermal cycling conditions were used; 94^{o}C for 5 minutes, then 39 cycles of 94^{o}C for 1 minute, 62^{o}C for 1 minute, 72^{o}C for 2 minutes 42 seconds, followed by 72^{o}C for 7 minutes. Only 4 of the 6 primer sets worked (Figure 2.1).

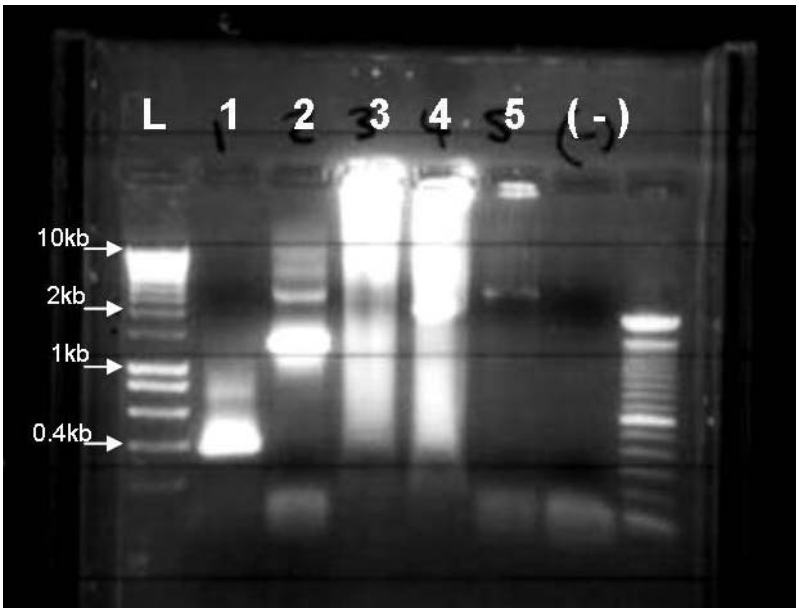


Figure 2.2

Initial PCR amplification of FMDV genome with 5 primer sets. Initial amplification of the final PCR products 1-5 from primer sets UKFMD/For and Rev 1-7 prior to optimization. Non specific products were visible in lanes 1 to 4.

It was possible to combine the primer sets 5 and 6 to create one long PCR product, and a new primer was designed to pair with UKFMD/For3 to amplify the capsid region (UKFMD/Rev7 Table 2.1). These five primer sets worked (Figure 2.2) but showed non

specific priming under the PCR conditions used (94°C 4 minutes, 40 cycles of 94 °C 1 minute, 65 °C 1 minute, 72 °C 3 minutes, followed by 72 °C 7 minutes).

2.3.2.c Optimisation of PCR reactions

To optimise the PCR for fragments 3 and 4 the annealing temperature was increased to 68°C and it was determined that the Mg^{++} concentration used was correct.

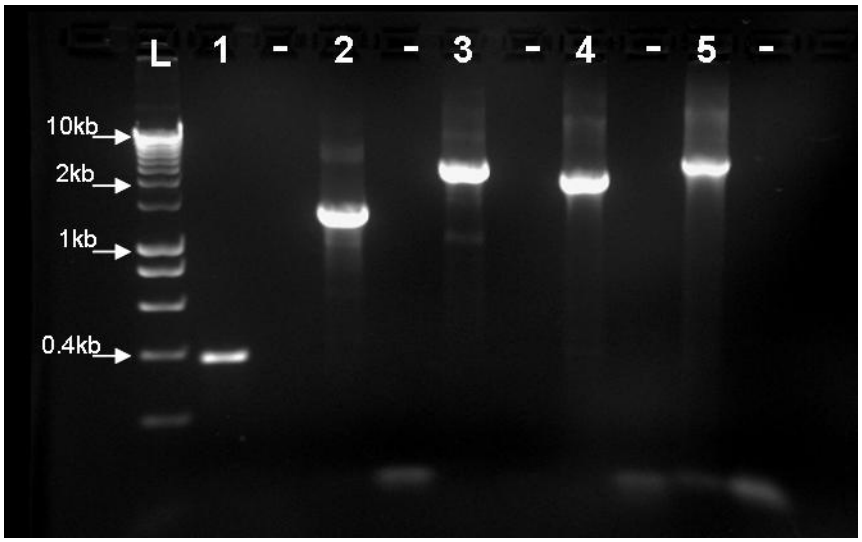


Figure 2.3

Overlapping PCR products spanning the entire FMDV genome. Optimised amplification of five overlapping PCR products spanning the complete genome of FMDV from epithelium sample UKG 126/2001.

However, in order to improve the yield of fragment 5, the amplification programme was modified lowering the annealing temperature to 60°C. When PCR reactions with primers UK/FMD Forward and Reverse 1-4 were run on the higher annealing temperature programme and fragment 5 (UK/FMD Forward 5 and Reverse 6) was run on the lower annealing temperature good clean PCR products were obtained (Figure 2.3).

The PCR products were cleaned up directly without running through an agarose gel and this improved the success of subsequent sequencing reactions. The primer sets amplify overlapping fragments the location and size of which are illustrated in Figure 2.4.

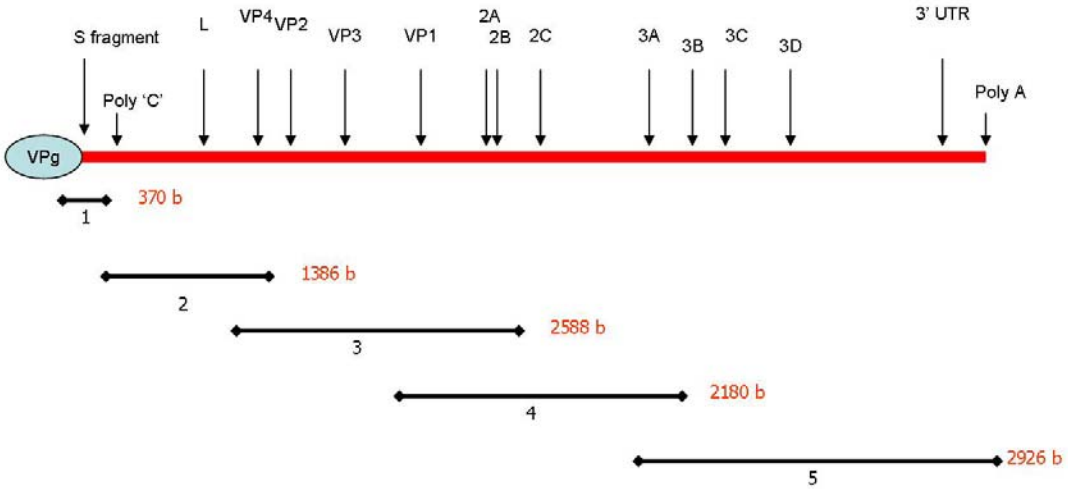


Figure 2.4

Schematic showing size and position of 5 overlapping PCR products spanning the complete genome of FMDV. Vertical arrows indicate the start of specific UTR features and coding regions.

2.3.3 Amplification of complete genome in 2 PCR products

The PCR reaction to amplify the s fragment was optimised as described in section *2.3.2*. The remainder of the genome is ~7840 nt in length, the poly protein is encoded by a stretch of 6996 nt, and if the final product is to be sent off site it must be no more than 6560 nt in length. It was decided to attempt to amplify as large a proportion as possible that incorporated the entire poly protein, and also design primers that would amplify a product no more than 6560 nt in length (i.e. 80% of the FMDV genome).

2.3.3.a Initial attempts to amplify large fragment by PCR

Many different combinations of primer sets were tried, however the only result obtained was a high molecular weight smear. Different variables were investigated to try and

resolve this problem including Taq concentration (Figure 2.5a), and annealing temperature and Mg^{++} concentration (Figure 2.5b), with no success.

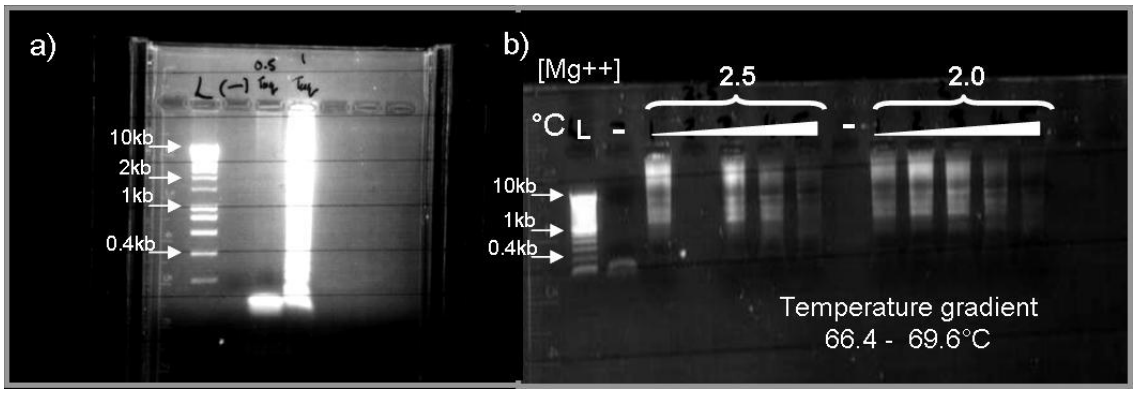


Figure 2.5

Attempt to optimise long PCR. a) Agarose gel depicting the PCR products from PCR reactions differing in Taq concentration from 0.5µl to 1µl. b) PCR reactions with 2.5µl and 2.0µl $MgSO_4$ with temperatures at 0.64°C increments from 66.4 to 69.6°C

A literature search suggested that a touch down or the reverse PCR thermal cycling programme could improve yields of long PCR products (Don et al. 1991). This was investigated as shown in Figure 2.6 with no success. It was possible that the intricate RNA secondary structure was causing the problem and so the effect of adding DMSO to the PCR reaction was investigated. Again no improvement was seen, Figure 2.7.

Ultimately, a clean PCR product of a correct size using a normal PCR thermal cycling protocol was obtained. This suggested that the success of the long PCR was determined by the quality of the cDNA sample, as this was the only variable. Using this successful sample the candidate primer sets were tested (Figure 2.8), determining that it was possible to amplify a product as large as 7258 nt (88.5% of FMDV genome using primer set Forec6 and Revec43, Figure 2.8f), and also identifying it as the best candidate fragment for sending to Wellcome Trust Sanger Institute (6526 nt, primer set Forec6 and UKFMD/Rev5).

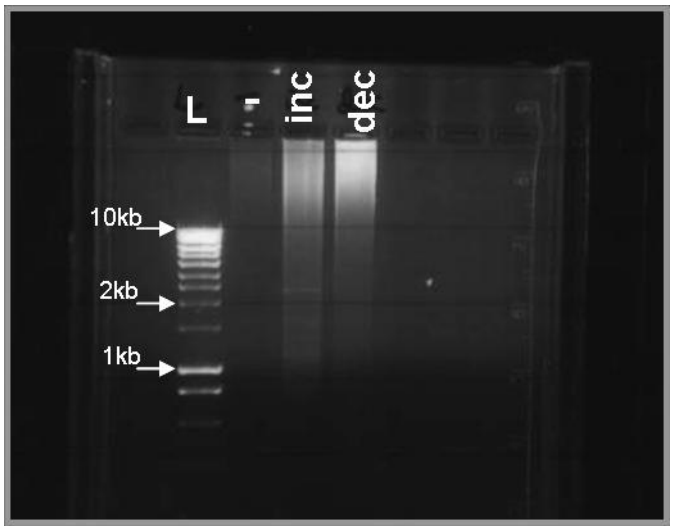


Figure 2.6

Investigation of touch down PCR. Agarose gel depicting PCR products from PCR reactions run on dec (lane 3); touch down thermal cycling conditions (9 cycles annealing at 70°C, 14 cycles annealing at 68°C, and 19 cycles annealing at 66°C) and inc (lane 4); the reverse (9 cycles at 66°C, 14 cycles at 68°C, and 19 cycles at 70°C). The negative control is shown in lane 2.

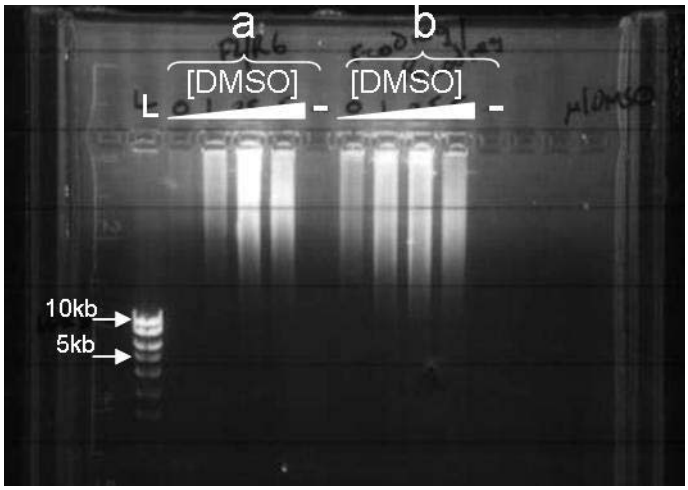


Figure 2.7

Investigation of the effect of DMSO. Agarose gel depicting the effect of adding DMSO to the PCR reaction using primer sets a) UKFMD/For 2 and UKFMD/Rev 6, b) Fcodreg and Rcodreg.

2.3.3.b Final amplification of 6526nt fragment for 5 samples and transportation to the Sanger Institute

To improve the RNA quality of the virus samples the RNA extraction was repeated and the pellet was re-suspended in only 12µl of nuclease free water, and all of this was used for the subsequent reverse transcription and PCR reactions. This was successful as shown in Figure 2.9, and these 5 samples were sent to the Sanger Institute in ethanol after phenol/chloroform extraction (according to the IAH bio security regulations), for precipitation at the Sanger Institute.

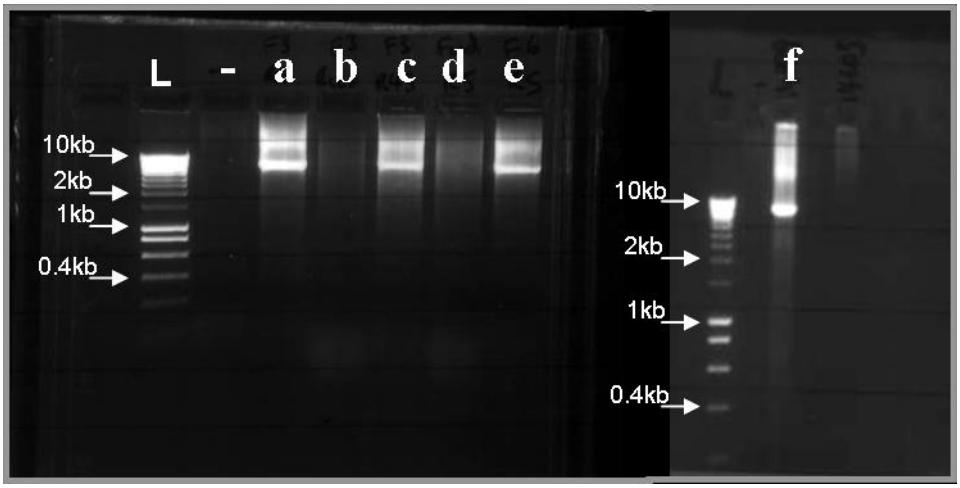


Figure 2.8

Determination of successful primer sets for long PCR. Agarose gel depicting PCR products produced from different primer sets (see Table 2.1 and 2.2); a) UKFMD/For 3 and UKFMD/Rev 5, b) UKFMD/For 3 and Rcodreg, c) UKFMD/For 3 and Revec 43, d) Fcodreg and UKFMD/Rev 5, e) Forec6 and UKFMD/Rev 5, and f) Forec6 and Revec43.

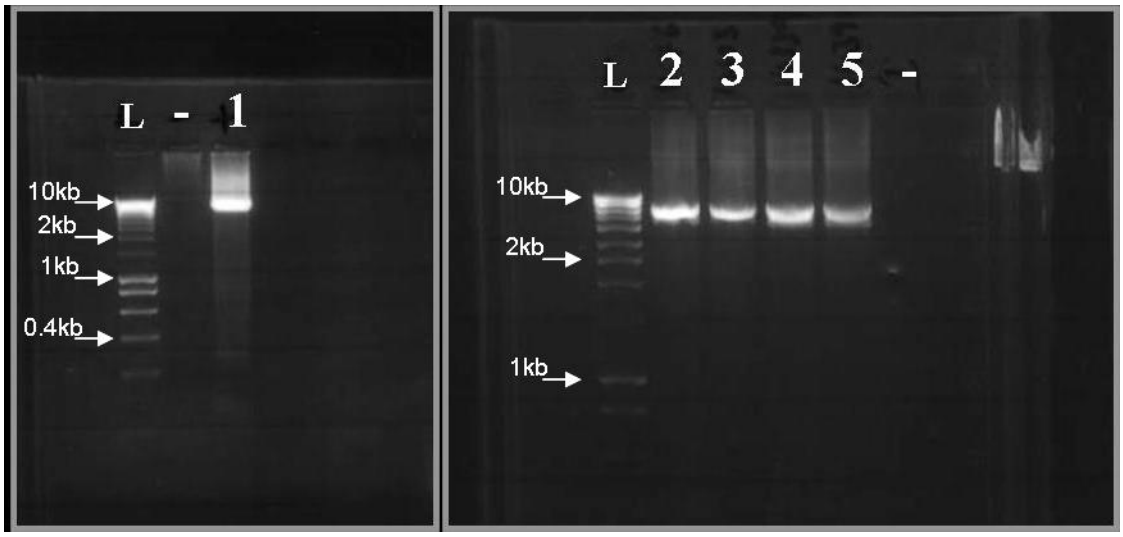


Figure 2.9

Amplification of FMDV genome for transportation to the Sanger Institute. Agarose gels depicting PCR products produced from primers Forec 6 and UKFMD/Rev 5 (6526b) from samples; 1) UKG 127/2001, 2) UKG 11676/2001, 3) UKG 14603/2001, 4) UKG 14339/2001, 5) UKG 14391/2001.

2.3.4 Sequencing of PCR products

To achieve maximum redundancy, allowing for the occasional failure of a sequencing reaction without the need to repeat the whole process, 84 sequencing primers were designed (42 forward and 42 reverse) approximately 200nt apart and offset by approximately 100nt resulting in a primer every 100nt (Table 2.2). A liquid handling robot (Qiagen Biorobot 3000) was used to set up the sequencing reactions in order to reduce operator error and also to make the method suitable for high throughput. The robot programme set up the reactions adding the master-mix first, followed by the DNA, then the primer, using only one box of tips. This method for sequencing the complete genome results in an average of 5x coverage of each nucleotide. The final method used for the remainder of the project including details concerning reagent volumes and concentrations is detailed in the methods section 3.3, Chapter 3, and the protocol is included as Appendix 1.

2.3.5 Assessing confidence of sequence data

2.3.5.a Reproducibility of sequencing via overlapping PCR products

This complete method was tested for reproducibility by a repeat of the whole process from a single piece of epithelium. The epithelium for virus UKG 1961/2001 was cut in half and each half was sequenced independently.

2.3.5.b Comparison of sequence data generated by two methods

Each sequenced determined by the Sanger Institute was directly compared to that of the same isolate generated at the IAH.

Table 2.1 Primer sequences for PCR

NAME	*SEQUENCE (5'-3')*
UKFMD/FOR1	TTG AAA GGG GGC GTT AGG GTC TCA
UKFMD/REV1	GGG TGA AAG GTG GGC TTY GGG T
UKFMD/FOR2	CCC AAG TTT TTA CCG CCT TTC CCG
UKFMD/REV2	GTT GAT AAT GCT TCC AGT GTT GCC TG
UKFMD/FOR3	CCA CGC TGG CAT CTT CCT GAA AG
UKFMD/REV3	GAG CTT GTA CCA GGG TTT GGC CTC A
UKFMD/FOR4	GTG TTG GAC CTG ATG CAA ACC CC
UKFMD/REV4	GTC TCT TGC GAG TCT CGC GGA TC
UKFMD/FOR5	TTC AAG CCT CAA CCG CCC CTC
UKFMD/REV5	CCT CAA ACA TGA TGT TCA TTG CGT CAC T
UKFMD/FOR6	GAG AGT GGT GCT CCC CCG ACT G
UKFMD/REV6	GGC GGC CGC TTT TTT TTT TTT TTT
UKFMD/REV7	CCA GTG GCC AGT TCC TCA AAT G
FCODREG	CTT CTA CGC CTG AAT AGG TGA C
RCODREG	CCT TTC GGA CTT TTC ACT CCT AC

Table 2.2 Primer sequences for sequencing and PCR

Name	*Direction*	*5'nt*	*Sequence (5' to 3')*
Forec1	Forward	1	TTGAAAGGGGGCGTTAGGGTC
Forec 2	Forward	186	CGTGACTGGTTAATACTCTTACC
Forec 3	Forward	377	CCAAGTTTTTACCGCCTTTCCC
Forec 4	Forward	600	GCTTGAGGAGGACTTGTACAAAC
Forec 5	Forward	757	AGCACTGTTGCTTCGTAGCGG
Forec 6	Forward	926	CACTGGTGACAGGCTAAGGATG
Forec 7	Forward	1147	CACTGTTCTTATCACGAGCAC
Forec 8	Forward	1332	GATGCTATCAAACAATTGGAAG
Forec 9	Forward	1584	GAGGACTTTTACCCCTGGACG
Forec 10	Forward	1763	CAATTACTACATGCAGCAGTACCAG
Forec 11	Forward	1959	ACCGAGGAGACCACTCTTCT
Forec 12	Forward	2129	GTTCTTCAAAACCCACTTGTTC
Forec 13	Forward	2329	CAGAACTTTGCTCTATTGACAAG
Forec 14	Forward	2552	CAACATCGCCCCTACCAACG
Forec 15	Forward	2723	GTTCACCAACTTCCTTGATG
Forec 16	Forward	2912	CACCATCAACCTGCACTTCATG
Forec 17	Forward	3128	GGAGACCACAAATGTACAGG
Forec 18	Forward	3306	GTTGAGAACTACGGTGGTGAG
Forec 19	Forward	3489	CTAGAAGTGGCAGTGAAACAC
Forec 20	Forward	3693	GTGAGAGGTGACCTGCAAG
Forec 21	Forward	3914	GCTCAAGTTGGCAGGAGACG
Forec 22	Forward	4081	CCACTGGAGTGAAGGCTATCAG
Forec 23	Forward	4265	GAAGATCTCCGACTCGCTCTCC
Forec 24	Forward	4444	CCATTCTCAAGAACGGCGAGTG
Forec 25	Forward	4640	GTGTTTGAAGAGCGGGAACATC
Forec 26	Forward	4838	ACCTGACCCTGACCACTTCGAC
Forec 27	Forward	5030	CAACCTGTACTCGGGCTTCAC
Forec 28	Forward	5203	ACGACTGTGCCCTTCTCAACG
Forec 29	Forward	5441	CTTTGAGGGGATGGTGCATGAC
Forec 30	Forward	5606	CAAGAGACAGCAGATGGTGGATG
Forec 31	Forward	5805	GAAGAACAACCACAAGCTGAAGG
Forec 32	Forward	5987	GAAGAAACCTGTCGCTTTGAAAG
Forec 33	Forward	6195	GACAAGATCATGTTGGACGG
Forec 34	Forward	6400	ACGCTGATGTTGGGAGACTG
Forec 35	Forward	6562	CTTTCATCGTCGGCACTCAC
Forec 36	Forward	6759	GGTGTGTTTAACCCCGAATTTGG
Forec 37	Forward	6931	TGCATAGCGTGCTGGGTACG
Forec 38	Forward	7160	CTTCCTGAAGGACGAGATTCGC
Forec 39	Forward	7351	GATTTGGCACGCATTTTGCTC
Forec 40	Forward	7574	TTGTTCCGCAACAAGCATCATC
Forec 41	Forward	7773	CCAGCTGACAAAAGCGACAAAG
Forec 42	Forward	7977	TCTGGACCTGACGAGTACCG
Revec	Reverse	8183	GAAATAGGAAGCGGGAAAAAGCC
Revec	Reverse	7897	GGTCTTCGAAGCCATCACAG
Revec	Reverse	7707	CTTGCAACCACGATGTCGTC
Revec	Reverse	7503	TCAGGATCCACTCAGCGTTTG
Revec	Reverse	7279	CACAAAATCTGCCAATCATCATC
Revec	Reverse	7064	ACCACGGCGTTTCCCCTG
Revec	Reverse	6885	CCTCCTCAGACATCTTTGTGTTTC
Revec	Reverse	6671	GTGTGGTTCGGGATCGATGTG
Revec	Reverse	6473	TCCGTCCATGCACACTACAATG
Revec	Reverse	6284	GAGCATGTCCTGTCCTTTCAC
Revec	Reverse	6110	GTCGAGGATGAGCTCAACAG
Revec	Reverse	5928	GTTTCTGTCTCTCCATCGGACC

Revec	Reverse	5714	GGTCTCCAGAGGGTTCTTTTCC
Revec	Reverse	5546	CTCAAAGTTTTCCTTCAGGCG
Revec	Reverse	5341	ACCTTCTCGTGGAGCTCGAC
Revec	Reverse	5119	TTGGCGCTCACGTCAATGTC
Revec	Reverse	4956	CCGTAGTTGAAACCATTTGGG
Revec	Reverse	4780	AGCACGTTCGCAAGGAAACTC
Revec	Reverse	4575	GCTGCTTTTCAAGGATGCC
Revec	Reverse	4399	CTCTCAAGGTCTTCGGGTGTG
Revec	Reverse	4196	GTCCTTTGACCGTGCTGCTAC
Revec	Reverse	4063	GACACCAACCGGTTAAAGTCG
Revec	Reverse	3809	CCTCTTCATGCGGTAAAGCAG
Revec	Reverse	3624	CCGTGTAAGGCAGTGCAAGC
Revec	Reverse	3432	GGGTTTGCATCAGGTCCAAC
Revec	Reverse	3215	GCTAGCTAGAACGACCAGTGCG
Revec	Reverse	3047	TGTCCCACTCCGCATGAATG
Revec	Reverse	2818	CTGTCTGAGTCCGTCTTTGTGG
Revec	Reverse	2642	ACCACCGTAACCGTCGCTAC
Revec	Reverse	2440	CCAACAAAGGGCACAGTGATG
Revec	Reverse	2263	CAACCGTTTCTCATGTAAGCATAAG
Revec	Reverse	2065	GCTGTTGCGTACCCGTAAGTG
Revec	Reverse	1895	CCAGTCATTGTTCTGAGTGTTGG
Revec	Reverse	1671	CTTTCCATTCTCCGTTGAGC
Revec	Reverse	1480	GTCCCGTCTACCATGCACAC
Revec	Reverse	1288	GGTTCATCAACGTACCTAAACAAC
Revec	Reverse	1055	GGTCACCTATTCAGGCGTAGAAG
Revec	Reverse	864	GTTGCACACATGGTGGGTC
Revec	Reverse	700	CCCCTCTAGACCTGGAAAGACC
Revec	Reverse	483	GGGAAAGGCGGTATCATCTTG
Revec	Reverse	371	GGGTGAAAGGTGGGCTTCG
Revec	Reverse	168	GTTCGTCCGCACAGTCCAG

* Numbered according to AJ539141

2.4 Results

2.4.1 Amplification of complete genome by overlapping PCR products

A method to amplify the complete genome of Pan Asia O UKG FMDV through 5 overlapping PCR products, each of <3kb has been established. The detailed protocol is included as Appendix 1. Sequencing of these PCR products using the primers detailed in Table 2.2 provides good coverage of the genome, with 64 primer determined nucleotides. Figure 2.10 outlines the overall protocol for this method. This strategy can now be used to investigate FMDV nucleotide evolution during the UK 2001 FMD outbreak.

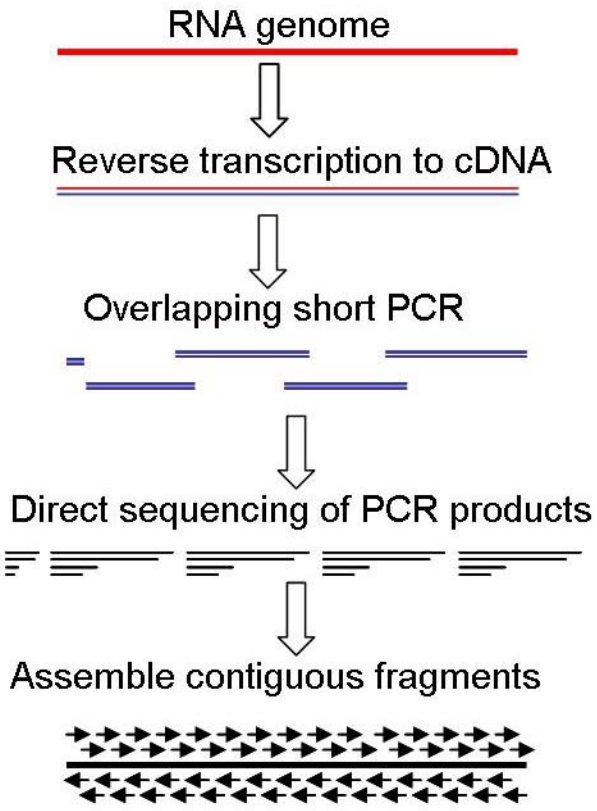


Figure 2.10

Schematic of sequencing strategy through overlapping PCR products

2.4.2 Amplification of long PCR fragments

The region of the FMDV genome that represents the complete poly protein was successfully amplified in a single PCR reaction. Additionally a PCR primer set was designed to amplify 80% of the genome, which allowed the exploration of sequencing strategies involving external collaborators.

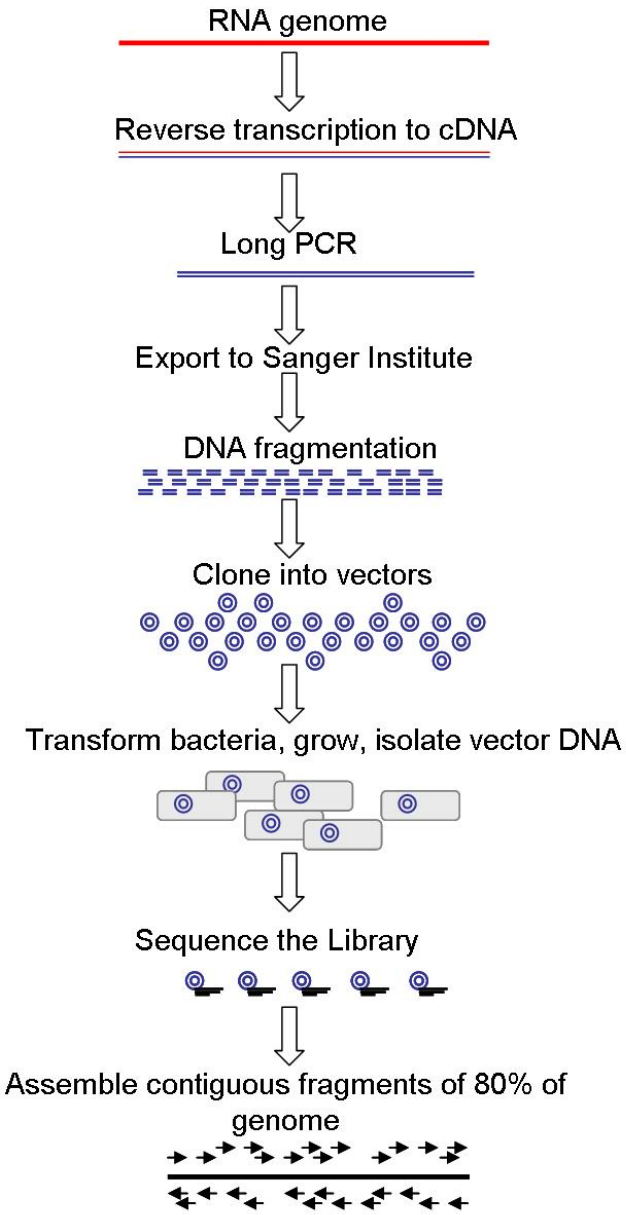


Figure 2.11

Schematic of sequencing protocol in collaboration with the Wellcome Trust Sanger Institute

The Sanger institute use the ‘shot gun’ sequencing method, as outlined in Figure 2.11. Although it was only possible to sequence 80% of the genome, this method is advantageous because prior knowledge of the expected sequence is not required.

2.4.3 Assessment of the reproducibility of the complete protocol

One piece of epithelium was sequenced twice independently. When the two sequences were compared they were found to be identical. The sequences provided by the Sanger Institute for the 6526nt PCR product from UKG samples 127/2001, 11676/2001, 14603/2001, 14339/2001 and 14391/2001, were compared to the complete genome sequences generated by the method of overlapping PCR products. The pairs of sequences were all found to be identical.

2.5 Discussion

A method to sequence the complete genome of UK Pan Asia O serotype FMDV using overlapping PCR products has been developed. The complete sequencing of virus from an epithelium sample can be completed within a week and costs approximately £500 (in reagents) per FMDV genome. Using the robot for assisting in the sequence reaction set up decreases the risk of human error; however it was found that using a multi-pipette instead can improve the proportion of successful reactions. This could be because the pipette is more accurate in dispensing liquid, or that the reaction is mixed better, or because it is possible to add the DTCS Quickstart master mix last to the reaction when setting it up by hand (the robot has to add it first to avoid contamination of the other reactions).

A successful pilot study was undertaken to sequence a large proportion of the virus genome in collaboration with the Wellcome Trust Sanger Institute, assessing the possibility of utilising their genome sequencing expertise and resources to study FMDV. Although this has been useful, it is important to note that in providing the PCR product much of the time consuming work was done at IAH Pirbright. Another disadvantage of this programme for sequencing was the limitations enforced by bio security, resulting in it only being possible to sequence 80% of the FMDV genome which is smaller than the poly protein. However, the method of nucleotide sequencing employed by the Sanger Institute does not require prior knowledge of the expected sequences for sequence primer design, which would be a clear advantage especially when sequencing FMDV strains de novo.

The introduction of error through the amplification process, for example through Taq infidelity, was of initial concern. However, the method for complete genome sequencing from overlapping PCR products was found to be reproducible, and this conclusion is

reinforced by the fact that four sequences determined at the Sanger Institute from a different PCR product were identical to those produced at the Institute for Animal Health for the same infected samples.

The method detailed here did not always work if there was a low initial virus concentration. For this reason if the method is to be used in the future, it may be worthwhile incorporating a cDNA amplification step (Carrillo et al. 2005). Another option would be to investigate the probability of introducing mutations upon propagation in cell culture; if this was negligible it could form an alternative strategy to increase virus concentration.

The method is very specific to UK Pan Asia O serotype FMDV, because it was designed to investigate the UK 2001 outbreak of the disease. If complete genome sequencing is to be used as an epidemiological tool in the future, it is essential to adapt the method such that it holds the capacity to sequence all serotypes. Another improvement to the method would be to optimise the PCR reactions further, so that they are all able to run on the same thermal cycling programme, which would make the protocol more user friendly. Novel sequencing technologies are continuously being developed such as microarray based, and pyrosequencing based technologies that may play a greater role in FMDV sequencing in the future.

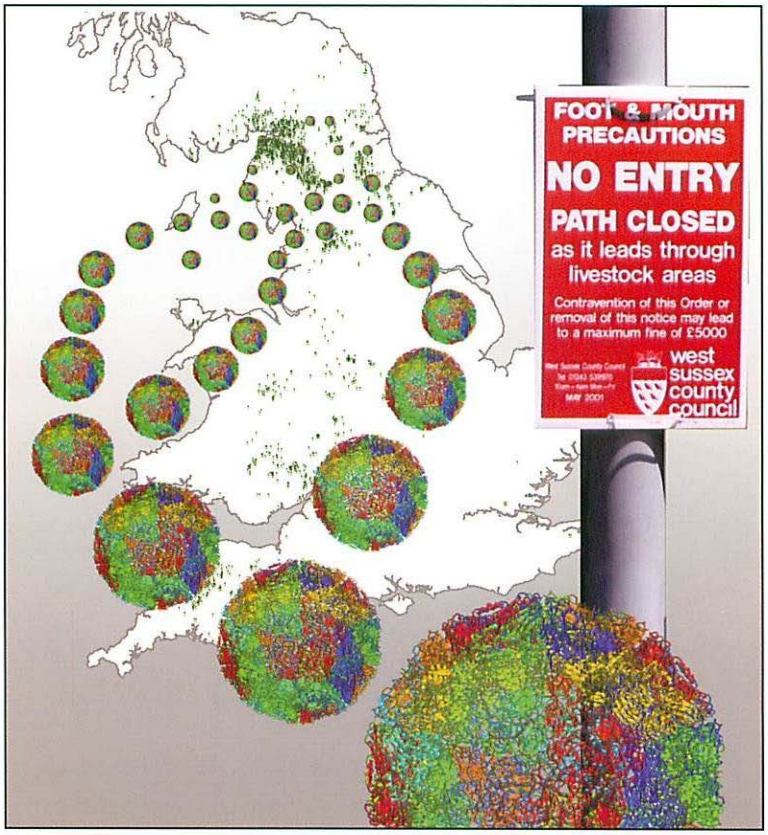
Chapter 3

Molecular epidemiology of the Foot-and-mouth disease outbreak in the United Kingdom in 2001

The work in this chapter has been published;

Cottam, E.M., Haydon, D.T., Paton, D.J., Gloster, J., Wilesmith, J.W., Ferris, N.P., Hutchings, G.H., King D.P. 2006 Molecular Epidemiology of the Foot-and-Mouth Disease Virus Outbreak in the United Kingdom in 2001 *Journal of Virology,* 80, 22, 11274-11282

VOLUME 81, NUMBER 1 JOURNAL OF VIROLOGY JANUARY 2007 Pages 1–438



Published
Twice Monthly
by the
American
Society
for
Microbiology

Journal of Virology

AMERICAN SOCIETY FOR MICROBIOLOGY

JANUARY 2007, VOLUME 81, NUMBER 1

Cover photograph (Copyright © 2007, American Society for Microbiology. All Rights Reserved.): The transmission history of foot-and-mouth disease virus (FMDV) can be elucidated from complete genome sequence analysis of the virus, enabling epidemiological tracing of the virus between infected premises. The United Kingdom map shows the location of premises infected during the 2001 FMDV outbreak. The structure is a representation of FMDV British field strain (serotype O), showing the alpha-carbon backbone (accession IFOD), courtesy of Nick Knowles, Institute for Animal Health, Pirbright, United Kingdom. *(See related article in November 2006: vol. 80, issue 22, p. 11274.)*

3.1 Summary

The objective of this study was to quantify the extent to which the genetic diversity of Foot-and-Mouth Disease Virus (FMDV) arising over the course of infection of an individual animal becomes fixed, transmitted to other animals, and thereby accumulates over the course of an outbreak. Complete consensus sequences of 23 genomes (each of 8200 nucleotides) of FMDV were recovered directly from epithelium tissue acquired from 21 farms infected during the UK 2001 FMDV outbreak over a period of nearly 7 months. Analysis of these consensus sequences revealed very few apparently ambiguous sites, but clear evidence of 197 nucleotide substitutions at 191 different sites. We estimated the rate of nucleotide substitution to be 2.26×10^{-5} per site per day (95% CI $1.75\text{-}2.80\times10^{-5}$), and nucleotide substitutions to accrue in the consensus sequence at an average rate of 1.5 substitutions per farm infection. This is a sufficiently high rate that detailed histories of the transmission pathways can be reliably reconstructed. Coalescent methods indicated that the date at which FMDV first infected livestock in the UK was 7th February 2001 (95% CI Jan 20th – Feb 19th), identical to estimates obtained on the basis of purely clinical evidence. Nucleotide changes appeared to have occurred evenly across the genome and within the open reading frame the ratio of non-synonymous to synonymous change was 0.09. The ability to recover particular transmission pathways of acutely acting RNA pathogens from genetic data will help resolve uncertainties about how virus is spread, and could help in the control of future epidemics.

3.2 Introduction

Foot and mouth disease virus (FMDV), discovered by Loeffler and Frosch in 1898, is a member of the genus aphthovirus, in the family *Picornaviridae*. The virus replicates within pigs, cattle, sheep and other cloven-hoofed animals (Domingo et al. 2002; Hughes 2004; Racaniello 1996). FMDV RNA polymerase has a high error rate, estimated to be in the order of 10^{-3}–10^{-5} misincorporations/nucleotide/genome replication, which is not mitigated by a proof reading mechanism (Carrillo et al. 1998; Domingo et al. 2001; Haydon et al. 2001). Thus nucleotide changes during a single replication event are likely to be common since the virus genome is 8,500nt long (including poly 'A' and poly 'C' tracts) and replicates via a negative strand intermediate. The high mutation rate of this virus, coupled with a fast replication rate and extensive population size, results in rapid evolution of FMDV (Domingo et al. 2003).

In 2001, the UK livestock industry was devastated by an epidemic of FMDV, caused by the Pan Asia O strain of the virus (Gibbens et al. 2001; Gibbens & Wilesmith 2002). The disease was rapidly and widely disseminated throughout the UK and a total of 2030 farms were declared 'infected premises' (IP). The extent of the outbreak and speed with which the virus spread hindered attempts to trace the exact origin of infection for many premises; numerous cases were simply attributed to 'local spread'. Samples of infected tissue were collected from the majority of IP prior to culling of the animals and stored at the Institute for Animal Health, Pirbright. These samples provided a unique opportunity to study the micro-evolution of this virus strain over the 7 months that this epidemic persisted.

It is evident from previous studies, both *in vitro* (see Chapter 1; section *1.4.3*), and *in vivo* (Carrillo et al. 1998), that FMDV mutation and substitution rates are high. It is much less clear as to what extent the genetic diversity, anticipated to be generated over the course

of infection within an individual animal, becomes fixed, transmitted to other animals, and thereby accumulates over the course of an outbreak. Quantitative models of viral genetic change over short time periods suggest this rate of accumulation should be readily measurable and informative of transmission pathways at high resolution (Haydon et al. 2001). Strong tests of this prediction would be provided by evidence that the consensus sequence (acquired directly from tissue samples, rather than through prior passage in cell culture), which reflects the average of the viral genetic diversity within a sample, evolves at a sufficient rate to enable reliable tracing of transmission pathways.

Studies considering the nucleotide sequence encoding the VP1 capsid protein of FMDV alone have provided valuable insight into the emergence of various strains and serotypes worldwide over recent decades (Bronsvoort et al. 2004; Knowles & Samuel 2003; Sangare et al. 2003). Changes observed within VP1 over the course of single outbreaks have also been previously documented (Christensen et al. 2005; Jinding et al. 2006; Mohapatra et al. 2002; Sangare et al. 2003; Tsai et al. 2000), but the relatively short length of the sequence encoding VP1 (633nt) together with the short infectious period, has limited the power of these analyses to reconstruct transmission pathways at high resolution. Complete genome analysis has the potential to resolve transmission histories and improve the precision of epidemiological investigations. The ability to determine the exact source of infection of IP from the UK 2001 outbreak would help us understand the route of transmission of infection between farms. Such information is invaluable for the control of future epidemics.

Here we report 23 complete consensus genomes, sequenced directly from clinical epithelial samples collected from 21 IP from the UK 2001 outbreak. These samples were carefully chosen with the aim of first characterising the extent of variation within the outbreak, secondly analysing the relationship between time and accumulation of genetic

change, and thirdly to determine if sequence data can map known transmission events and aid analysis of spatio-temporal clusters of IP with uncertain transmission histories.

3.3 Methods

3.3.1 Virus samples

The consensus FMDV genome was amplified by RT-PCR and sequenced from clinical epithelium samples taken from 21 infected premises (Figure 3.1). All epithelium samples had been previously found positive by virus isolation. IP were sampled from every month of the outbreak.

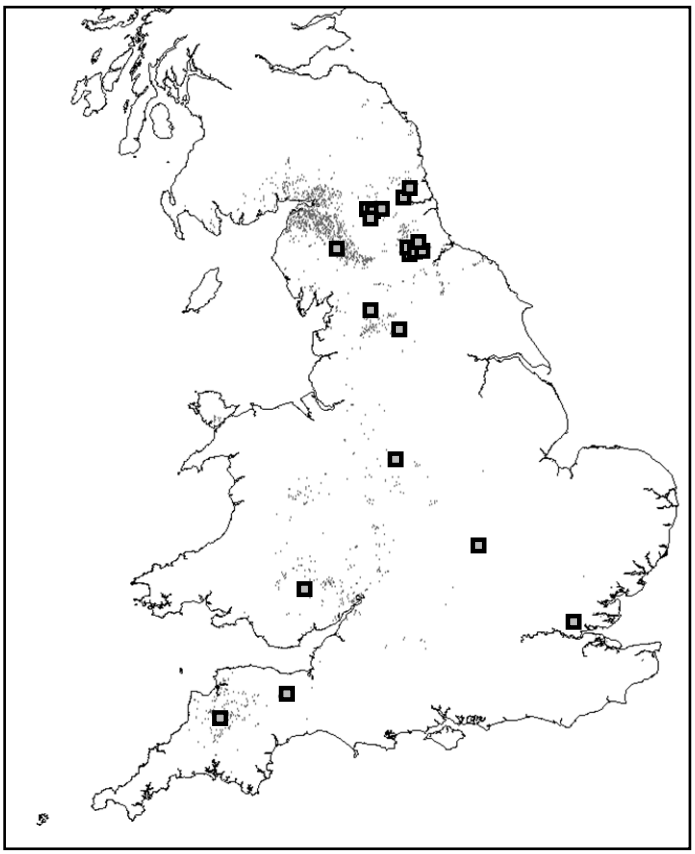


Figure 3.1

Location of FMDV infected premises investigated. The figure shows a map of Great Britain with total designated FMD infected premises (·) in 2001; 20 of those from which virus has been sequenced are highlighted (□).

The samples were from the earliest IP and also from a representative IP of eight geographical clusters sampled at a later time. The known transmission events studied involved IP at the start of the outbreak for which transmission histories were clearly established from contact tracing. Uncertain transmission histories were studied by

investigating a cluster from County Durham. Accession numbers and sequence coverage are detailed in Table 3.1.

3.3.2 Epithelium Sample Suspension

As described in Chapter 2, Section 2.4.1a.

3.3.3 RT-PCR amplification of samples

TRIzol (Invitrogen) RNA extraction was performed as described previously (Reid et al. 1998). Ten µl RNA, 4µl 10mM primer UKFMD/Rev6 ($^{5'}$GGC GGC CGC TTT TTT TTT TTT TTT$^{3'}$), 4µl 10mM random hexamers (Promega), 2µl 10mM dNTP mix was incubated at 68°C for 3 minutes, then on ice for 3 minutes. Eighteen µl of freshly prepared RT mix (4µl 10x RT buffer (Invitrogen), 8µl 25mM $MgCl_2$, 4µl 0.1M DTT, 2µl RNase OUT (Invitrogen)) was added to the sample followed by 2µl Superscript III Reverse Transcriptase (Invitrogen). The sample was then incubated at 42°C for 4 hours after which the reaction was stopped by incubation at 85°C for 5 minutes. The cDNA was then cleaned using QIAquick PCR Purification Kits (Qiagen), eluting in 30µl of elution buffer before storage at -20°C. Five overlapping PCR fragments covering the FMD genome were amplified from each sample, using 47.5µl of master mix (5µl 10x buffer, 2µl $MgSO_4$, 1µl 10mM dNTP mix, 1 µl 10mM forward primer, 1µl 10mM reverse primer, 0.2µl Platinum Taq Hi-Fidelity (Invitrogen), 37.3 µl nuclease-free water) plus 2.5µl cDNA. The five primer sets were as follows 1) forward $^{5'}$TTG AAA GGG GGC GTT AGG GTC TCA$^{3'}$ reverse $^{5'}$GGG TGA AAG GTG GGC TTY GGG T$^{3'}$, 2) forward $^{5'}$CCC AAG TTT TTA CCG CCT TTC CCG$^{3'}$ reverse $^{5'}$GTT GAT AAT GCT TCC AGT GTT GCC TG$^{3'}$, 3) forward $^{5'}$CCA CGC TGG CAT CTT CCT GAA AG$^{3'}$ reverse $^{5'}$CCA GTG GCC AGT TCC TCA AAT GC$^{3'}$, 4) forward $^{5'}$GTG TTG GAC CTG ATG CAA ACC CC$^{3'}$ reverse $^{5'}$GTC TCT TGC GAG TCT CGC GGA TC$^{3'}$, 5) forward $^{5'}$TTC AAG CCT CAA CCG CCC CTC$^{3'}$ reverse $^{5'}$GGC GGC CGC TTT TTT TTT TTT TTT$^{3'}$ (see also Chapter 2, Table 2.1). Primer sets 1-4 were run on a PCR programme cycle of initial denaturation at 94°C for 2 minutes,

then 39 cycles of 94°C 30 seconds, 68°C 30 seconds, 72°C 3 minutes, and ending with incubation at 72°C for 7 minutes. Primer set 5 was run on a cycle of initial denaturation at 94°C for 2 minutes, then 39 cycles of 94°C 45 seconds, 60°C 45 seconds, 72°C 3 minutes, and finishing with incubation at 72°C for 7 minutes. PCR products were cleaned up using QIAquick PCR Purification Kits (Qiagen), eluting in 30µl elution buffer. In order to estimate DNA content, 2µl of the PCR product was run on a 1.2% agarose gel at 90v for 35 minutes alongside a quantitative ladder (HyperLadder I, Bioline). The DNA concentrations of these products were adjusted to give the recommended copy number for sequencing; (fragment 1: 20ng/µl, fragment 2: 70ng/µl, fragment 3: 100ng/µl, fragment 4: 100ng/µl, fragment 5: 100ng/µl), and stored at -20°C.

3.3.4 Sequencing reactions

Forty-two forward and 42 reverse sequencing reactions (primers are listed in Table 2.2, Chapter 2) were performed in a 96 well Beckman Sequencing plate using 8µl of master mix (CEQ Dye Terminator Cycle Sequencing with Quick Start Kit, Beckman Coulter), 4µl of 1mM primer, and 8µl DNA per reaction. A Qiagen Biorobot 3000 was used to set up the reactions. The plate was run on a programme of 30 cycles of 96°C for 20 seconds, 50°C for 20 seconds, and 60°C for 4 minutes. Following thermal-cycling, the reactions were cleaned up by ethanol precipitation before running on the Beckman Coulter Sequencing machine.

3.3.5 Data Analysis

The raw data files were assembled into a contig and edited using Seqman (DNASTAR). All further sequence manipulation was performed using Bioedit, and DNAsp freeware. The phylogenetic reconstruction (tree shown in Figure 3.2) was carried out using maximum likelihood methods as implemented in Paup* (version 4b 10, (Swofford 2001)), assuming an HKY model of base substitution (variable base frequencies and variable transition and transversion frequencies (Hasegawa et al. 1985)) with rate heterogeneity.

Bayesian posterior probabilities were obtained from 500,000 generations sampling every 100th tree using the software package Mr Bayes (Huelsenbeck & Ronquist 2001). The genealogical relationships shown in Figure 3.3 were based on statistical parsimony as implemented in the software package TCS (Clement et al. 2000) assuming gaps to be missing nucleotides (Figure 3.3). For this statistical parsimony analysis ambiguities were replaced with the conserved base. A molecular clock was fitted using MCMC techniques implemented in the software package BEAST (Bayesian Evolutionary Analysis Sampling Trees) (Drummond et al. 2002)) assuming a relaxed clock (Ho et al. 2005) with exponentially distributed rates and exponentially increasing population size, and the HKY model of base substitution (Hasegawa et al. 1985) with rate heterogeneity. The number of substitutions over the phylogeny was estimated using DNASp, and the non-synonymous: synonymous ratios (*dN/dS*) were estimated using Hy-Phy software package (Pond et al. 2005). Hy-Phy was also used for partition analysis to determine if selective pressures (*dN/dS* ratios) differed significantly between genes using a Muse-Gaut 94 substitution model (Muse & Gaut 1994). The Kolmorogov-Smirnov (KS) test was used to investigate the positioning of mutations along the genome. All sequences have been submitted to GenBank and assigned the accession numbers DQ404158-DQ404180.

3.4 Results

Twenty-three clinical samples selected from the UK 2001 outbreak (Table 3.1) were examined by genome sequencing of FMDV. All sequences were unique and comprised the complete genome excluding the poly 'C' and poly 'A' tracts, and 64 primer determined nucleotides (24nt at 5' end, 19nt and 21nt either side of internal poly 'C' tract). The nucleotide sequence changed considerably during the outbreak with variation arising at 191 nucleotide sites. Ambiguities within the sequences were rare, occurring in only three isolates (DQ404180, DQ404160, DQ404167). DQ404180 contained three ambiguous sites: two between pyrimidines T/C at nucleotide positions 2833 and 5734 and one between purines A/G at position 4744. DQ404160 contained an ambiguity between T/C at position 1270. DQ404167 contained a T/C ambiguity at position 6207 and also an ambiguity between nucleotides and a deletion at bases 8111-8115. This supports the view that there is a predominant consensus sequence present within an epithelium sample (or one that is preferentially amplified by PCR). Four deletions were identified within the 5' untranslated region (UTR), three of which were inherited by progeny viruses recovered later in the course of the outbreak. The deletion that was present in one isolate only was one base and followed a single nucleotide insertion 20nt upstream. A deletion of 5 nucleotides was also identified in the 3'UTR of one isolate (DQ404167), although present as an ambiguity with some of the population containing the full sequence.

Table 3.1 Details of the 23 FMDV consensus genomes sequenced in this study

Accession No.	*Infected Premise No.*	*World Reference Laboratory No.*	*Animal*	*Date of Sample isolation*	*County*	*Sequence length (nt)*	*Total nt sequenced*	*Average times coverage of each base*
DQ404180	1	UKG 11/2001	Pig	19/02/2001	Essex	8193	40,050	4.89
DQ404179	4	UKG 126/2001	Pig	22/02/2001	Northumberland	8193	42,722	5.21
DQ404178	4	UKG 127/2001	Pig	22/02/2001	Northumberland	8193	48,838	5.96
DQ404177	4	UKG 128/2001	Pig	22/02/2001	Northumberland	8193	32,044	3.91
DQ404176	6	UKG 150/2001	Cattle	23/02/2001	Northumberland	8193	44,636	5.45
DQ404175	7	UKG 173/2001	Cattle	24/02/2001	Devon	8193	44,296	5.41
DQ404174	1700	UKG 438/2001	Sheep	28/02/2001	Ireland	8193	42,686	5.21
DQ404173	16	UKG 220/2001	Cattle	26/02/2001	Northampton	8195	42,760	5.22
DQ404172	38	UKG 621/2001	Cattle	01/03/2001	Staffordshire	8193	44,194	5.39
DQ404171	1070	UKG 4569/2001	Sheep	06/04/2001	Northumberland	8193	47,008	5.74
DQ404170	1448	UKG 7675/2001	Sheep	23/04/2001	Durham	8193	39,528	4.82
DQ404169	1404	UKG 7038/2001	Sheep	20/04/2001	Durham	8193	35,038	4.28
DQ404168	1575	UKG 9011/2001	Cattle	10/05/2001	W. Yorkshire	8193	43,915	5.36
DQ404167	1597	UKG 9327/2001	Sheep	14/05/2001	Durham	8188	48,213	5.89
DQ404166	1654	UKG 9788/2001	Cattle	28/05/2001	Durham	8193	42,169	5.15
DQ404165	1692	UKG 9964/2001	Cattle	03/06/2001	Durham	8193	42,147	5.14
DQ404164	1757	UKG 11676/2001	Cattle	16/07/2001	Devon	8193	44,887	5.48
DQ404163	1945	UKG 14339/2001	Cattle	31/07/2001	Powys	8191	37,185	4.54
DQ404162	1956	UKG 14476/2001	Sheep	08/08/2001	W. Yorkshire	8192	45,421	5.54
DQ404161	1948	UKG 14391/2001	Cattle	14/08/2001	Cumbria	8189	44,293	5.41
DQ404160	1970	UKG 14524/2001	Cattle	23/08/2001	Northumberland	8187	42,745	5.22
DQ404159	1976	UKG 14603/2001	Cattle	26/08/2001	Northumberland	8188	45,752	5.59
DQ404158	2027	UKG 15101/2001	Sheep	17/09/2001	Northumberland	8187	45,178	5.52

3.4.1 Maximum likelihood phylogenetic analysis of complete genome sequences

Groups of viruses that reflect the known geographic spread of FMD were identified by phylogenetic reconstruction of the outbreak using maximum likelihood methods (tree shown in Figure 3.2). The transition transversion ratio was estimated to be 7.61 (the alpha parameter governing the gamma distribution for rate heterogeneity was estimated as infinity). All internal branches within the phylogeny were supported by high Bayesian posterior probabilities in excess of 96%. Virus in different geographical areas evolved along separate lineages resulting in a large breadth of variation within the outbreak with 197 nucleotide substitutions at 191 sites along the genome. Thus six sites were identified to have been mutated twice, higher than that expected on the basis of a Poisson process with mean 197/8196 in which all sites were assumed to be equally mutable (in which an average of 2.2 sites would be expected to receive multiple hits).

3.4.2 Fine-scale statistical parsimony analysis (TCS)

Statistical parsimony analysis (Figure 3.3) revealed that virus isolated from the first infected premise (IP4) and the last (IP2027) differed by 44 silent (synonymous) and 8 non-silent (non-synonymous) substitutions, and there were also 2 nucleotide deletions (one of 4 bases and one of 2 bases in the 5'UTR). At a finer resolution, viruses from closely related farms differed by between 1 and 5 nucleotides. Virus sequences obtained from 3 pigs sampled on the same day on a single premise (IP4), where disease had been present longer than on other farms, differed in consensus by 1-3 nucleotides.

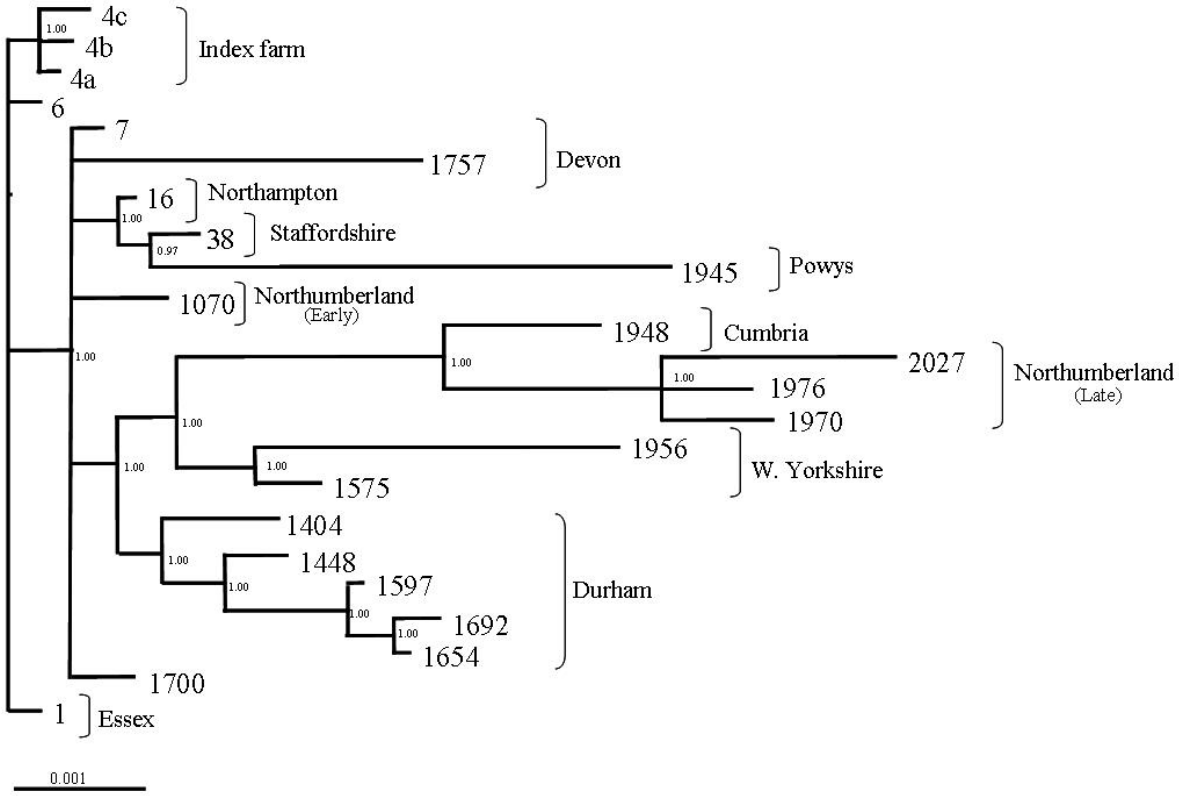


Figure 3.2

Maximum likelihood phylogenetic reconstruction of 23 virus sequences, named by the infected premises (IP) number from which they were taken. The county where each IP was situated are indicated by brackets. Bayesian posterior probabilities as indicated, were obtained from 500,000 generations sampling every 100th tree.

3.4.3 Genomic distribution of nucleotide substitutions

Twenty-eight substitution sites were identified within the non-coding regions at the 5 and 3 prime end of the genome, and 163 substitution sites within the open reading frame. Of these, 40 resulted in non-synonymous change, and 123 were synonymous, resulting in an overall non-synonymous to synonymous ratio (*dN/dS,* estimated per synonymous and non synonymous site) of 0.09. The distribution of synonymous changes throughout the genome could not be distinguished from a uniform distribution (Figure 3.4) (KS test, D=0.0424, P>0.5). Non-synonymous changes appeared more clustered (Figure 3.4),

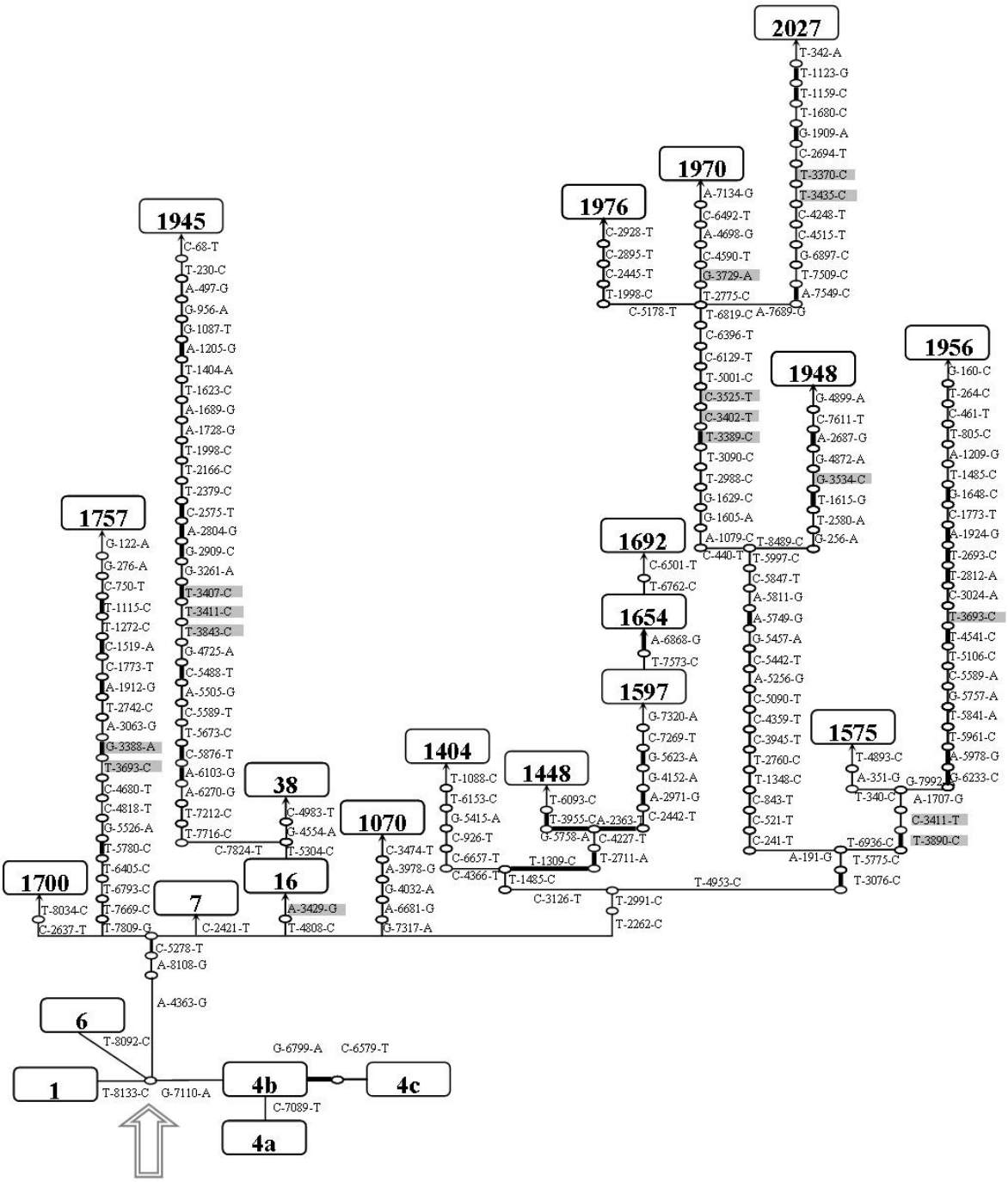


Figure 3.3

TCS analysis of 23 United Kingdom Pan Asia O FMDV sequences. Statistical parsimony analysis of 23 sequences by TCS; each sequence is represented by the IP number from which it was isolated. Each connecting branch line represents a nucleotide substitution, with each dot representing a putative ancestor virus. Non-synonymous mutations are shown in boldface. The arrow represents the introduction of virus into the United Kingdom via IP4, where replication occurred for ~2 weeks before clinical samples were collected. IP4 was the source of virus for IP1, where the disease was first detected, and also for IP6, from where

virus was disseminated throughout the United Kingdom. IP2027 is the last IP from the outbreak. Grey shading indicates changes located within VP1.

although the KS test (D = 0.1845, P>0.2) again did not distinguish their distribution from a uniform one. Only a single amino acid position (position 42 in VP1) underwent two independent substitutions. These were on two distinct geographical genetic lineages as a result of different nucleotide substitutions in adjacent sites: in Northumberland, this was a change from valine to alanine, and in Devon from valine to isoleucine. Contrary to previous studies (Carrillo et al. 2005), VP1 was not found to be more variable than other parts of the genome, incurring less substitutions than other genes such as VP3 (Figure 3.4).

3.4.4 Pattern of dN/dS ratio throughout genome

The observed *dN/dS* ratios (corrected per synonymous and non synonymous site) of each genomic region differed as shown in Figure 3.4 (ranging from as low as 0 for 2A, or 0.1 for 2C to 1 for VP4). However, fitting models assuming different *dN/dS* ratios for each genomic region (using partition analysis in HyPhy) did not significantly increase the likelihood of the model fit (chi-sq.= 7.724, d.f.=11, P=0.74) and the null hypothesis that all *dN/dS* ratios were equal could not be rejected on the basis of these data. VP4 had the highest *dN/dS* ratio overall, equal to 1, and thus there was no evidence of positive selection occurring at an amino acid level anywhere in the genome.

3.4.5 Rate of nucleotide substitution over time

Total nucleotide changes from the ancestor virus were shown to accrue linearly with time; synonymous changes accumulated faster than non-synonymous changes (Figure 3.5). A relaxed molecular clock for the rate of substitution of all nucleotide changes (estimated using BEAST was found to advance at a rate of 2.26×10^{-5} per site per day (95% CI $1.75\text{-}2.80\times10^{-5}$).

This analysis also predicted that the most recent common ancestor to the viruses sequenced was present 12 days before recovery of the earliest FMDV sequenced in this study. This would date the start of the outbreak to the 7th February 2001 (95% CI Jan 20th – Feb 19th) which is identical to estimates obtained on the basis of purely clinical evidence (Anderson 2002; Scudamore 2002). Previous studies have suggested that there may have been 35 sequential IP-IP transmission events between IP4 and the final case (IP2027) (Haydon et al. 2003), suggesting that substitutions were fixed across the genome at an average rate of 1.5 nucleotide changes per IP transfer.

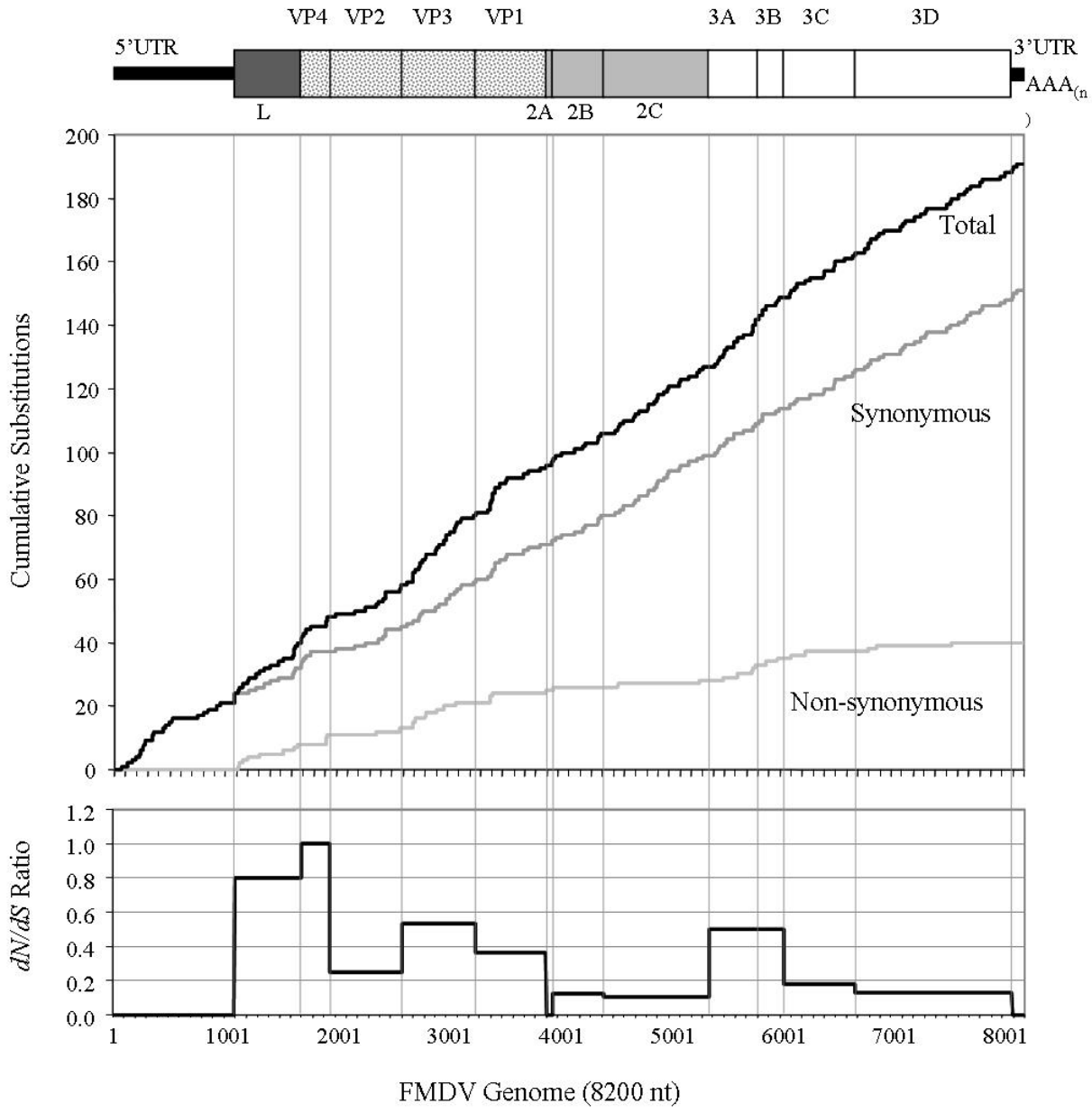


Figure 3.4

Synonymous and non-synonymous mutations across the genome. Shown is a graph of cumulative substitutions against nucleotide positions, depicting the rate of substitution for total (black), synonymous (grey), and non-synonymous (light grey) point mutations. The rate of accumulation of silent changes throughout the genome is constant and faster than the rate of amino acid substitutions which also appears to vary more. The *dN/dS* ratio for each genomic region is shown in the lower graph, but likelihood analyses provide no support for different *dN/dS* ratios for different genes.

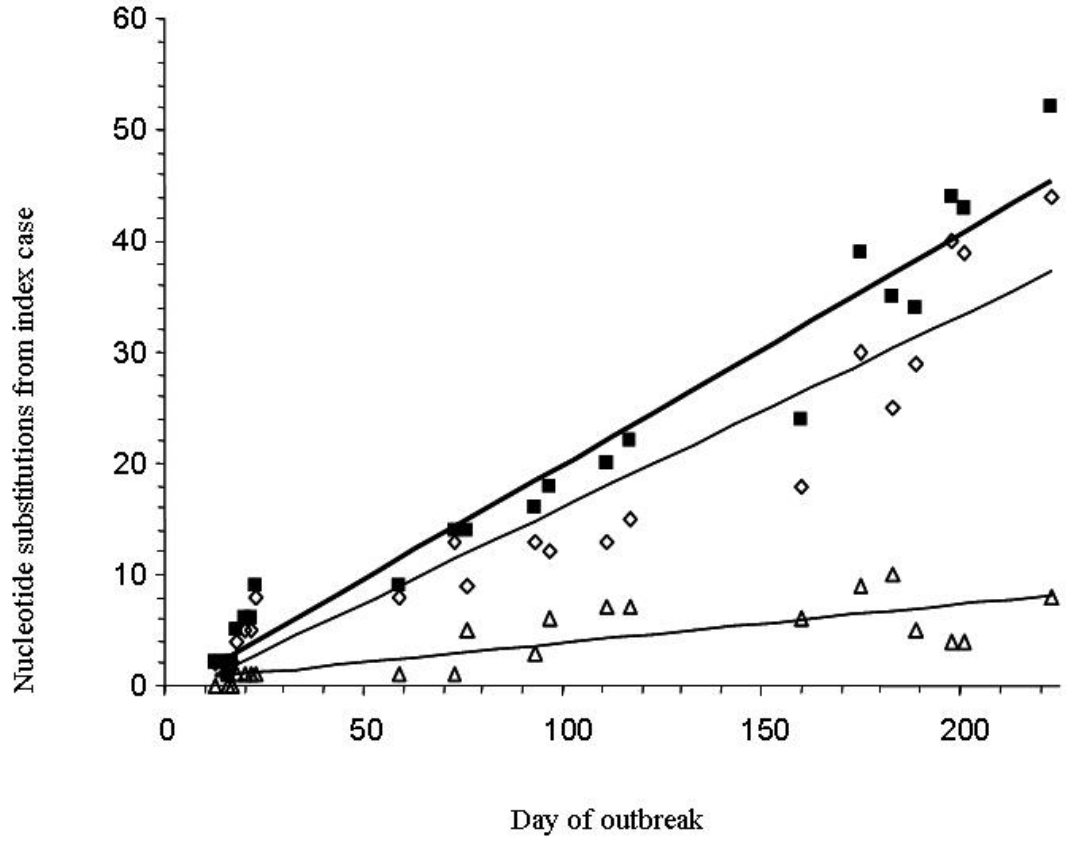


Figure 3.5

Accumulation of substitutions with respect to time. Shown is the accumulation of mutations (a measure of genetic distance) increasing linearly with time, compatible with a molecular clock. ■, all substitutions; ◊, silent substitutions; Δ, non silent substitutions.

3.4.6 Use of sequence data to trace transmission pathways

The potential of genetic analysis to enable tracing of FMD transmission history within the outbreak was investigated. A known chain of transmission events between four farms was studied where the routes of virus transmission had been established with confidence using contact tracing, and also a cluster of five farms for which the routes of transmission were uncertain.

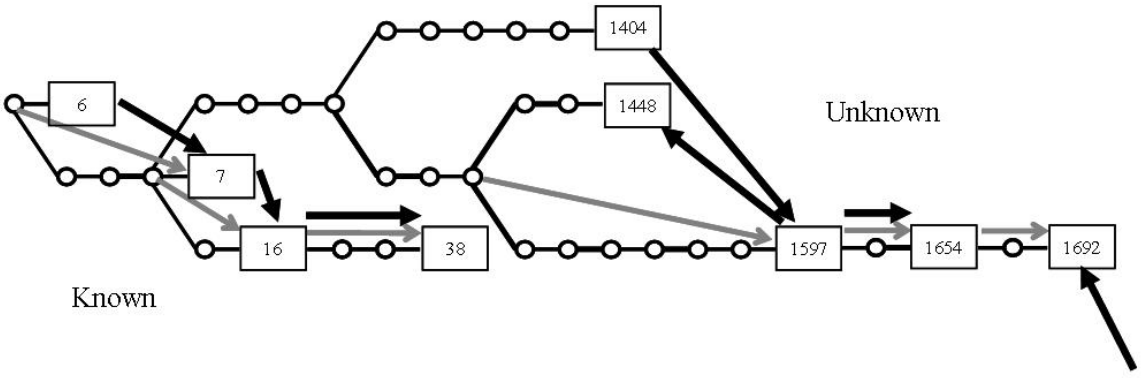


Figure 3.6

TCS analysis of two clusters of infected premises. Statistical parsimony analysis of known and unknown virus transmission events. Each connecting branch line represents a nucleotide substitution, with each dot representing a putative ancestor virus. Black arrows indicate the transmission route determined by traditional contact tracing, grey arrows indicate the route supported by the genetic data. The data show the close genetic relationship of IP1692 to IP1597 and IP1654, although epidemiological tracing suggested IP1692 was infected from a more distantly related source.

The genetic data were consistent with the results of the conventional contact tracing for the known chain of transmissions where virus spread from IP4 to IP6 via an airborne route, and then via sheep movement from IP6 through markets to IP7, and subsequently to IP16 and IP38 via a lorry (Figure 3.6). However, parsimony based analysis of a cluster of IP in County Durham, for which the transmission sequence was less certain, suggested that the ancestor virus of a certain farm differed from that suggested by previous tracing studies which had suggested that IP1692 was unrelated to IP 1404,1448,1597,1654 (Figures 3.6, 3.2 and 3.3). The genetic data indicated that IP1692 was very closely related to the others, as shown in Figure 3.6, linking it to the other four in the cluster.

3.5 Discussion

By studying virus from 21 premises, infected during the 2001 UK FMDV epidemic we have provided the first intra-epidemic analysis of FMDV by full genome sequencing. This has demonstrated the potential of such studies to provide detailed and reliable descriptions of the sequence of transmission events at high spatial and temporal resolutions and provided useful insights into the mode and tempo of FMD viral micro-evolution.

Complete consensus sequences were clearly amplified with very few ambiguities found, indicating that this method was not affected by the proposed heterogeneous nature of virus populations. Sequence was obtained directly from epithelium samples avoiding passaging virus through animals or cell culture. The virus isolate from the abattoir in Essex (IP1) was compared to an isolate sequenced previously O UKG 35/2001 (AJ539141) (Mason et al. 2003b) from the same abattoir. O UKG 35/2001 had been passaged through 2 pigs prior to direct sequencing. Notably the virus consensus sequence derived directly from epithelium of an animal from IP1 is very close to O UKG 35/2001 differing only at the three ambiguous sites identified within this isolate. It also differed from the remainder of the UK viruses sequenced at these same 3 sites, suggesting that it was an intermediate between the UK viruses and O UKG 35/2001.

The genetic variation observed between viruses was mainly due to synonymous point substitutions, the majority of which were transitions. These nucleotide changes were found to occur evenly across the whole genome (as indicated by the KS test). These two factors suggest that the evolution of FMDV demonstrated to have occurred during the 2001 UK outbreak may have arisen largely as a result of genetic drift subject to purifying selection. Fixation of mutations could have occurred due to bottleneck events on transmission between animals or IP. The study includes only 23 sequences sampled

over a short time scale, which has resulted in limited statistical power to discriminate differences in substitution rates and selection pressures between genes.

The TCS statistical parsimony analysis showed clearly the genetic evolutionary history of the virus. However, it could not definitively identify the geographic location or IP on which a nucleotide substitution had occurred. This is because the exact time that virus was transmitted from a farm in relation to the time at which animals were culled and viral samples collected, is unknown. During the 2001 UK outbreak there was strong pressure to cull animals as fast as possible following identification of infection but never-the-less there were lengthy delays between the time that some IP were estimated to have become infected and the time that samples could be collected from them. When transmission from an IP occurs a long time prior to acquisition of viral samples it will remain uncertain on which IP substitutions occurred. This is apparent in our data, for example we do not have the 'real' ancestor to the outbreak, as the farm on which the outbreak began was not identified as infected until up to 3 weeks after the virus is estimated to have been introduced into Britain (Figure 3.3). This problem, along with the observation that virus recovered from closely-housed animals can differ by 1-3 nucleotides and is likely to pass through a 'bottle neck' on passage between farms, suggest that if this tool is to be used in future outbreaks a greater number of samples taken from each and every farm may improve the resolution of this method.

Virus samples were sequenced from pigs, cattle and sheep. Although no evidence was found of the effect of host tropism within these genetic data, such an effect cannot be confidently excluded using these data alone. The route by which each individual animal was infected is unknown. For example the virus could have been circulating in sheep before being subsequently identified and sequenced from a cow, which would complicate the results.

Nucleotide change conformed to a relaxed molecular clock corresponding to 0.00825 substitutions/site/year, thus 0.9% of the genome is predicted to change per annum. Additional analyses conducted assuming various strict molecular clocks, and alternate demographic models fitted the data with only slightly reduced average posterior likelihoods and produced results that were very similar. These models are suggestive of a constant rate of viral replication and transmission during the outbreak. Independent genetic determination of the timing of the first case of the outbreak, matched previous estimates based on clinical evidence. Such a tool for detecting the timing of an initial case in an outbreak could, in principle, prove very useful in identifying how virus is introduced into the country. The results reported here were consistent with other studies; for example those of serotype 'O' virus in Taiwan which found a divergence of 0.2-0.9% between VP1 genes sequenced over a two months period (Tsai et al. 2000), serotype SAT-2 which was found to change at a rate of 0.009 substitutions/nucleotide/year (Bastos et al. 2003), and serotype 'C1' which changed at an estimated rate of 0.0004-0.045 substitutions/nucleotide/year (Sobrino et al. 1986). They also correspond closely to rates measured over longer time periods with the 'O' serotype of the virus seen to change at a rate of 1.43% per year in the Philippines, and at 0.43% per year in Turkey for example (Haydon et al. 2001).

It is well established that detailed high-resolution transmission histories can be reconstructed for some RNA viruses. Transmission events have been reconstructed between individuals infected by viruses with long infectious periods using sequence data, for example for HIV virus (Leitner & Albert 1999; Zhang et al. 1997), and Hepatitis C virus (Spada et al. 2004). In the recent outbreak of SARS, human to human tracing was investigated using virus genetic sequencing (Liu et al. 2005a; Wong et al. 2004). Broad resolution of rabies (Bernardi et al. 2005; Kobayashi et al. 2005; Susetya et al. 2003) and rhinovirus (Savolainen et al. 2002) transmission networks have also been studied using genetic sequencing. Using full genome sequences enables tracing of transmission

histories of the virus at high resolution (between farms), reliably and with confidence. The capacity to conduct animal disease tracing using genetic data is of considerable value, particularly considering the difficulties of identifying transmission pathways by other means. Much of FMDV transmission between farms is poorly understood and the results of this study demonstrate the potential of this methodology to help in unravelling the transmission routes utilised by the virus. Unknown transmission histories were resolved as illustrated by the investigation of the cluster from County Durham. However, less detailed yet informative analyses were also possible. For example, there was uncertainty over the relationship between Northumberland cases that arose in April and those that arose mid-August during the epidemic. Ineffective decontamination and cleaning of formerly infected IP could have led to propagation of subsequent outbreaks. However, the late Northumberland cases (IP 1970, 1976, 2027) were shown to be only distantly related to the early Northumberland case at IP1070 (Figure 3.3) and actually originated from Cumbria.

This study shows that complete genome sequencing can be used to resolve inter-farm transmission, and as shown previously (Christensen et al. 2005) VP1 sequences alone are unlikely to contain sufficient genetic information to enable reconstruction of transmission pathways at the highest levels of resolution. Between the first and the last IP there were estimated to be 35 inter-farm transmissions during which there were 52 nucleotide changes (i.e. 1.486 substitutions/inter-farm transmission), but only 13 of these substitutions fell within the capsid genes (0.371 substitutions/inter-farm transmission), and only 5 in VP1 (0.143 substitutions/inter-farm transmission). Hence only full genome sequences have the capacity to detect more than one nucleotide change per farm transfer. However, with the advance of high-speed sequencing technology, full genome sequencing has the potential to be used in real time during future epidemics to further inform control policy.

Chapter 4

Integrating genetic and epidemiological data to determine transmission pathways of Foot-and-mouth disease virus

The work described in this chapter has been accepted for publication;

Cottam E.M.[1,2], Thébaud G.[2*], Wadsworth J.[1], Gloster J.[3], Mansley L.[4], Paton D.J.[1], King D.P.[1], Haydon D.T.[2], Integrating genetic and epidemiological data to determine transmission pathways of Foot-and-mouth disease virus, Proceedings for the Royal Society of Biology

[1]Institute for Animal Health, Ash Road, Pirbright, Surrey, United Kingdom, GU24 0NF, [2]Division of Environmental and Evolutionary Biology, University of Glasgow, Glasgow, United Kingdom, G12 8QQ, *Current address: INRA, UMR BGPI, CIRAD TA A 54/K, Campus de Baillarguet, 34398 Montpellier cedex 5, FRANCE, [3]Met Office, Fitzroy Road, Exeter, United Kingdom, EX1 3PB, currently working at 1, [4]Animal Health Divisional Office, Strathearn House, Broxden Business Park, Lamberkine Drive, Perth, PH1 1RZ.

4.1 Summary

Estimating detailed transmission trees that reflect the relationships between infected individuals or populations during a disease outbreak often provides valuable insights into both the nature of disease transmission and the overall dynamics of the underlying epidemiological process. These trees may be based on epidemiological data that relate to the timing of infection and infectiousness; or genetic data, that show the genetic relatedness of pathogens isolated from infected individuals. Genetic data is becoming increasingly important in the estimation of transmission trees of viral pathogens due to their inherently high mutation rate. Here we propose a maximum likelihood approach that allows epidemiological and genetic data to be combined within the same analysis to infer likely transmission trees. We apply this approach to data from 20 farms infected during the UK 2001 foot-and-mouth disease outbreak, using complete viral genome sequences from each infected farm, and information on when farms were first estimated to have developed clinical disease, and when livestock on these farms were culled. Incorporating known infection links due to animal movement prior to imposition of the national movement ban results in reduction of the number of trees from 41,472 that are consistent with the genetic data, to 1,728, of which 4 represent 95% of the total likelihood calculated using a model that accounts for the epidemiological data. These trees differ in several ways from those constructed prior to the availability of genetic data.

4.2 Introduction

Genetic data from RNA viruses have been used increasingly for tracing disease transmission pathways, taking advantage of their inherent capacity to evolve quickly. Such studies have been carried out with viruses such as HIV (Leitner & Albert 1999; Zhang et al. 1997), Hepatitis C virus (Bracho et al. 2005; Spada et al. 2004), SARS coronavirus (Liu et al. 2005a; Wong et al. 2004), Ebolavirus (Walsh et al. 2005), Rhinovirus (Savolainen et al. 2002) and Noro-viruses (Dowell et al. 1995). These genetic data can be used to complement field epidemiological studies that utilise traditional contact tracing information and the relative timing and spatial proximity of infection events to each other.

Analysis of the overlapping periods within which individuals or groups are infected and/or infectious can help to decide the most likely direction of transmission, while genetic pathogen data can identify which infected individuals or groups are most closely epidemiologically linked. Although there may be numerous combinations of transmission pathways that are consistent with the genetic data or epidemiological data alone, an analysis combining both types of data can lead to identification of a much smaller set of plausible transmission pathways. Analytical methods to integrate these two types of data sets are beginning to be developed (Wallace et al. 2007), but the increasing speed and economy with which viral genetic data can be generated during epidemics requires the development of wider range of approaches if full advantage of these data sources is to be taken in improving the tracing of transmission pathways.

Recently, it was shown that Foot-and-mouth disease virus (FMDV) transmission can be traced from farm-to-farm using complete genome sequencing (Cottam et al. 2006). Viruses were sequenced from farms infected at the beginning of the 2001 UK FMDV outbreak and the genetic data were shown to be consistent with the transmission

pathways established from contact tracing studies. However for the remainder of the outbreak, and following the national ban on animal movement (NMB), the spread of virus is much less well understood, with infection assumed to be transmitted either by airborne spread or by mechanical transfer on people or inanimate objects (fomites). Indeed, the precise source and route of infection for the vast majority of the 2030 premises infected during the UK 2001 epidemic remain unknown.

During the epidemic, substantial amounts of epidemiological data were recorded for the farms involved. Although the contact tracing data are difficult to use to pinpoint the precise origin of infection with high confidence, the information concerning the timing of infection is clearer. For every farm involved, the time at which clinical disease began on a farm (estimated from lesion ageing) and the date on which animals were culled was recorded. This information can be used to estimate the most likely date on which a farm was infected and, combined with data on when herds were slaughtered, gives the most likely period over which a farm would have been infectious. These temporal data have been used previously to estimate sequences of transmission events (Haydon et al. 2003), referred to henceforth as transmission trees. While many of the overall characteristics of these transmission trees (for example estimates of the average number of susceptible farms infected by an infectious farm) are robust to variations in their precise structure, the very large numbers of different transmission trees that are consistent with these temporal data render them mostly unhelpful with respect to identifying particular transmission events with confidence.

An integrated analysis, combining both data on the timing of infection and genetic sequence data from each infected farm should improve the resolution and confidence with which virus transmission routes can be identified. It may also highlight anomalous epidemiological information, indicative of unidentified intermediate infected premises missing from chains of transmission.

This study focuses on a set of farms in County Durham that were infected early in the 2001 UK FMDV epidemic. For each of the premises included in the study we used data on the timing of infection and animal culling to establish distributions describing likely infection dates and infectious periods, together with complete genomes sequences acquired from viruses sampled from these infected premises that provided information on their relative relatedness. We have then combined the inferences from these two types of data to determine statistically a set of most likely transmission trees, which we show to be a much smaller set than obtained through analysis of either data type alone. The increasing use of genetic data for forensic epidemiological purposes will require the integration of genetic and epidemiological data, and this study proposes an initial approach to this problem.

4.3 Methods

4.3.1 Infected premises included in study

The fifteen infected premises in the Durham area (established to be FMDV positive through laboratory testing) included in this study were identified from Defra's Animal Health and Welfare FMD Data Archive (http://footandmouth.csl.gov.uk/) which also provided epidemiological data (including lesion age and date of cull). The relative locations of these premises are shown in Figure 4.1. There had been no recorded movements of animals between any of the infected premises, the movements of FMD susceptible livestock having been prohibited throughout Great Britain since 23 February 2001. The details of the clinical samples included in this study are described in Table 4.1, including source animal, farm type, and details of livestock holdings. For farm D, the epithelium was cut in half, and both halves sequenced independently to investigate the repeatability of the amplification and sequencing method used.

4.3.2 Complete genome sequencing of foot-and-mouth disease viruses

The full genome sequence of the viruses recovered from 4 of the infected premises (F, G, I, J) in this cluster had been determined previously together with 7 genomes from 5 other infected premises from the start of the outbreak included in this study (1-5) (DQ404172, DQ404173, DQ404175-DQ404180, DQ404165, DQ404166, DQ404167, DQ404169, DQ404170) (Cottam et al. 2006). Eleven more viral genomes were sequenced for this study following the method described previously (Cottam et al. 2006) (EF552688-EF552697, and EU214601, Table 4.1). Briefly, a virus suspension was made from original epithelium material, RNA was extracted using TRIzol reagent (Invitrogen), the RNA was reverse transcribed using Superscript III Reverse Transcriptase (Invitrogen), and the cDNA was cleaned up using GFX PCR DNA and gel band purification kit (Amersham Biosciences). Five overlapping PCR products were

obtained spanning the whole genome, cleaned up using the GFX PCR DNA and gel band purification kit (Amersham Biosciences), and sequenced using the Beckman CEQ8000 Coulter system. The raw data files were assembled into a contig and edited using SeqMan (DNAStar), and all further sequence manipulation was performed using Bioedit and DnaSP freeware. The genealogical relationships were based on statistical parsimony as implemented in the software package TCS (Clement et al. 2000). The tree was rooted to the closest FMDV strain SAR/19/2000 (Mason et al. 2003b).

4.3.3 Enumeration of transmission trees consistent with the sequence data

Rooting the sequence genealogy with SAR/19/2000 indicated the general direction of the transmission events (away from the node between farms 1 and 2). We took into account the known transmission history between farms 1 to 5 and restricted the tree configurations to those that included these 4 transmission events as a fixed link. Prior to the movement ban, infected animals could be transferred between farms and we allowed for the possibility of virus transmission without mutation. After the movement ban, farm-to-farm transmission likely required more infection cycles and, given the rate of molecular evolution of FMDV, the possibility of farm-to-farm transfer without mutation was considered to be negligible. Using the most parsimonious genealogy constructed by TCS, a list of plausible farms on which each putative virus haplotype could have been located was constructed, assuming no back-mutation and only a single lineage present on each farm. When this list included more than one farm, all transmission trees consistent with this ambiguity were counted. So, for example, haplotypes preceding ('upstream of') the most recent common ancestor of two farms could have arisen on either farm, whereas haplotypes proceeding ('downstream of') this common ancestor can (under reasonable assumptions) be assigned to either one farm or the other. The likelihood of each tree was estimated based on the available epidemiological data as described below.

Table 4.1 Farm and animal source details for each of the 22 FMDV consensus genomes included

Virus Sample	Accession no.	Animal *	Date of report	Date of cull	Oldest Lesion age (days)	No. and type of susceptible animals present on farm*	No. and type of infected animals reported*
1a†	DQ404179	P	24/02/2001	25/02/2001	10	571P	400P
1b†	DQ404178	P	24/02/2001	25/02/2001	10	571P	400P
1c†	DQ404177	P	24/02/2001	25/02/2001	10	571P	400P
2†	DQ404176	C	25/02/2001	25/02/2001	7	101BC, 366S	51BC
3†	DQ404175	C	24/02/2001	27/02/2001	4	600BC, 1500S	50BC, 17S
4†	DQ404173	C	26/02/2001	28/02/2001	3	558S	6S
5†	DQ404172	C	01/03/2001	03/03/2001	2	450BC	1BC
A	EF552688	S	31/03/2001	02/04/2001	2	167DC, 1600S	3S
B	EF552689	S	02/04/2001	03/04/2001	2	113BC, 130S	130S
C	EF552690	C	20/04/2001	20/04/2001	3	8BC, 400S	3BC
D	EF552691	C	11/05/2001	11/05/2001	1	188DC	2DC
E	EF552692	S	22/04/2001	24/04/2001	7	110BC, 152S	3S
F†	DQ404170	S	23/04/2001	24/04/2001	4	234BC, 340S	17S
G†	DQ404167	S	14/05/2001	14/05/2001	4	5BC, 672S, 2P	35S
I†	DQ404166	C	28/05/2001	28/05/2001	3	186BC, 3909S, 27P	2BC, 1S
J†	DQ404165	C	03/06/2001	04/06/2001	2	456DC	2DC
K	EF552693	C	01/04/2001	04/04/2001	1	215BC, 383S	1BC
L	EF552694	C	10/04/2001	10/04/2001	2	107BC, 124S	1BC
M	EF552695	S	17/05/2001	18/05/2001	5	46BC, 188S	14S
N	EF552696	C	12/04/2001	14/04/2001	1	6BC	1BC
O	EF552697	S	15/04/2001	15/04/2001	2	39DC, 47BC, 197S	3S
P	EU214601	S	30/04/2001	01/05/2001	4	88BC, 176S	3S

*Animals described by one letter coding; P Pig, S Sheep, C Cattle, DC Dairy cattle, BC Beef cattle.

†Virus genomic sequences published previously

4.3.4 Epidemiological data

The first part of the analysis required estimating a distribution describing the likelihood that an individual farm was infected on a particular date, and a distribution describing the likelihood that a farm was a source of infection on a particular date (this has been described previously as the temporal risk window; (Taylor et al. 2004; Thrusfield et al. 2005)). This analysis required two functions: $I_i(t)$, describing the probability that the ith farm was first infected at time t; and $L(k)$, the probability that the first infected individual on a given farm incubates virus for k days prior to becoming infectious (and here we assume this function to apply to all infected farms). From these two functions it is possible to estimate a further function, $F_i(t)$, describing the probability that the ith farm is a source of infection at time t.

The mean incubation period was chosen to be 5 days in common with other studies of the UK 2001 outbreak (Keeling et al. 2001), and the distribution of incubation periods, $L(k)$, was assumed to follow a discrete form of the gamma-distribution with scale and shape parameters of 3.00 and 1.67 respectively (this results in a 95% probability of incubation periods between 2 and 12 days, in accord with previous estimates (Gibbens & Wilesmith 2002). Farms were assumed to be a source of infection immediately after the incubation period and uninfectious once culled. This assumption is justified by the imposition of intense farm biosecurity, cleansing and disinfection following livestock culling.

The most likely date of infection for each farm was estimated to be the date on which disease was reported to be present on the farm less the number of days of the oldest lesion on the farm less five days for virus incubation. We represented the uncertainty around the most likely date of infection arising from (unknown) error in the lesion dating, and possible variation in the incubation period by $I_i(t)$, a discrete form of a beta distribution. Three pieces of information were used to inform the shape of $I_i(t)$: the

estimated most likely date of infection of farm *i* - which determined the mode of $I_i(t)$, the most likely infection date of the primary case (which determined the earliest possible infection time) and the reporting date of the *i*th farm (less two days to allow for a minimum incubation period), which determined the very latest possible infection time.

$F_i(t)$, the probability that the *i*th farm is infectious at time *t* was then calculated from $I_i(t)$ and $L(k)$ as follows:

$$F_i(t) = \sum_{j=0}^{C_i} \left(I_i(j) \cdot \left(\sum_{k=1}^{t-j} L(k) \right) \right)$$

where C_i was the time at which the *i*th farm was culled (time in this study is measured in days since 26th January 2001). This expression sums over the probability of farm *i* becoming infected on day *j*, *and* completing the incubation period in no more than *t-j* days. Now it is possible to calculate the likelihood that farm *j* infected farm *i*, assuming that farms cannot be multiply infected and that there are only *n* possible sources of infection:

$$\lambda_{ij} = \left(\sum_{t=0}^{\min(C_j, C_i)} I_i(t) \cdot F_j(t) \right) \Bigg/ \sum_{\substack{k=1 \\ k \neq i}}^{n} \left(\sum_{t=0}^{\min(C_j, C_k)} I_i(t) \cdot F_k(t) \right)$$

This equation sums the product of the likelihoods that farm *i* was infected, and farm *j* infectious over all possible days that transmission to *i* could have occurred. The denominator is required to ensure the λ_{ij} sum to 1. These λ_{ij} can be calculated for all possible pairs of farms, and used to compute the overall log-likelihood of any single transmission tree. We proceeded by computing the likelihoods of all transmission trees consistent with the sequence data and to identify which of these trees was the most likely based on the epidemiological data. A subset of trees comprising the top 95% of the

distribution of tree likelihoods was identified. For each transmission link in each tree in this top 95% we also computed the ratio of the likelihood of the tree to the likelihood of the tree that included the next most likely alternative source of infection for the link. This ratio reflects the confidence in the inferred source of infection relative to other possible sources.

4.3.5 Rate of nucleotide substitutions per nucleotide per day

A molecular clock was fitted to the virus sequence data from farms A-P using Markov Chain Monte Carlo techniques implemented in the software package BEAST (Bayesian evolutionary analysis sampling trees) (Drummond et al. 2002). A relaxed clock was fitted with exponentially distributed rates and exponentially increasing population size with the HKY model of base substitution (Hasegawa et al. 1985) with rate heterogeneity assumed.

4.3.6 Number of nucleotide substitutions detectable between consecutive farm infections

The distribution of the number of nucleotide substitutions detected between consecutive farm infections was estimated using the most likely transmission tree determined as described above. The most recent common ancestor of viral genotypes sampled from the source and recipient of infection was assumed to have been present on the source farm. Hence, the number of nucleotide substitutions per farm transfer was estimated as the number of nucleotide substitutions that arose between this common ancestor and the virus sampled on the recipient farm. We included 17 of the 19 transmission events comprising the most likely transmission tree in the analysis (the two transmission events of virus to farms A and K were excluded from the distribution as it is very likely that they were infected from farms outside of the study area).

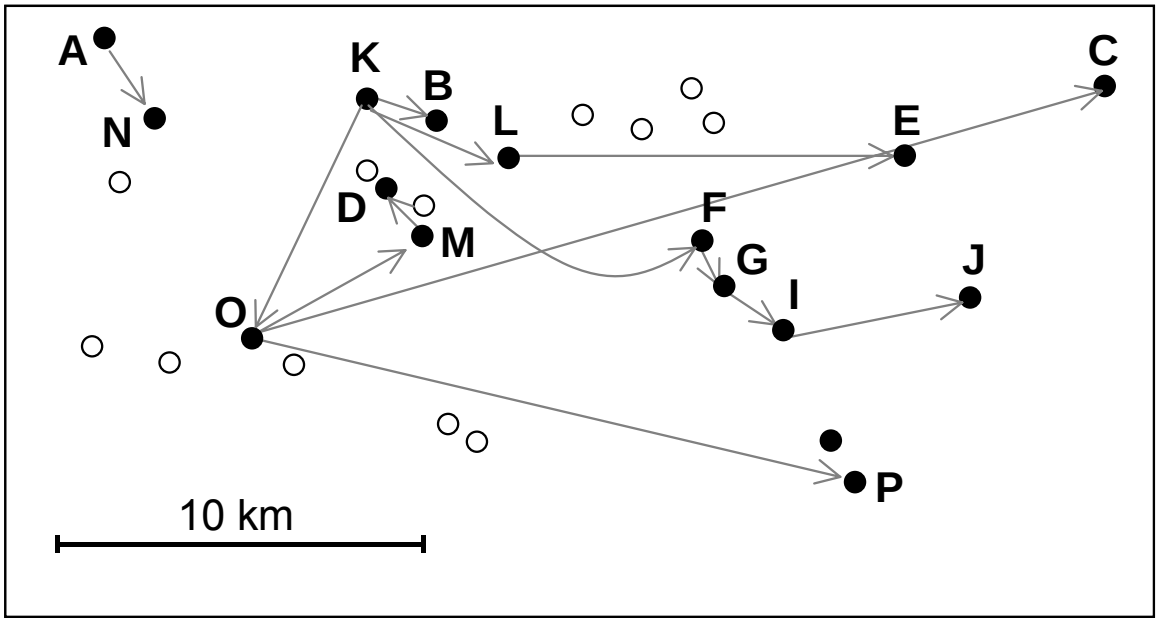


Figure 4.1

Map showing the spatial relationship of 15 infected premises confirmed by laboratory testing (filled circles) and 12 infected premises (determined by clinical observations) that were subsequently found to be negative for virus by laboratory testing (open circles). A-P indicate the infected premises from which virus has been sequenced. The direction of most likely transmission events as determined by this study are shown by the grey arrows. Scale indicated in kilometres.

4.4 Results

4.4.1 Complete genome sequencing of virus isolates and genetic transmission tracing

Epidemiological data from 20 infected premises were included in this study. For five of these infected premises the transmission tree was determined previously by tracing of direct animal movements (farms 1-5), for the other 15 the transmission routes were unknown (A-P). The 11 new sequences (A, B, C, D, E, K, L, M, N, O, and P) were between 8193 and 8195 nt in length with no ambiguous nucleotides, determined with an average of 4.8 times coverage of each nucleotide. Between all 15 sequences (A-P) there were 85 variant nucleotides, of which 24 resulted in non-synonymous changes. Five of the 85 variant nucleotides occurred within the VP1 capsid gene. The epithelium sample from farm D (which was sequenced twice independently) yielded identical sequences. A statistical parsimony tree depicting the genetic relationship between the viruses sequenced was generated for the 23 sequences, rooted using the genome sequence of the closest previously sequenced virus (from South Africa) as shown in Figure 4.2.

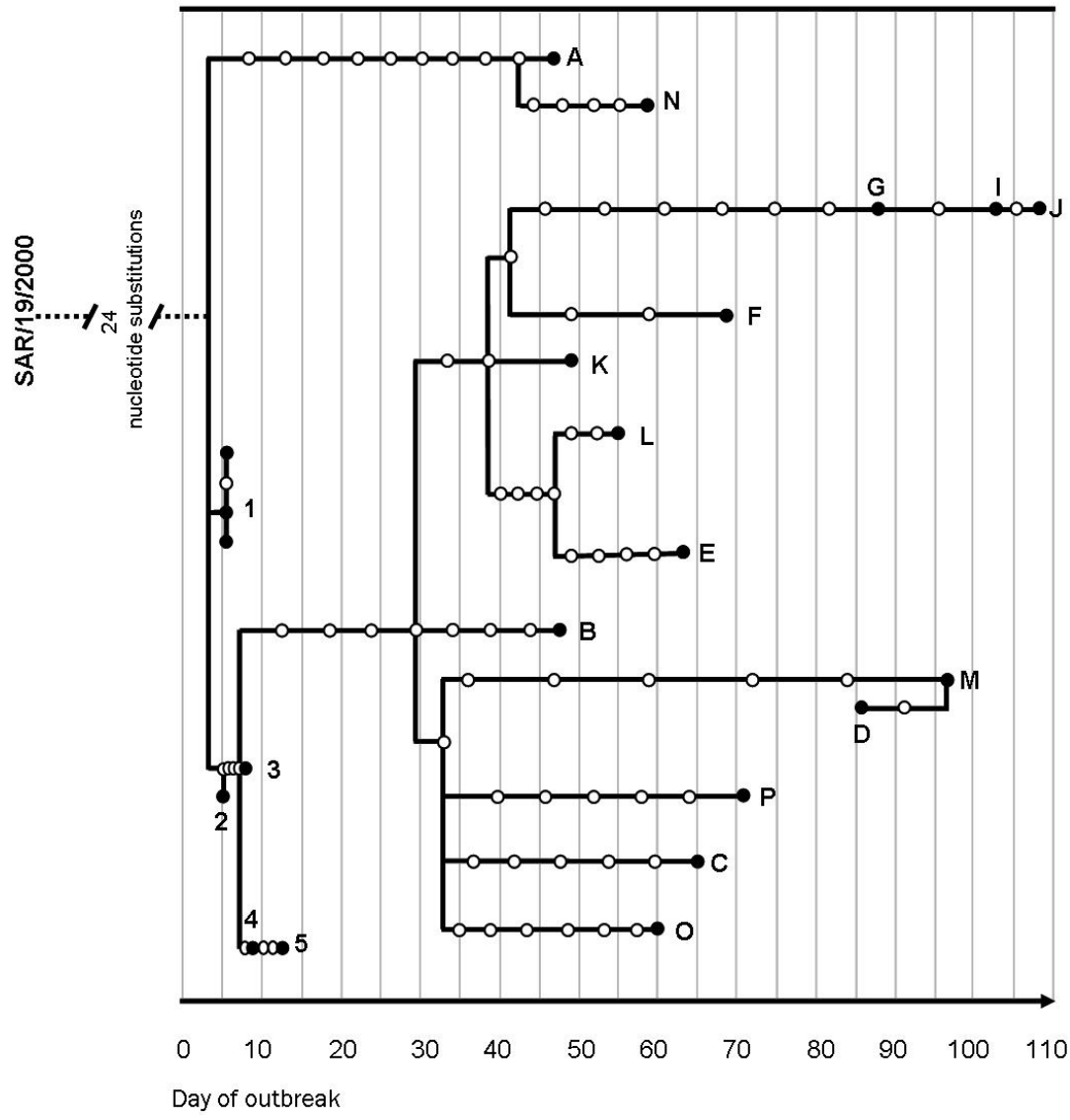


Figure 4.2

Statistical parsimony analysis of 22 United Kingdom Pan Asia O FMDV complete genome sequences rooted to the closest relative SAR/19/2000 by TCS; each connecting branch line represents a nucleotide substitution, with each dot representing a putative ancestral virus haplotype (filled dots indicate sequenced haplotypes). The farms infected through movement of infected livestock are represented by the numbers 1-5 (representing Defra infected premises numbers 4,6,7,16 and 38 respectively), and those from a cluster in Durham are represented by the letters A-P. Day zero corresponds to the 26th January 2001.

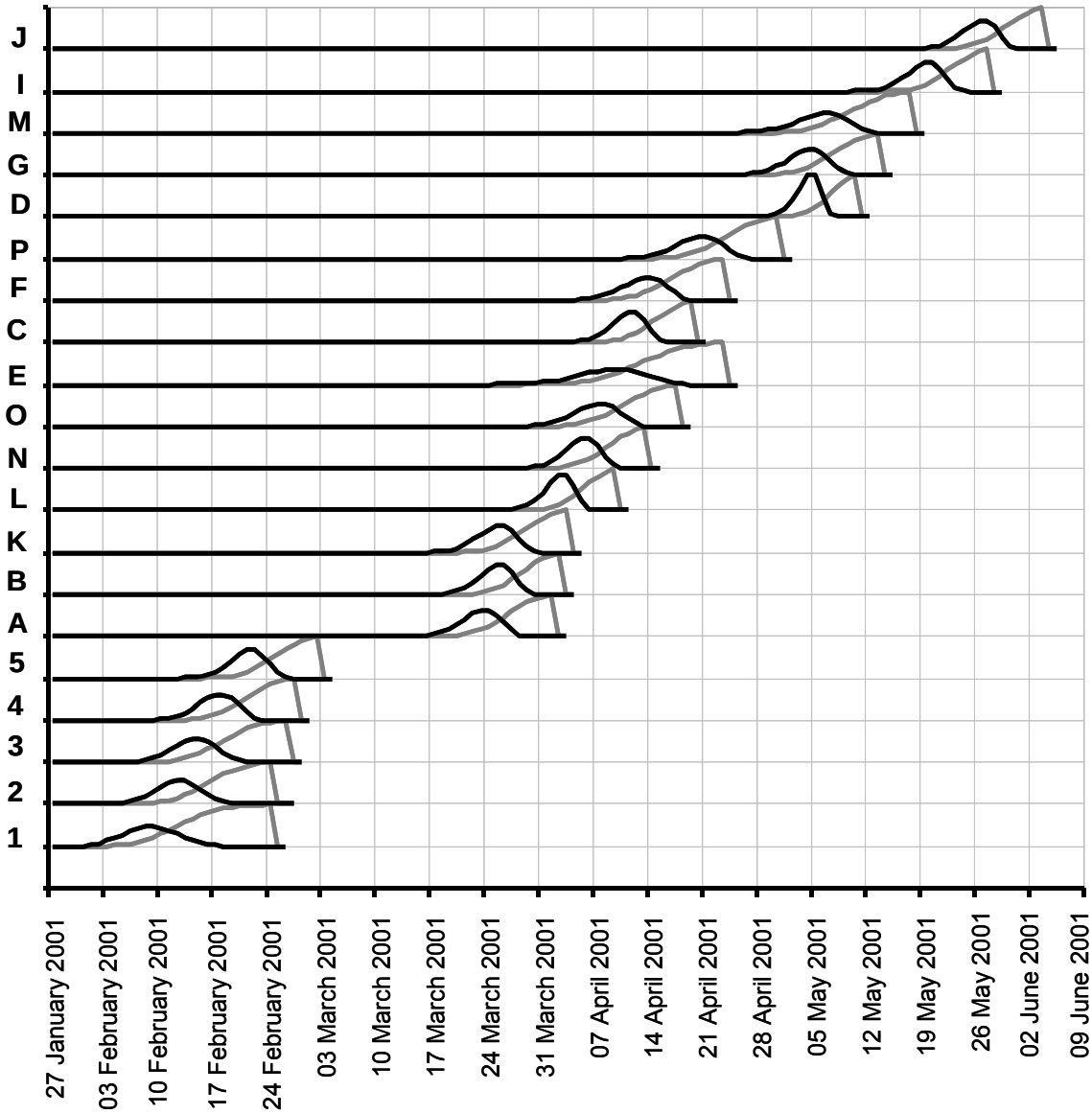


Figure 4.3

Temporal infection profiles of 20 farms infected with FMDV in 2001. Black lines indicate the likelihood of infection on any particular day, $I_i(t)$; and grey lines the likelihood that a farm was infectious on a particular date, $F_i(t)$, prior to culling.

4.4.2 Temporal infection profiles of infected premises and transmission tracing

For each farm the distributions of likely date of infection and the most likely period of infectiousness were calculated and are shown in Figure 4.3. The overlaps of infection and infectious periods can be clearly noted, providing a low-resolution impression of possible transmission pathways, together with some obvious breaks in the chain of

transmission suggesting the presence of additional infected farms or inaccuracies in the estimates of time of infection.

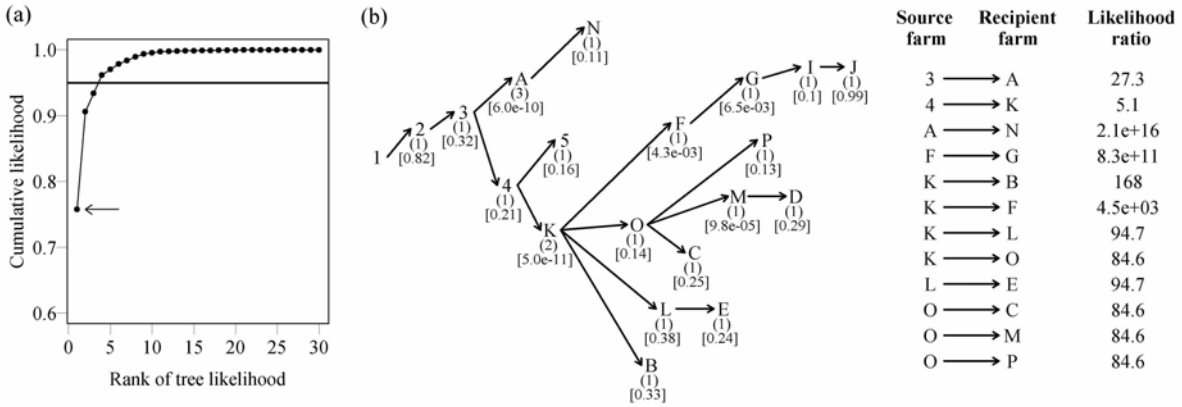


Figure 4.4

Most likely transmission trees for the sequenced farms. (a) Cumulative likelihood distribution of the transmission trees, where the four most likely trees comprise more than 95% of the sum of all tree likelihoods (arrow: the most likely tree). (b) Most likely tree, with for each farm the number of different sources of infection among the top 95% set of trees (in parentheses), and λ_{ij}, the likelihood that the recipient farm was infected by the indicated source farm (in square brackets). The legend indicates the likelihood ratio between the most likely tree and the next most likely tree with a different source farm for a given infected farm, when the source of infection is not assumed.

4.4.3 Likelihood analysis of integrated epidemiological and genetic data sets

With no knowledge of any transmission routes the number of alternative trees consistent with the genetic data is 41,472. This number reduces to 1,728 if the known transmission links between farms 1-5 are fixed. Four of these 1,728 trees account for 95% of the likelihood (see Figure 4.4a). For each farm, a maximum of 3 alternative sources of infection can be identified among these 4 most likely trees, but the most likely source of infection is always at least 80 times more likely than any other source (except for Farms A for which farm 3 is the alternative source and K for which farms 1 and 2 are alternative sources, Figure 4.4b).

4.4.4 Rate of nucleotide substitutions per nucleotide per day

A relaxed molecular clock for the rate of substitution of all nucleotide changes from farms A-P was estimated to be 2.76 x 10^{-5}/site/day (95% CIs 1.00 x 10^{-5} to 4.37 x 10^{-5}).

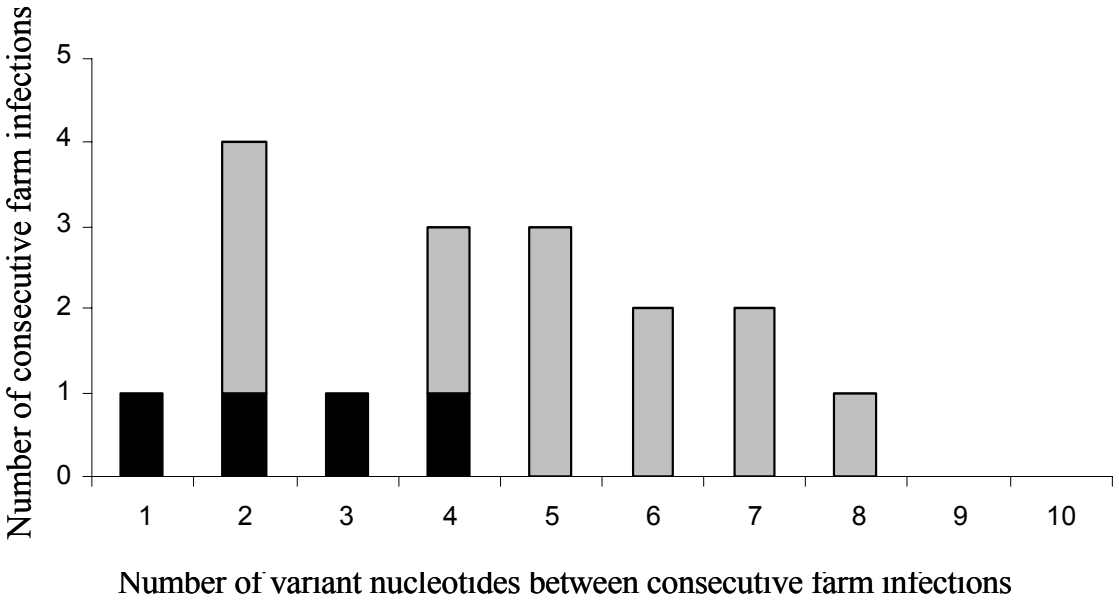


Figure 4.5

Distribution of the number of variant nucleotides between viruses recovered from consecutively infected farms). Those that represent transmissions that occurred before the NMB are shown in black. The number of variant nucleotides was determined from the common ancestor of source and daughter farm, as described in material and methods, using the transmission events shown in Figure 4.4b. The distribution has a mean of 4.3, SD=2.1.

4.4.5 Number of nucleotide changes upon farm-to-farm transfer

The distribution of the number of nucleotide changes per farm transfer is shown in Figure 4.5. The distribution has a mean of 4.3 nucleotide substitutions, with a standard deviation of 2.1. If this distribution is partitioned into transmission events pre and proceeding the NMB then the mean number of substitutions per transmission link

proceeding the NMB (farms A-P) is 4.85 and is significantly higher than that preceding the NMB, represented by farms 1-5, of 2.5 substitutions (t test; t=2.19, *P*=0.045).

4.5 Discussion

The method described in this study makes a preliminary attempt at integrating information from genetic data with the relative timing of infection events to compute the most likely transmission tree that reflects the spread of infection between farms. The study has demonstrated the power of forensic genetic tracing for FMDV, and highlights the future challenges in analyses of this sort.

Using only the known transmission events relating farms 1-5, 1,728 possible transmission trees remain consistent with the rooted parsimony tree identified using the genetic data. This is because while it is possible to identify viruses that share a common ancestor, it is often not possible to identify the farm on which the ancestors were present, and therefore the exact direction of transmission events. However, the difficulties in interpreting the genetic data can be overcome in part by the addition of epidemiological data: here, the most likely time of infection and infectiousness of a farm, which enables the likelihood of particular transmission events to be estimated. This integration of information from the two data sets enables identification of the most likely trees at a resolution far greater than is possible using either data set alone.

In this study we have generated complete genome sequences from all infected farms from a geographic area of approximately 100 km^2 found to be FMDV positive by laboratory analysis. Interpretation of the results is complicated because both the epidemiological timing data and the genetic data suggest that there may have been more than a single disease introduction event in this area.

Our analysis indicates particularly low probabilities of infection between some pairs of infected farms linked in the most likely transmission tree (see Figure 4.4b). These links include those between farms 3 to A, 4 to K, O to M, K to F, and F to G. In some cases

we suspect that additional intermediary infectious farms must exist that are either: 1) outside of the geographic area of our study, or 2) located within the study area but from which infected tissue samples were not collected (within the study area there were 108 farms identified as 'dangerous contacts' or 'contiguous premises' and on which livestock were culled prior to the 28th May, but even if these farms were infected, they may not have been very infectious). Based on the epidemiological activity to the north of our study area throughout March, we believe the former may be a reasonable explanation for how farms A and K became infected.

Explanations for the source of infections to farms M, F and G are more problematic as there are almost no alternative plausible local sources of infection. Furthermore, the possibility of airborne (and potentially longer-range) spread occurring between farms C, E, F, G, I, and J was considered improbable on the basis of wind strength and direction in relation to potential sources of infection, and is not thought to have played a role in dissemination of virus to these premises. This forces us to question the accuracy of the data relating to timing of infection. Lesion dating is imperfect with an unknown amount of error associated with it. However, the dominant source of error associated with using lesions to estimate the date of infection is that the oldest lesions may be overlooked entirely (and it is for these reasons that the function describing the likelihood that a farm was infected on a particular day, $I_i(t)$, was only minimally constrained). This is particularly likely on farms with livestock that include sheep where the clinical symptoms of disease are often mild. Farms M, F, and G all maintained substantial sheep herds in which disease was eventually confirmed (Table 4.1) and it is possible that infection was present on these farms considerably earlier than estimated, but went unreported.

From the data generated during this study we can estimate the distribution of the number of nucleotide changes that arose between consensus sequences recovered from source and recipient infected farms. When this number is unusually large it suggests that virus

has been replicating in some other livestock population, either on unidentified intermediate farms, or on the recipient farm but in a population in which disease had previously been overlooked. For example, the large number of changes observed on farm A (more than double the average expected for a single farm-farm transmission event) is suggestive of such an intermediate. It is important to recognize that the number of changes detectable between consecutively infected farms will likely depend on the mode of infection. In this study, the number of nucleotide substitutions per infection generation interval is lower prior to the imposition of the national movement ban, compared to after, probably because there are less viral replication cycles associated with infections arising as a result of movement of infected animals, compared to fomite associated transmission. Data of this sort could be used in future investigations to infer the presence of undetected sources of infection, or to determine the most likely number of infected premises in a transmission chain. However, for this to be feasible it is important that this distribution is characterized in more detail.

The most likely transmission tree shown in Figure 4.4b raises some interesting points relating to this particular case study. This tree differs from the original contact tracing tree proposed by Defra (which is inconsistent with the genetic data). First, the separate lineage of infected premises represented by A and N is anomalous, and when compared to the remainder of the sequences available from the 2001 UK outbreak (Cottam et al. 2006), represents transmission of the virus directly into the area from the source of the outbreak, and not via Longtown and Hexham markets (as is thought to be the case for all other infected premises throughout the UK (Gibbens et al. 2001)). This finding suggests further study of the outbreak is necessary to determine the origins of this previously unidentified chain of transmission events.

The rate of nucleotide substitution noted in this study for virus from farms A-P over this short time period is 2.76×10^{-5} (95% CIs 1.00×10^{-5} to 4.37×10^{-5}) which is very close to

that estimated previously (2.26×10^{-5}, 95% CIs 1.75×10^{-5} to 2.80×10^{-5}) for the whole of the UK 2001 outbreak (Cottam et al. 2006). In principle, it would be possible to evaluate the likelihood of the transmission trees based on the genetic data, after the imposition of a molecular clock. However, until more is known about the extent of nucleotide variation from single animals, herds, and different types of livestock on a single farm, we are inclined to view a more detailed and complex quantitative analysis of 'one-genotype per farm' data with some caution.

Here we have presented a simple method by which the transmission tree-space is qualitatively restricted through the use of the genetic data, and the likelihood of trees remaining in this space quantitatively evaluated by the use of epidemiological data. This likelihood, estimated from the epidemiological data could be made more sophisticated by accounting for the mode of transmission (for example incorporating information about the possible movement of animals between farms), or the livestock composition on particular farms (for example, pigs are generally regarded as more infectious than cattle or sheep, whereas cattle are more susceptible than sheep or pigs (Alexandersen et al. 2003b; Alexandersen et al. 2003c; Alexandersen et al. 2002)). Farm infectiousness is not explicitly quantified in this analysis. While infectiousness is unlikely to remain constant over the course of an infection, available data do not enable a parameterization of possible change (Savill et al. 2007). Here we have simply assumed that the probability a farm was a source of infection varies over time in a way dictated by the timing of infection. It is plausible that susceptibility varies between farms, and if this were quantified it could also be incorporated into future models.

Ultimately, the analysis of these sorts of data will be best conducted by development of a single likelihood model that accounts for both the evolutionary and epidemiological dynamics at the various different scales at which they occur. Developing models that

integrate different data types in this way will be a difficult but exciting future research challenge.

Chapter 5

Analysis of Foot-and-mouth disease virus nucleotide sequence variation within naturally infected epithelium

5.1 Summary

The frequency of mutants within a virus population will affect the rate at which mutations are fixed in virus populations. Successful replication and the ability to escape from host immune responses are likely to shape the frequency and distribution of mutants within a population. Additionally, the mutant frequency and distribution could differ depending on the host animal, and size and age of the virus population. In this study, genetic variation within Foot-and-mouth disease virus (FMDV) populations in lesions from naturally infected animals was investigated, through the analysis of 41 complete capsid sequences. Comparisons were made between virus sequences obtained from a 3 day old cow lesion and a 1 day old sheep lesion; investigating evolutionary rates, mutation type, frequency and distribution. The overall mutant frequency in this study was 3.06 x 10^{-4} mutations per nucleotide sequenced. The virus sequences obtained from the cow lesion had a high *dN/dS* ratio of 1.228, whereas the virus sequences from the sheep lesion had a lower *dN/dS* ratio of 0.187. Changes at important neutralising antigenic sites were found within the virus sequences from the cow lesion. Also, an example was found of previously dominant sequences being present as a minor population. The effect of two reverse transcriptase enzymes on the mutant frequency and distribution observed was compared, with no significant difference found between the two sequence sets. This is the first investigation of genetic variation within a single virus population from a naturally infected host, and the findings encourage further research in this area.

5.2 Introduction

Genetic variability of FMDV populations within an animal have yet to be determined, although current reports of the mutation rate of the virus suggest it should be high (Carrillo et al. 1998; Domingo et al. 2001; Haydon et al. 2001). The genetic composition of *in vitro* cell culture populations has been extensively studied for serotype C (Airaksinen et al. 2003; Arias et al. 2001; Pariente et al. 2001; Sierra et al. 2000) and once for serotype O (Gu et al. 2006). However, it is not known how well these studies represent what occurs within an infected animal where the virus could be subject to different selection pressures and constraints.

In Chapters 3 and 4, it was suggested that nucleotide changes become fixed within the virus consensus sequence due to the transmission of the virus through 'bottlenecks' leading to a founder effect in the new population. If this is true, then the mutant diversity that surrounds the consensus would be a determinant of the rate and manner in which the virus consensus sequence evolves. Thus, it is imperative to study the extent and characteristics of the sequence diversity surrounding the consensus sequence, in order to further our understanding of the genetic evolution of the virus in the field.

The most widely used method to investigate genetic variation within a virus population is molecular cloning of PCR products and nucleotide sequencing of clones produced by *E.coli* from the PCR fragments. Although this method is sensitive to artificial mutations introduced by the various amplification enzymes, when high fidelity enzymes are used it can provide an estimate of the upper limit of genetic sequence variability within a sample. This method has been compared previously to 'biological' cloning, which involves plaquing of individual viruses and sequencing of plaques, and was found to give statistically indistinguishable results (Arias et al. 2001).

Previous studies have investigated virus variation within a host for several different RNA viruses, demonstrating that the extent and character of the swarm of genetically variant viruses within a population is often highly variable between hosts, and at different disease stages within the same hosts, and can also directly correlate with specific virus pathogenicity. For example, a study of Hepatitis C virus (HCV) noted that an increase in the genetic diversity within a virus population was associated with progressive hepatitis (Farci et al. 2000). Another study of HCV suggested that genetic bottlenecking and hence a reduction in virus intra-population diversity may have assisted patient recovery and resistance to the persistent state that the virus can cause (Quer et al. 2005). Also, in studying HCV, it was found that the pattern of within host evolution could differ between hosts, and that non-synonymous variation could depend on the stage of infection (Alfonso et al. 2005). The importance of bottleneck events and selection and hence evolution within hosts of a virus population was again emphasised in a combined study of HCV and HIV (Poon et al. 2007), which found that circulating virus is shaped by the transmission rate and the immune diversity of the population. The extent to which a bottleneck can affect virus diversity during HIV-1 transmission was shown in a 99% depletion in diversity estimated using coalescent theory (Edwards et al. 2006), which has important implications for virus diversity and evolution. Such extreme bottlenecking would require close to all virus variants to be fully viable within a population, and also from a diverse population will encourage the fixation of mutation and hence virus evolution. Another virus for which intra-host genetic variation of the virus population has been studied is West Nile Virus, which is transmitted by a mosquito vector. It was found that the virus showed different levels of diversity within either the vector or the main animal host, and that an increase in diversity could lead to a decrease in pathogenicity (Jerzak et al. 2007). This phenomenon of differences between viruses and between host environments affecting viral diversity has also been shown to be true of plant viruses (Schneider & Roossinck 2000; Schneider & Roossinck 2001).

The majority of the above studies have investigated viruses with long infectious periods and the relevance to FMDV, which causes an acute disease over a short infection period, is not known. However, studies on Poliovirus have shown that a variation in the fidelity of the virus polymerase, and hence a change in genetic diversity of the resulting population, can have a profound affect on the pathogenicity of the virus (Pfeiffer & Kirkegaard 2005; Vignuzzi et al. 2006). Previous studies on FMDV have shown that genetic evolution of the consensus virus sequence can occur during persistent infection of the virus within a host (Vosloo et al. 1996). Also, although no changes were found within the consensus sequence of the VP1 protein, antigenic variants arose during acute infection of pigs, detected *in vitro* as antibody escapes (Carrillo et al. 1990). A further study noted that a mixed population is maintained within a host and also can be seen in secondary lesions (i.e. not at the point of inoculation) suggesting that more than one virus will initiate a lesion (Carrillo et al. 1998).

As has been shown for other RNA viruses it is feasible that the within population variation of FMDV differs depending upon the host environment and stage of disease, and may also be affected by the rate of transmission events and the bottlenecks that arise as a result. For this reason within host genetic variation and evolution of FMDV needs to be studied further to assess its potential to improve our understanding of the fixation of mutations into the consensus sequence, variation in disease severity between hosts, the progression to persistence (a previous study has shown that genetic change is a factor in persistence (Horsington & Zhang 2007)), and the probability of immune and vaccine escape by the virus.

The aim of this study was to investigate further the nature of genetic evolution of FMDV at the level of a single population within a single lesion of a naturally infected host. An increased knowledge of genetic variability within a virus population could potentially assist estimates of population genetic parameters such as mutation, replication,

recombination and selection. It is the interaction of these parameters that determines the population genetic dynamics illustrated by the mutant frequency within a population, and also determines the rate at which the virus changes genetically in the field.

Within this study the performances of two different reverse transcriptase enzymes were compared by testing specific hypotheses concerning evolutionary rates and *dN/dS* ratios between the two populations of clones (generated from the same starting epithelium material) using likelihood ratio tests. Furthermore, the genetic variation surrounding the virus consensus sequence within the capsid region upon natural infection of both a cow and a sheep (infected during the 2001 UK FMD epidemic) was investigated, providing an estimate of mutant frequency. The cloned sequences have enabled sophisticated statistical examination of the phylogenetic relationships within two separate virus populations, with respect to *dN/dS* ratios. Additionally the distribution of mutations observed was analysed, with a significantly high number of mutations occurring within antigenic sites within one population.

5.3 Methods

5.3.1 Virus samples

The virus populations investigated were from epithelium originating from a cow UKG 14524/2001 from Infected Premises (IP) 1970, and a sheep UKG 15101/2001 from IP2027. The consensus sequence of the complete genome for both had been determined previously (DQ404160 and DQ404158 respectively). The two IP were situated in Northumberland and infected close to the end of the epidemic. At the time of infection IP1970 held 200 sheep and 62 beef cattle. Three cattle were reported to have been infected, with the oldest lesion being 3 days. The farm was examined on 23rd August 2001 and culled the following day. A sample of tongue epithelium was sent to IAH Pirbright. IP2027 was also a mixed holding with 1341 sheep and 172 beef cattle. The farm was examined on the 24th September 2001 and culled the same day, with 9 sheep showing lesions representative of FMD. Epithelium from a 24 hour old lesion on the dental pad was sent to IAH Pirbright.

5.3.2 Original epithelium sample suspension

The epithelium samples were ground using a pestle and mortar and a 10% suspension in phosphate buffer was made as described in Chapter 2, section 2.3.1a.

5.3.3 TRIzol RNA extraction

Two hundred µl of the epithelial tissue suspension was added to 1ml TRIzol (Invitrogen) and the RNA extracted as described previously (Reid et al. 1998). The RNA was re-suspended in 12µl nuclease free water.

5.3.4 Superscript III reverse transcription

Three µl of 10mM UKFMD/Rev 6 primer and 2µl 10mM dNTP were added to the 12µl RNA suspension and incubated at 68^{o}C for 3 minutes and then on ice for 3 minutes. Twenty one µl of freshly prepared RT mix (4µl 10x RT Buffer (Invitrogen), 8µl 25mM MgCl2, 4µl 0.1M Dithiothreitol (DTT), 2µl RNase OUT (40 units/µl, Invitrogen), 3µl nuclease free water) were added to the sample followed by 2µl Superscript III reverse transcriptase (200 units/µl, Invitrogen). The sample was then incubated at 42°C for 4hrs, after which the reaction was terminated by incubation at 85°C for 5 minutes.

5.3.5 Accuscript High Fidelity reverse transcription

Three µl of 10mM UKFMD/Rev 6 primer (see Table 2.1, Chapter 2), and 2µl 10mM dNTP were added to the 12µl RNA suspension and incubated at 68°C for 3 minutes and then on ice for 3 minutes. Twenty one µl of freshly prepared RT mix (4µl 10x RT Buffer (Stratagene), 4µl 0.1M Dithiothreitol (DTT), 2µl RNase OUT (40 units/µl, Invitrogen), 11µl nuclease free water) were added to the sample followed by 2µl Accuscript High Fidelity reverse transcriptase (Stratagene). The sample was then incubated at 42°C for 4hrs, after which the reaction was terminated by incubation at 85°C for 5 minutes.

5.3.6 Amplification of complete capsid sequence by PCR

The cDNA was cleaned using GFX PCR DNA and gel band purification kit (Amersham Biosciences) to remove residual UKFMD/Rev 6 primer, eluting the cDNA in 35µl of elution buffer (10mM Tris·Cl, pH 8.5: QIAGEN). The complete capsid PCR reaction contained 5µl 10x Hi-Fi buffer, 2µl 50mM $MgSO_4$, 1µl 10mM dNTP, 1µl 10mM UKFMD/For3, 1µl UKFMD/Rev 7 (see Table 2.1, Chapter 2), 0.2µl Platinum Taq Hi-fidelity (5units/µl, Invitrogen), 35.8µl nuclease free water and 4µl cDNA. Thermal cycling was as follows; initial denaturation at 94°C for 2 minutes followed by 39 cycles of 94°C for 30 seconds, 68^{o}C for 30 seconds, and 72°C for 3 minutes, ending with incubation at

72°C for 7 minutes. The PCR product (2588 bp) was then cleaned using GFX PCR DNA and gel band purification kit (Amersham Biosciences).

5.3.7 'A' Tailing of blunt ended PCR product

Fourteen µl of tailing reaction mix (2µl 10x PCR buffer (Promega), 2µl 25mM MgCl2, 2µl 0.2mM dATP, 2µl Taq DNA polymerase (5 units/µl, Promega), 6µl nuclease free water) was added to 6µl of cleaned PCR product. The reaction was incubated for 30 minutes at 70^{o}C.

5.3.8 pGEMT-Easy Vector Ligation

One µl of 10x rapid ligation buffer (Promega), 1µl pGEMT-easy vector (50ng), 1µl T4 ligase (3 units/µl, Promega) and 4µl of nuclease free water were added to 3µl of the tailed PCR product. The ligation reaction was incubated overnight at 4^{o}C.

5.3.9 Transformation of competent DH5α E. coli cells and selection of Ampicillin resistant E.coli with dysfunctional galactosidase

Five hundred µl of chemically competent DH5α *E. coli* cells (Invitrogen) were thawed on wet ice. One hundred µl of cells were gently added directly to the appropriate ligation reaction. The reaction was incubated on ice for 30 minutes, followed by heat shock treatment in a 37^{o}C water bath for exactly 45 seconds followed by 2 minute incubation on ice. One ml of SOC media (bacterial growth medium) was added to the cells, before incubating at 37^{o}C for 1hour with shaking at 225rpm. Two hundred µl of transformed cells were spread onto petri dishes containing LB agar with 100µg/ml Ampicillin and 90µl X-gal. The plates were inverted and incubated over night at 37°C.

5.3.10 Picking and growing of E. coli colonies

White colonies were picked with a sterile pipette tip and dropped into 12ml of LB broth containing Ampicillin 100µg/ml and incubated overnight in the shaking incubator at 37°C at 225rpm.

5.3.11 Qiagen mini-prep plasmid extraction

The plasmids were extracted from the *E. coli* cells using the Qiagen mini-prep kit following the manufacturer's instructions and eluting in 50µl.

5.3.12 Restriction digest using EcoR1

Two µl of the eluted plasmid was added to 18ul of fresh restriction digest reaction mix (2µl 10x restriction digest buffer H (Promega), 0.5µl EcoR1 (12 units/µl, Promega), 15.5µl ddH_2O), and incubated at 37°C for 30 minutes.

5.3.13 Sequencing of plasmid clones in Beckman Coulter sequencing plates

One µl of plasmid (approximately 500ng) followed by 7µl of nuclease free H_2O was added to each well of a Beckman Coulter sequencing plate. The plasmids were then denatured using a thermal-cycler to heat the samples to 94°C for 2 minutes before returning to room temperature. Four µl of 1mM sequencing primer was then added to each reaction followed by 8µl of the DTCS Quick Start Master Mix (Beckman Coulter). The plate was run on a programme of 30 cycles of 96°C for 20 seconds, 50°C for 20 seconds, and 60°C for 4 minutes. Following thermal-cycling, the reactions were ethanol precipitated before running on the Beckman Coulter Sequencing (CEQ800) machine.

5.3.14 Data analysis

The raw .scf data files were assembled using the software package DNASTAR, and all further sequence manipulation was performed using Bioedit. The location and number of substitutions were determined using DnaSP (Rozas & Rozas 1999). The minimum mutation frequency was calculated as the number of unique substitutions/total number of nucleotides sequenced. The Pi (π) statistic, defined as the average number of nucleotide differences per site between any two nucleotide sequences chosen randomly from the sample population (Nei & Li 1979), was calculated using DNAsp. Maximum likelihood trees were determined using Phylip version 3.66 (Felsenstein 2004), assuming a constant rate among sites. Likelihoods of these trees were compared using models of nucleotide substitution that differed in their assumptions regarding evolutionary rates and *dN/dS* ratios between branches using the codeml programme within the PAML3.15 freeware package (Yang 1997). Codeml was set to run assuming the standard genetic code, with the equilibrium codon frequencies in the codon substitution model calculated from the average nucleotide frequencies at the three codon positions, and with no variation in *dN/dS* ratios between sites. Estimated model likelihoods were compared by performing likelihood ratio tests. When comparing two evolutionary models, the simpler (null) model, having p_0 parameters and associated log-likelihood l_0, and the more complicated (alternative) model having p_1 parameters,with associated log-likelihood l_1 twice the log-likelihood difference $2\Delta l$ =2(l_1-l_0), has a χ^2 distribution with degrees of freedom (d.f.)= p_1-p_0 if the null model is true. Thus the test statistic $2\Delta l$ was compared to the χ^2 distribution with appropriate d.f. to test whether the more complex model is supported over the simpler model.

Possible differences in the two sets of sequences resulting from the different reverse transcriptase enzymes (Superscript III and Accuscript) were investigated to assess the enzymes' performance, by comparing nested models using likelihood ratio tests. Firstly comparing evolutionary rates between the two sequence sets, by including sequences

from these two populations in the same alignment and constructing a single tree and comparing the model likelihood of the null hypothesis that all branches in the tree are restricted to the same rate, with the alternative model likelihood of individual rates for the two virus populations, with the *dN/dS* ratio restricted to one rate for the entire tree. Secondly *dN/dS* ratios were compared, evaluating the model likelihoods of one *dN/dS* ratio for the entire tree, or two different *dN/dS* ratios, one for each population. The hypothesis that the *dN/dS* ratio of virus population UKG 14524/2001 differed from that of UKG 15101/2001 was then tested using the same method as detailed above. To test the hypothesis that the distribution of mutations within the two populations UKG 14524/2001 and UKG 15101/2001 differed, the distribution of mutations across the clones in both virus populations were compared to a Poisson distribution using the chi-squared test with 100,000 Monte Carlo bootstrap replicates. The exact binomial test was used to determine whether or not a significantly high number of mutations had occurred within sites of known antigenic importance.

5.4 Results

5.4.1 Overview

Complete capsids from 26 individual clones from UKG 14524/2001 (13 from cDNA synthesised by Superscript III reverse transcriptase and 13 from cDNA synthesised by Accuscript reverse transcriptase) and 15 individual clones from UKG15101/2001 were sequenced. Twenty nine unique variant nucleotides, an insertion of a single nucleotide, and a deletion of 31 nucleotides were found following comparison of the cloned sequences to their respective consensus sequences. The substitution mutations are listed in Table 5.1. The sequences from UKG 14524/2001 have a minimum mutation frequency of 2.79×10^{-4} mutations per nucleotide sequenced, and the sequences from UKG 15101/2001 have a minimum mutation frequency of 3.94×10^{-4} mutations per nucleotide sequenced. Properties of the two sets of clone sequences are summarised in Table 5.2.

5.4.2 Maximum likelihood phylogenetic analysis of sequences

Maximum likelihood phylogenies were constructed for the sequences representing the clonal population of UKG 14524/2001, UKG 15101/2001, and both sequence sets combined, and are shown in Figure 5.1 a,b,c respectively. The two phylogenies from the different animal lesions vary in shape, with that representing UKG 14524/2001 containing more highly divergent sequences.

5.4.3 Comparison of cloned sequences generated by the two reverse transcriptase enzymes

To determine whether or not the performances of the two reverse transcriptase enzymes differed, the model likelihood of a single evolutionary rate for all of the sequences from virus population UKG 14524/2001 was compared to the likelihood of separate rates for

the sequences generated by Supercript III and those generated by Accuscript. No significant difference was found between the two groups (*P*=0.559, χ^2=0.341, d.f.=1). Additionally the model likelihood of separate *dN/dS* ratios for the two sequence sets was not supported over the model likelihood of one *dN/dS* ratio for all (*P*=0.863, χ^2=0.03, d.f.=1). Thus the clones generated by the two enzymes did not differ significantly in number or type of mutations.

5.4.4 dN/dS and Ts/Tv ratios between two virus populations

The *dN/dS* ratio per nucleotide site for the cloned population representing UKG 14524/2001 was 1.228. For UKG 15101/2001 it was 0.187. The results are summarised in Table 5.2. Although these appear quite different, maximum likelihood analysis did not significantly support a model of 2 separate dN/dS ratios in preference to a single ratio model for all (*P*=0.097, χ^2=1.572, d.f.=1). The transition to transversion ratio for the two sets of sequences UKG 14524/2001 and UKG 15101/2001 was 28.53 and 8.41 respectively. No statistical difference in terms of within population genetic diversity was detected between the two virus populations.

Table 5.1 Variant nucleotides observed in all sequences

Clone	*Capsid protein*	*Nucleotide position*	*Nucleotide change*	*Amino acid change*
UKG 14524 S 09	**VP1**	**2017**	**G-A**	**V-I**
UKG 14524 S 10	VP3	1168	T-C	S-P
UKG 14524 S 11	VP3	1053	T-C	-
UKG 14524 S 12	VP2	551	C-T	A-V
UKG 14524 S 12	VP4	156	G-T	-
UKG 14524 S 13	**VP3**	**1075**	**C-T**	**H-Y**
UKG 14524 S 13	VP1	1588	T-C	S-P
UKG 14524 A 08	VP3	1462	T-C	-
UKG 14524 A 09	VP2	350	T-C	V-A
UKG 14524 A 10	**VP2**	**650**	**T-C**	**I-T**
UKG 14524 A 11	VP2	575	T-C	V-A
UKG 14524 A 11	*VP3*	*1080*	*C-T*	-
UKG 14524 A 12	*VP3*	*1080*	*C-T*	-
UKG 14524 A 12	VP3	1530	A-G	E-G
UKG 14524 A 12	VP3	1567	C-T	STOP
UKG 14524 A 13	VP2	308	G-A	R-H
UKG 14524 A 13	*VP3*	*1080*	*C-T*	-
UKG 14524 A 13	VP1	1970	G-A	C-Y
UKG 15101 S 05	VP3	1098	C-T	-
UKG 15101 S 06	VP1	1929	G-T	-
UKG 15101 S 07	VP2	654	C-T	-
UKG 15101 S 08	VP3	1491	T-G	-
UKG 15101 S 09	VP3	1518	T-C	-
UKG 15101 S 10	VP3	1262	A-G	K-R
UKG 15101 S 11	VP2	385	A-G	S-G
UKG 15101 S 12	VP2	606	C-T	-
UKG 15101 S 13	VP1	1668	G-A	-
UKG 15101 S 14	VP1	2136	G-A	-
UKG 15101 S 15	VP4	240	C-T	-
UKG 15101 S 15	VP3	1352	A-G	D-G
UKG 15101 S 15	VP1	1741	G-A	A-T

Shown in **bold** are non-synonymous changes in antigenic sites, and in *italics* are those changes that occurred in more than one sequence.

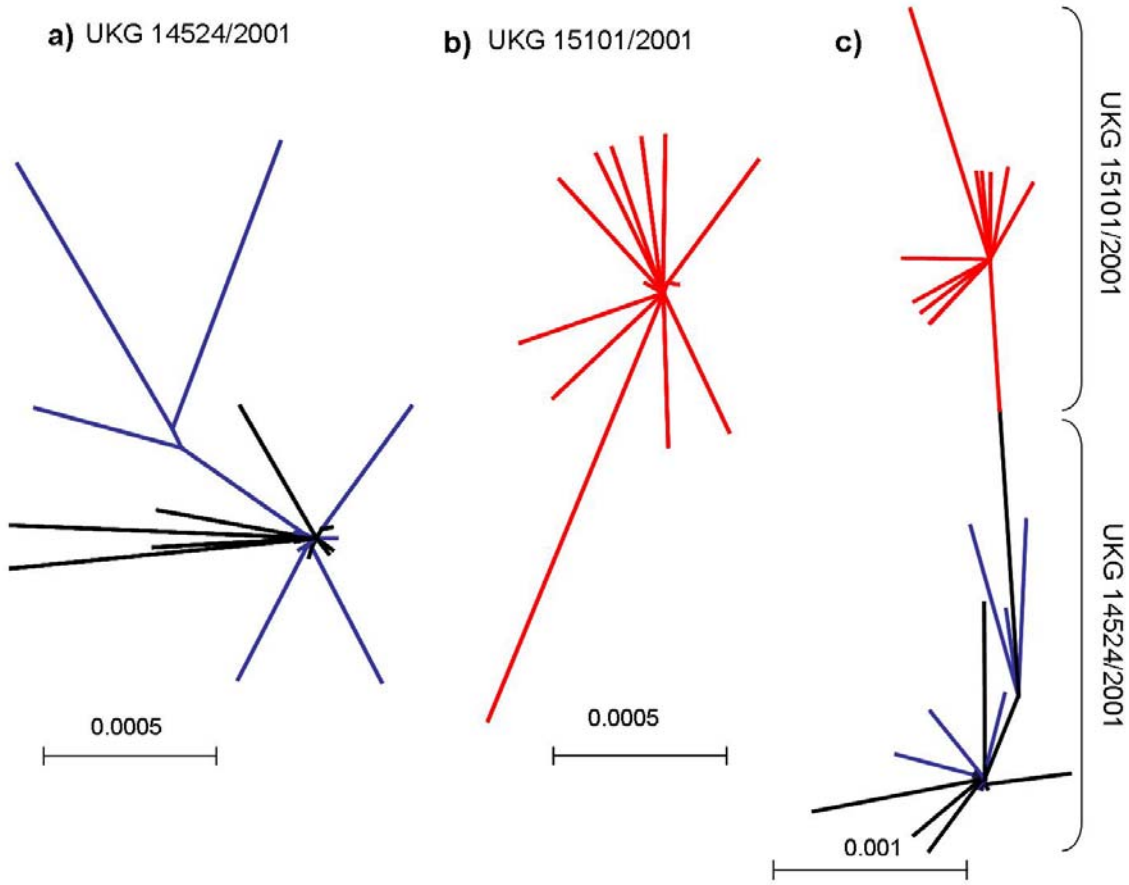


Figure 5.1

Maximum likelihood analysis of complete capsid sequences from UKG 14524/2001 and UKG 15101/2001. a) Maximum likelihood analysis of 26 complete capsid sequences representing sample UKG 14524/2001; in blue are the sequences obtained from cDNA transcribed by Accuscript reverse transcriptase, in black are sequences obtained from cDNA transcribed by Superscript III reverse transcriptase, b) Maximum likelihood analysis of 15 complete capsid sequences representing sample UKG 15101/2001, c) Maximum likelihood analysis of all 41 capsid sequences from UKG 14524/2001 (blue and black) and UKG 15101/2001 (red).

Table 5.2 Genetic characteristics of two natural populations of FMDV

	No. of clones	*No. of mutant free clones*	*Minimum mutant frequency x10-4*	*Ts:Tv ratio*	*No. of non-synonymous mutations*	*No. Synonymous mutations*	*dN/dS*	π^{abc}
14524	26	15	2.79	28.53	12	4	1.228	[a]6.2×10^{-4} [b]8.0×10^{-4} [c]5.6×10^{-4}
Accuscript	13	7	3.14	99.00	7	2	1.352	
Superscript III	13	8	2.45	12.06	5	2	1.086	
15101	15	4	3.94	8.41	4	9	0.187	[a]8.0×10^{-4} [b]2.2×10^{-3} [c]3.3×10^{-4}

π^{abc} calculated for all nucleotide sites[a], synonymous sites[b], and non-synonymous sites[c]

5.4.5 Distribution of nucleotide changes between cloned sequences

Figure 5.2 shows the genetic relationship of each cloned sequence for both the two populations from two different RT enzymes (a, b), and the total sequences representing UKG 14524/2001 (c), and also the sequences representing UKG 15101/2001 (d).

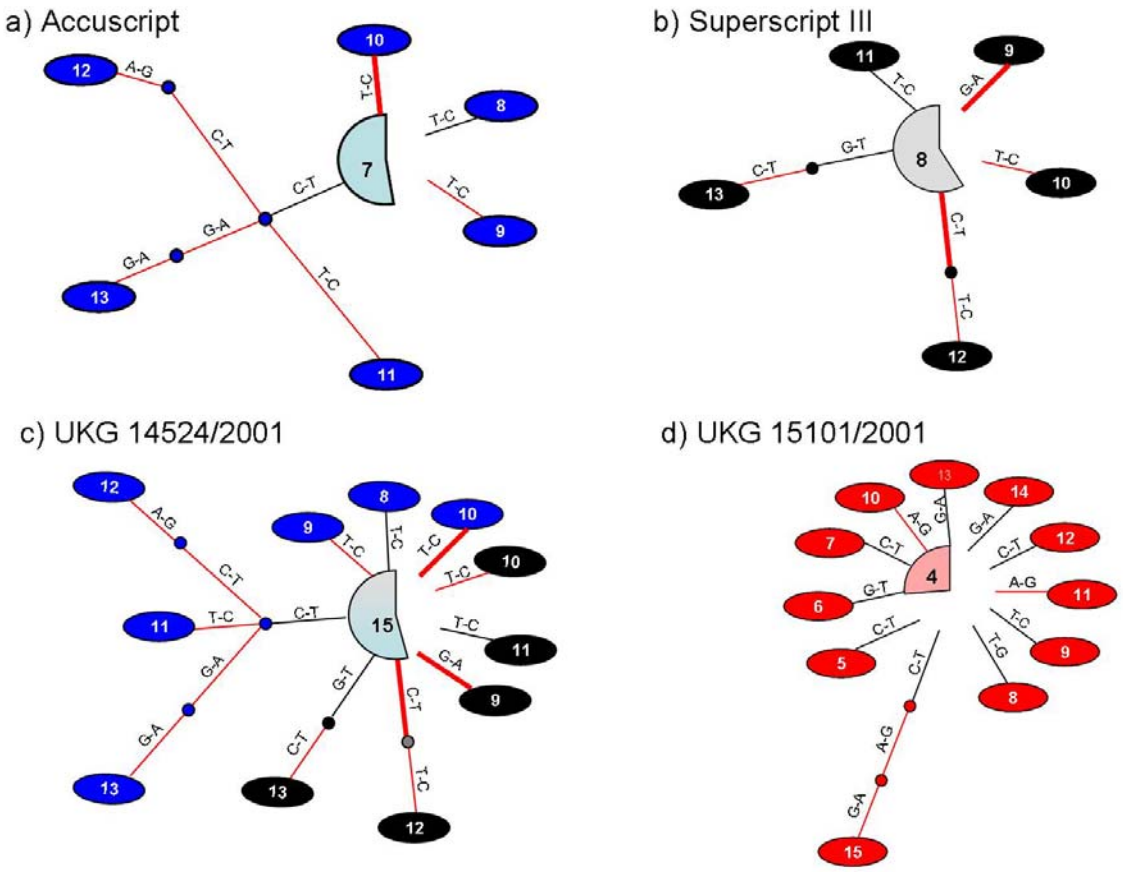


Figure 5.2

Statistical parsimony analysis of capsid sequences. Shown are the genetic relationships between sequences, with each branch representing a single nucleotide substitution: black branches represent synonymous changes, red non-synonymous, and in thick red are those that occurred within antigenic sites. In the centre of each tree the proportion of sequences that were identical to the consensus is represented by a pie diagram, with the exact number printed in black. Each oval represents a sequence and each dot a putative ancestral sequence. a) 13 sequences from cDNA transcribed by Accuscript reverse transcriptase from UKG 14524/2001, b) 13 sequences from cDNA transcribed by Superscript III reverse transcriptase from UKG 14524/2001, c) 26 sequences representing UKG 14524/2001, d) 15 sequences representing UKG 15101/2001.

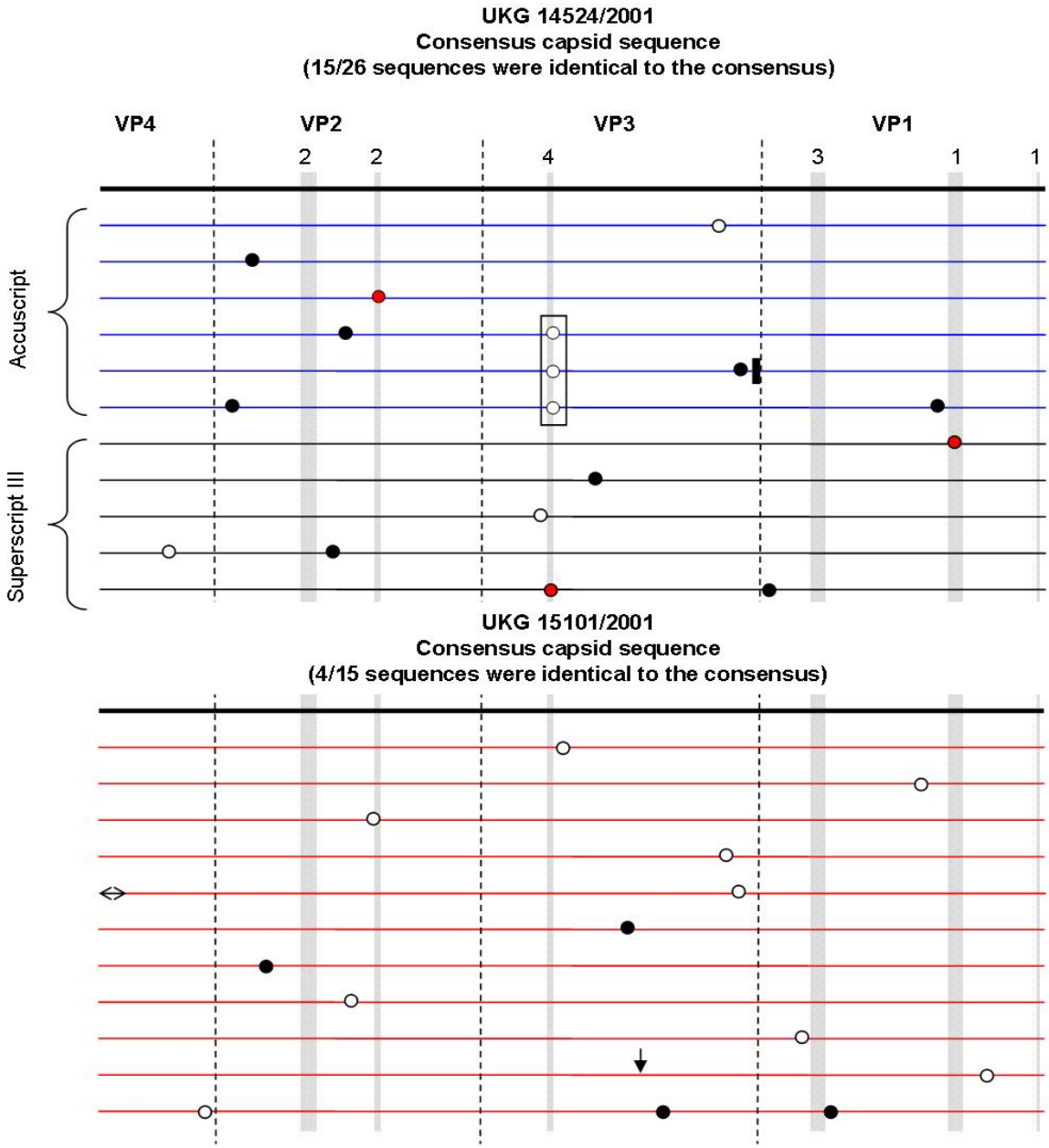


Figure 5.3

The schematic shows all mutant sequences reported in this study compared to their respective consensus sequences. Synonymous substitutions are represented by white circles, non-synonymous by filled circles and non-synonymous substitutions within antigenic sites by red circles. The insertion is indicated by a down arrow, the substitution resulting in a stop codon by a filled rectangle, and the deletion by a double ended arrow. The sections encoding the separate capsid proteins VP1-4 are indicated by the dashed lines, with antigenically important residues represented by grey shading.

Within the UKG 15101/2001 virus population a lower proportion of sequences were found to have no mutations (n=4) than the proportion that had one or more (n=11). This

distribution of mutations is contrary to what would be expected from a Poisson distribution where, for the number of mutations seen, 6 of the sequences should be mutant free, however the mutant distribution was not found to significantly differ from a Poisson distribution (chi-squared test: P=0.071, χ^2=7.036, 100,000 Monte Carlo bootstrap replicates). In contrast, more UKG 14524/2001 cloned sequences (n=15) were found to be identical to the consensus than harbouring 1 mutation (n=6), with the least number being those with 2 (n=3) or more (n=2) mutations. This clearly does not differ from a Poisson distribution (chi-squared test: P=0.290, χ^2=3.566, 100,000 Monte Carlo bootstrap replicates).

5.4.6 Genomic location of variant nucleotides

The relative positions within the capsid coding region of all the mutations found, for both groups of sequences, are shown in Figure 5.3. The only mutation found in more than one sequence was the silent C-T mutation in three sequences from sample UKG 14524/2001. Overall, taking into consideration coding region length, the majority of mutations occurred within the VP3 region (12 out of the 29 substitutions, compared to 8 in VP1). Three non-synonymous mutations were observed in known antigenic sites within the cloned sequences of virus population UKG 14524/2001. Using the Exact Binomial Test it was found that the probability of observing 3 of the 12 amino acid changes seen in virus population UKG 14524/2001 in antigenic sites is extremely low at 0.0007, hence the occurrence of three mutations within antigenic sites is unusual. One of these non-synonymous substitutions resulted in an amino acid change to a critical residue in the VP1 antigenic site 1 G-H loop, another was found to occur in the antigenic site 4 within VP3 and was an amino-acid change that has been seen before in O/PHI/5/95 (N Knowles, personal communication), and the third lay between two residues that form the β-EF loop of antigenic site 2 in VP2.

5.5 Discussion

This study has provided an insight for the first time into the mutant distribution of FMDV genomes within populations analysed from naturally infected animals, infected with Pan Asia O FMDV. The average minimum mutant frequency found in this study was 3.06 x 10^{-4} mutations per nucleotide sequenced. This is close to an *in vitro* study, for serotype O FMDV that sequenced cloned capsids, which found a mutation frequency of 2.1 x 10^{-4} mutations per nucleotide sequenced (Gu et al. 2006). There have been more studies on serotype C virus in which the mutant frequency ranges from 0.7 x 10^{-4} to 7.7 x 10^{-4} mutations per nucleotide sequenced (Airaksinen et al. 2003; Arias et al. 2001; Pariente et al. 2001; Sierra et al. 2000). Those studies that compared mutation frequencies in the VP1, or P1 region to the 3D region found a decreased mutation frequency in the latter, which suggests that the mutation frequency found within the capsid is likely the highest that would be seen throughout the genome. The distribution of the substitutions across the sequences and the presence of identical sequences to the consensus provides support for the dominance of the consensus sequence. The majority of substitutions seen in this study were in VP3, and this corresponds to the majority of fixed mutations noted during the UK 2001 outbreak (Cottam et al. 2006).

It is possible that the mutations observed could be an artefact of the amplification process, introduced, for example, by the reverse transcription enzyme, the PCR enzyme, the *E.coli* replication system, or even during the sequencing reaction. A high fidelity PCR enzyme was used that is reported to give an overall reverse-transcription PCR error rate of 7.7 x 10^{-6} mutations per nucleotide copied (Arezi & Hogrefe 2007; Malet et al. 2003). This would result in a mutation being present in less than 2% of capsid sequences. In this experiment in total we found approximately 54% of sequences to be mutant which is clearly higher. It has also been suggested that the reverse transcriptase can be responsible for mutations seen and this was investigated by assessing the performance

of Accuscript (Stratagene), which is reported to have proof reading qualities, and Superscript III (Invitrogen). No statistical difference was seen between the sequences generated by the two enzymes in terms of evolutionary rate (P=0.559) or *dN/dS* (P=0.863), suggesting that their performances did not differ. High sequence coverage of each nucleotide from different forward and reverse sequencing primers was used to avoid the introduction of errors through analysis of the sequencing reaction. In consideration of this it seems reasonable to suggest that the mutant frequency recorded here is the maximum mutant frequency because it represents the real mutant frequency plus minimal artificial mutants. However, this does assume that the PCR reaction was unbiased in terms of cDNA amplified, and that the RNA of the original sample is equally represented. Indeed, if only a low proportion of RNA sequence was sampled then it would be expected that many mutations would occur on multiple clones. However, all but one of the observed mutations were unique supporting the view that the RNA sampling was not biased.

For the virus population characterised from cow epithelium UKG 14524/2001 the *dN/dS* ratio was 1.228 which is relatively high, in contrast to the *dN/dS* ratio for mutations observed within the virus population from infected sheep epithelium UKG 15101/2001 of 0.187 which is much lower. This difference was not found to be significant due to low sample numbers. During the UK 2001 epidemic non-synonymous and synonymous mutations were fixed into the consensus sequence at a ratio of 0.09 (Cottam et al. 2006), which could thus be regarded as the ratio expected under normal purifying selection. It may be that transmission between animals presents stronger purifying selection than replication within an animal, however these data sets are small and more study is required to determine this. The two virus populations (UKG 14524/2001 and UKG 15101/2001) were not found to differ significantly in terms of diversity.

The distribution of substitutions across clones differs between the virus population from the cow epithelium UKG 14524/2001 and the sheep epithelium UKG 15101/2001, with only 26.6% being identical to the consensus in the population from UKG 15101/2001 compared with 57.7% for UKG 14524/2001. However, this data must be analysed with caution, because it relies on the assumption that all sequence changes are linked as shown and that no recombination has occurred within the PCR reaction (Yu et al. 2006).

For the virus population analysed from UKG 14524/2001 there were 3 changes that occurred in antigenic sites, which is higher than what would be expected at random. It is possible that this represents positive selection of variants that can escape the immune response, as supported by the high *dN/dS* ratio of 1.228. However, both virus populations were sampled from relatively early infections and it is unlikely that the adaptive immune response would be significant. An alternative explanation is that these sites represent parts of the capsid that are more easily mutated than others, either due to some unknown mechanism during replication, or due to relative resistance to deleterious mutation.

Three identical synonymous mutations were found within the 26 sequences from the virus population UKG 14524-2001. The mutations showed a thymine (T) residue, instead of the consensus cytosine (C). A comparison of the consensus with the sequences from the UK 2001 outbreak revealed that 'T' is found at this position in all the other viruses sequenced; indeed it may be more accurate to say that those sequences harbouring a 'C' at this position are the mutant sequences. This observation demonstrates an instance of a nucleotide change becoming fixed within the consensus. The relevance of this to the mode by which nucleotide changes become fixed into the consensus is discussed further in Chapter 7. The presence of three sequences containing the original nucleotide ('T') at this position suggests that those viruses with a 'C' at this residue replicated more abundantly than those with a 'T'. This may not

necessarily be due to increased fitness, but perhaps a result of the relative diversity of the starting population.

Overall 28 transition and 3 transversion mutations were found in this study. A recent paper suggested that poliovirus RNA dependent RNA polymerase also has a comparable transition to transversion bias (Freistadt et al. 2007), although slightly less (4.3 compared to 9.3) than observed in this study. It also corresponds to the overall transition to transversion ratio reported for sequences during the 2001 UK epidemic of 7.61 (Cottam et al. 2006).

Although the difference in the genetic make up of the virus populations between the infected cow and the sheep epithelium is intriguing it cannot be regarded as conclusive because only one sample from each representative animal has been sampled. However the findings of this chapter do encourage further research into virus evolution at the single population level *in vivo*, which may provide additional insight into the fixation of mutations during the evolution of the virus in the field.

Chapter 6

Investigation of the spontaneous mutation rate of Foot-and-mouth disease virus O_1BFS

6.1 Summary

The spontaneous mutation rate of Foot-and-mouth disease virus is thought to be a major contributor to the genetic diversity seen in virus isolates in the field. This rate, however, has not been estimated empirically to date. This chapter describes an approach to determine the spontaneous mutation rate of O_1BFS FMDV using fluctuation analysis. This involved assessing the rate at which mutants arise in parallel cultures, looking for monoclonal antibody escape mutants in the first instance. The initial attempt was hindered by the inefficient neutralising ability of the monoclonal antibody. Also discussed is the possibility of completing this experiment by looking for virus that escapes inhibition by siRNA.

6.2 Introduction

6.2.1 Overview

The mutation rate of FMDV per genome replication is an important parameter to quantify as accurately as possible because it underpins our understanding of virus evolution and replication. There are many estimates of mutant frequencies for FMDV that are all in the order of $x10^{-4}$ mutations per nucleotide sequenced (Airaksinen et al. 2003; Arias et al. 2001; Gu et al. 2006; Pariente et al. 2001; Sierra et al. 2000). Mutation rates calculated from mutation frequency estimates at a single time point should be analysed with caution because low probability events occurring early during the growth of a population are over-represented in subsequent viral generations. Indeed to obtain an accurate estimate of mutation rate it is important to accurately determine the number of virus replication events that have occurred to produce the mutation frequency observed. To date, there is no study that measures the rate of FMDV polymerase error taking into account the number of replication events; there are only studies that measure the frequency of mutants within a population at a fixed time point (see Chapter 1 section 1.3.5).

Despite no clear study of the spontaneous mutation rate of FMDV it is often reported to have a mutation rate of $x\ 10^{-3}$ – $x\ 10^{-5}$ per nucleotide per genome replication, in line with other RNA viruses as discussed in Chapter 1, section 1.3.5. A study which attempted to determine a rate from various studies on Poliovirus by applying fluctuation analysis (discussed below) again found a large variation in mutation rates, perhaps a result of the fact that the studies chosen were not specifically designed for this analysis and many assumptions concerning population size etc. were required (Drake 1993). However, a conclusion concerning the FMDV mutation rate, that ranges from $x\ 10^{-3}$ to $x\ 10^{-5}$ per nucleotide per genome replication, from observations of Poliovirus is extremely uninformative. At its lowest, the mutation rate would represent 99% or more of virus genomes being able to replicate error free; however, at its highest each genome incurs

an average of 10 mutations (nearly 1 per gene). These two scenarios would clearly result in different modes of virus evolution; one in which a quasi-species structure is relevant and the other where quasi-species are completely irrelevant. Another problem is the assumption that FMDV mutation rates will be the same as for Poliovirus, and other RNA viruses. Studies on experimental evolution suggest that a virus mutation rate will evolve to an optimum dependent upon the frequency of transmission through bottlenecks and also the degree of adaptation to the host that takes place before the next bottleneck (Manrubia et al. 2005). The transmission and host characteristics of FMDV differ considerably from Poliovirus, and thus the mutation rate of FMDV is likely to have evolved to be optimum for the frequency of transmission, host range and replication opportunities that it encounters. For these reasons it is imperative to determine the specific mutation rate of FMDV. Although this chapter discusses methods for O_1BFS, once optimised this method could be applicable in assessing the mutation rates of other FMD field viruses.

6.2.2 Fluctuation analysis

A method to measure the rate of FMDV polymerase error harnesses the theories of Luria and Delbruck (Jones et al. 1994); that if the number of new mutations occurring is well approximated by a Poisson distribution with a parameter *m* (the number of mutational events) which equals mutation rate (μ) multiplied by the total number of progeny (N): then the proportion of mutant free colonies (P_0) is equal to e^{-m}, and $\mu=m/N$. Thus by measuring P_0 and N at time zero and time t the mutation rate μ can be determined;

$$\mu = -ln\ (P_0)/\ (N_t-N_0)$$

This analysis is termed fluctuation analysis and has been used to study bacterial populations. However, it has been applied to RNA viruses indirectly (Drake 1993) and also in direct experiments estimating the mutation rate for measles virus (Schrag et al. 1999) and tobacco mosaic virus (Malpica et al. 2002). In Drake's study in 1993 he

attempted to assess the mutation rate of Poliovirus using data from experiments reviewed in Chapter 1 section 1.3.5. However, these studies were not designed to produce data for this analysis and many assumptions were required concerning population size. As a result, a wide range of estimates of the Poliovirus mutation rate were calculated; from 2.2×10^{-3} to 1.3×10^{-5} per nucleotide per genome replication (Drake 1993).

When applying the fluctuation method to RNA viruses, an assumption is that mutational events are distributed randomly across independent virus cultures according to a Poisson distribution. To begin with a number of mutant free cultures are established by inoculating cell cultures in micro-titre plates with a very small quantified virus population. The virus populations are then grown in parallel until a population size is reached resulting in approximately half the virus cultures containing the mutation of interest. It will be necessary to perform initial ranging experiments to determine the appropriate population size at which this is true. The population size at this point is accurately quantified by assessing a set of representative cultures, and the remainder of the cultures are screened for the presence of mutants. The average number of newly produced viral particles N is approximated by the total number present in each culture N_t minus the initial population N_0. The number of mutations is dependent upon the mutation rate μ, and hence each virus population should contain μN mutations. Based on the Poisson distribution the proportion of cultures with no mutations P_0 is equal to $e^{-\mu N}$. Thus the estimation of N and P_0 allows direct estimation of μ.

6.2.3 FMDV Monoclonal antibody escape mutants

To use the method described above to determine the mutation rate of Foot-and-mouth disease virus it was initially decided to screen for the spontaneous generation of monoclonal antibody (MAb) resistant mutants. Many FMDV MAb escape mutants have been characterised, and were found to escape the MAb through the alteration of a limited set of possible nucleotide substitutions. The MAb chosen for use in this study was D9 (a

murine IgG antibody raised to FMDV O_1Lausanne) which binds to antigenic site 1. It has been shown previously that three amino acid residues are important for escape from MAb D9, these are amino acids 144,148 (which lie either side of the RGD in the G-H loop region of VP1) and 154 of VP1 (Kitson 1990). A change at 144 results in the least efficient escape, whereas changes at 148 or 154 confer complete resistance. The only mutations found previously were from L to S at amino acid 144, L to R at amino acid 148 and K to M at amino acid 154. Analysis of O_1BFS sequence suggests that it should be neutralised by D9 Mab because it contains a V at amino acid 144, L at amino acid 148 and K at amino acid 154. It is not possible to change by a single point mutation from the V (GTG) to encode S and thus mutants are unlikely to arise by this route. However the change from L (CTG) to R (CGG) at 148 would require a single transversion from T to G, and the change from K (AAG) to M (ATG) at amino acid 154 would require a single transition mutation from A to T. Previous studies carried out by (Kitson 1990) suggest that these mutants would be viable.

6.2.4 siRNA escape mutants

As described in the results section of this chapter it was not possible to complete the experiment using the MAb. For this reason an alternative method was explored. The new concept was to use silencing RNA to inhibit the virus and look for mutants escaping its effect. Small interfering RNA can cause sequence specific destruction of a target RNA sequence; (Dallas & Vlassov 2006; Zamore et al. 2000; Zhou et al. 2006). Although usually used to silence protein expression by destroying the mRNA, siRNA technology has also been targeted to virus genomes preventing viral replication. Indeed, previous studies have shown specific inhibition of FMDV (Chen et al. 2004; Kahana et al. 2004; Liu et al. 2005b; Mohapatra et al. 2002).

siRNAs are double stranded RNA molecules 21-23 nucleotides in length that contain 2 nucleotide overhangs at the 3' ends. When inside cells they are integrated into a multi-

subunit protein complex called the RNAi inducing silencing complex (RISC) in an ATP dependant manner (Zamore et al. 2000). The RISC guides the siRNA to the target RNA sequence. In order to function the siRNA duplex is unwound with the anti-sense strand remaining bound to the RISC. Combinations of endo- and exo-nucleases then degrade the target RNA. The capacity for virus to escape siRNA action relies on nucleotide substitutions within the target sequence.

6.2.5 Experiment design

To complete the experiment to assess the mutation rate of FMDV, it was first essential that a method is in place to completely inhibit virus replication so that mutants could be screened for with ease. Secondly, a time course in suitably sized tissue culture dishes was performed, to establish the rate at which virus populations increase and the extent to which they could grow within the cell monolayer available. This provided a basis around which to perform the ranging experiments to determine at what population size half the cultures contained a mutant. Finally multiple replicas of the experiment would be completed to obtain a statistically accurate estimate for the mutation rate.

6.3 Methods

6.3.1 Cells

Baby-Hamster-Kidney 21 cells (BHK21) were used throughout the experiment. BHK21 cells were maintained in complete growth medium (500ml Glasgow Eagles supplemented with 10% Foetal Calf Serum (FCS), 25ml Tryptose Phosphate Broth (TPB), 5ml Glutamine, 5ml Penicillin/Streptomycin), incubated at 37^{o}C with 5% CO_2. Stock cells were split to a ratio of 1:10 every 48-60hrs, by addition of Trypsin and brief centrifugation to pellet cells, followed by re-suspension in complete growth medium and seeding of new culture flasks.

6.3.2 O1BFS virus

A cell culture isolate of the British Field Sample O_1BFS (donated by Terry Jackson) was 3 x plaque purified (see below; section 6.3.5) and used throughout the experiment.

6.3.3 MAb D9

Undiluted ascites fluid of MAb D9 from stocks held at the Institute for Animal Health was used. The concentration of MAb within this fluid was unknown, and difficult to quantify due to the presence of other proteins. However, it is known that ascitic fluid represents much more concentrated MAb than can be obtained from cell supernatants from hybridomas, and thus this MAb (D9) was the best candidate available for the completion of this experiment.

6.3.4 Sequencing of O_1BFS

Virus RNA was extracted and reverse transcribed as described in Chapter 3, section 3.3.3. The cDNA was cleaned using the GFX PCR DNA and gel band purification kit

(Amersham Biosciences). The capsid coding region was amplified using primers designed against the O_1BFS published sequence (AY593815) named O_1BFS Capsid/For1 and O_1BFS Capsid/Rev1 detailed in Table 6.1. Thermal cycling was as follows; initial denaturation at 94°C for 2 minutes followed by 39 cycles of 94°C for 30 seconds, 62°C for 30 seconds, 72°C for 3 minutes, ending with incubation at 72°C for 7 minutes. The PCR product was cleaned using the GFX PCR DNA and gel band purification kit (Amersham Biosciences) and sequencing reactions were set up as described previously in Chapter 2, section 2.3.1d. The sequencing primers used are detailed in Table 6.1. Raw data files were assembled into a contig using the DNASTAR software package and the consensus sequence was analysed with Bioedit freeware.

Table 6.1 Primers for amplification and sequencing of O_1BFS FMDV capsid

Name	Direction	5' nt position wrt AY593815	Sequence 5' to 3'
O1BFS Cap/For1	F	1520	catgtgcttggctgatttcc
O1BFS Cap/For2	F	1926	gactggttctccaaacttgc
O1BFS Cap/For3	F	2244	ccacaaaggtgtctacggcag
O1BFS Cap/For4	F	2560	cagatcaaggtgtatgccaac
O1BFS Cap/For5	F	2841	cagggtgcttgctcagtttg
O1BFS Cap/For6	F	3153	gagaccacaaatgtgcagg
O1BFS Cap/For7	F	3458	ggttccatcacacactttgg
O1BFS Cap/For8	F	3730	gtgaccttcaagtgttggctc
O1BFS Cap/Rev1	R	4155	cttgataagcttgtaccaggg
O1BFS Cap/Rev2	R	3678	cctgcactcaccgttgtacac
O1BFS Cap/Rev3	R	3375	cataatgaacgagacgtccgtg
O1BFS Cap/Rev4	R	3046	cattcagcatgaatgcagtgc
O1BFS Cap/Rev5	R	2753	cagccacatcaaggaggttgg
O1BFS Cap/Rev6	R	2429	gtaatgtgcgcagtcatgttcg
O1BFS Cap/Rev7	R	2123	gtctcgagaccagaagtgttcg
O1BFS Cap/Rev8	R	1880	ggaggttgtgtcggtggag

6.3.5 Plaque purification

BHK21 cells were seeded into 60mm tissue culture Petri dishes to give 60-70% confluence by the following day, and grown in Glasgow Eagles complete growth medium (10% FCS, TPB, Glutamine, Pen/Strep). The growth medium was then removed and the

cells were washed with serum free medium before inoculation with 200µl of virus, diluted appropriately in serum free Glasgow Eagles medium. The cells were incubated with the virus inoculum for 15 minutes before washing twice with serum free medium. The cells were then overlaid with 2ml of agar overlay (0.6g Indubiose A37 agarose (PALL Life Sciences) dissolved in 25ml nuclease free water using a microwave, mixed with 75ml Glasgow's overlay medium, 5ml TPB, 1.5ml FCS, 1.5ml Pen/Strep). Following incubation at 37^{o}C with 5% CO_2 for 48 hours the plaques were visible and were picked from the appropriate dilution showing distinct individual plaques. The plaques were picked using a sterile plastic pipette tip which was then immersed in 450µl of serum free medium for 30 minutes before pippetting up and down and storing at -70^{o}C. This process was carried out three times to obtain 3x plaque purified virus.

6.3.6 Plaque assays

Plaque assays were performed in duplicate in 60mm tissue culture Petri dishes, with cells seeded as described above. Serial 1 in 10 dilutions of virus were used to infect the cells, and the cells were overlaid with agar as described above. After 48 hours incubation, the cells were stained with methylene blue stain overnight, and then washed to visualise the plaques. Plaque forming units (PFU) were calculated from the number of plaques visualised at a countable dilution to give PFU per ml of the starting virus inoculum.

6.3.7 Monoclonal Antibody inhibition

Cells were seeded into 6 well tissue culture plates to give 60-70% confluence by the following day. The cells were inoculated as described above, however before inoculation the virus dilution was mixed with the appropriate dilution of MAb (diluted in Glasgow Eagles medium). After the inoculation, the cells were not washed (this was found to improve the neutralisation capacity of the MAb; data not shown). The cells were overlaid

with agar as described above and incubated at 37°C with 5% CO_2 for 48hrs before staining and plaque visualisation.

6.3.8 ELISA for assessing adherence of D9 MAb to fixed antigen

Polyclonal O_1BFS rabbit serum (IAH) and coating buffer (Carbonate-bicarbonate, Sigma) were used to coat the wells of a 96 well plate overnight at 4°C. The wells were washed three times with PBS before addition of BEI (Bromoethyleneimine, Sigma) inactivated O_1BFS virus antigen in PBS followed by 1hr incubation at room temperature. Following another three washes in PBS either D9 MAb, or guinea pig O_1BFS anti serum as a positive control, were added and diluted across the plate in PBS and Tween. The plate was then incubated at room temperature for 1hr. The wells were then washed three times in PBS prior to addition of the conjugate; anti-mouse Horse-Radish-Peroxidase for D9 and anti-guinea pig Horse-Radish-Peroxidase for the positive control. The plate was then incubated at room temperature for 1hr, followed by three washes in PBS. O-phenylenediamine (OPD) (Sigma) plus hydrogen peroxide was added to each well followed by incubation in the dark for 15 minutes. The reaction was then stopped by the addition of stop acid (H_2SO_4). The optical density (OD) was read at 492nm.

6.3.9 Time course

BHK21 cells were seeded into 24 well tissue culture plates, to give 60-70% confluence by the following day. The growth medium was removed and the cells were washed with serum-free medium before addition of the virus, followed by Glasgow Eagles medium supplemented with 1% FCS (total volume of 1ml per well). At specific time points the infection was stopped by freezing the plate at -20°C. The plate was then thawed and the contents of each well were harvested into a 2ml tube and stored at -70°C. For each time point, the virus population in PFU was calculated from two wells by duplicate plaque assays as described above.

6.3.10 siRNA Transfection

Glass cover slips were placed into the bottom of wells in a 24 well tissue culture plate and washed with 70% ethanol, and then Glasgow Eagles medium. Cells were then seeded into the wells to give 50% confluence the following day. Cells were transfected following the manufacturer's protocol for the TransIT-siPAK Plus Trial Kit (Mirus). Briefly, the transfection reagent was mixed with 50µl of serum free medium, followed by addition of siRNA for a final concentration of 25nM per well and incubated at room temperature for 15 minutes. The transfection reagent/siRNA complex mixture was then added drop wise onto cells in complete growth medium and the cells were incubated at 37^{o}C with 5% CO_2. Two different reagents were tested for their efficiency at transfecting the Label IT RNAi Delivery Control fluorescein; TransIT-siQUEST and TransIT-TKO.

6.3.11 Confocal microscopy

After 18 hours, the transfected cells were fixed with 4% paraformaldehyde in PBS, at room temperature for 40 minutes. The fixing agent was then removed and the cells were stained with diluted DAPI (1µl 1000x DAPI in 10ml ddH_2O) for 10 minutes at room temperature. The cover slip was then removed using forceps and washed in ddH_2O and blotted on tissue paper before placing cell side down onto a drop of vecta shield (Vector Laboratories Inc) on a glass microscope slide. After blotting with tissue paper the cover slip was sealed with clear nail varnish. The slides were then visualised using a confocal microscope.

6.4 Results

6.4.1 O_1BFS complete capsid sequence

The complete capsid sequence of the original virus stock prior to plaque purification was obtained. Analysis of the sequence data showed that there were two non synonymous changes in the antigenic site 4 of VP3 from the published sequence AY593815. These changes were both from an adenine to a guanidine and result in changes to amino acids $VP3^{56}$ and $VP3^{60}$, which are indicative of adaptation to growth in cell culture.

6.4.2 D9 MAb ELISA

As shown in Figure 6.1 the OD readings suggest that the MAb D9 activity is greatly reduced at a concentration less than x 10^{-3}. Although this represents MAb binding to fixed antigen, these ranging experiments indicated a range of dilutions to test on live virus i.e. no less than 1 in a 1000.

6.4.3D9 MAb inhibition

Three different MAb volumes were used (0.5µl, 2µl, 5µl corresponding to final dilutions of 2.5 x 10^{-3}, 1 x 10^{-2}, and 2.5 x 10^{-2} respectively) to inhibit different concentrations of O1BFS virus ranging form 1 x10^2 to 5 x10^6 PFU in 200µl. The results are shown in Figures 6.2 – 6.4. The graph in Figure 6.2 shows that 0.5µl of D9 MAb failed to completely inhibit even low virus numbers i.e. less than 100 PFU. Volumes of 2 or 5µl of D9 MAb were more successful, but were unable to inhibit virus by more than 50% at PFU higher than 10^4, and at PFU lower than this struggled to give 100% inhibition with 1 or 2 plaques often visible.

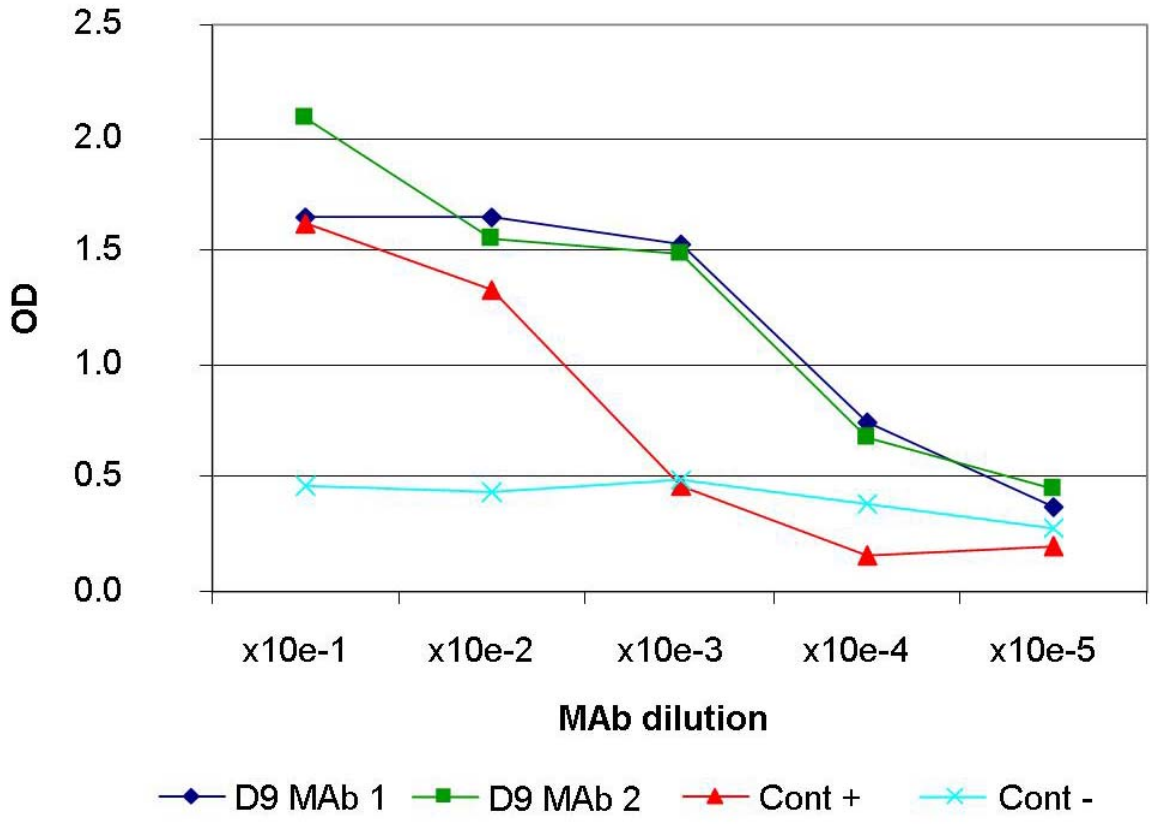


Figure 6.1

Analysis of D9 MAb by ELISA. The graph shows the optical densities measured following ELISA on dilutions of two separate aliquots of D9 monoclonal antibody ascitic fluid (D9 MAb 1 and 2), compared to guinea pig antiserum as a positive control (see material and methods section 6.3.8). The activity of D9 is greatly reduced at a dilution higher than one in a thousand.

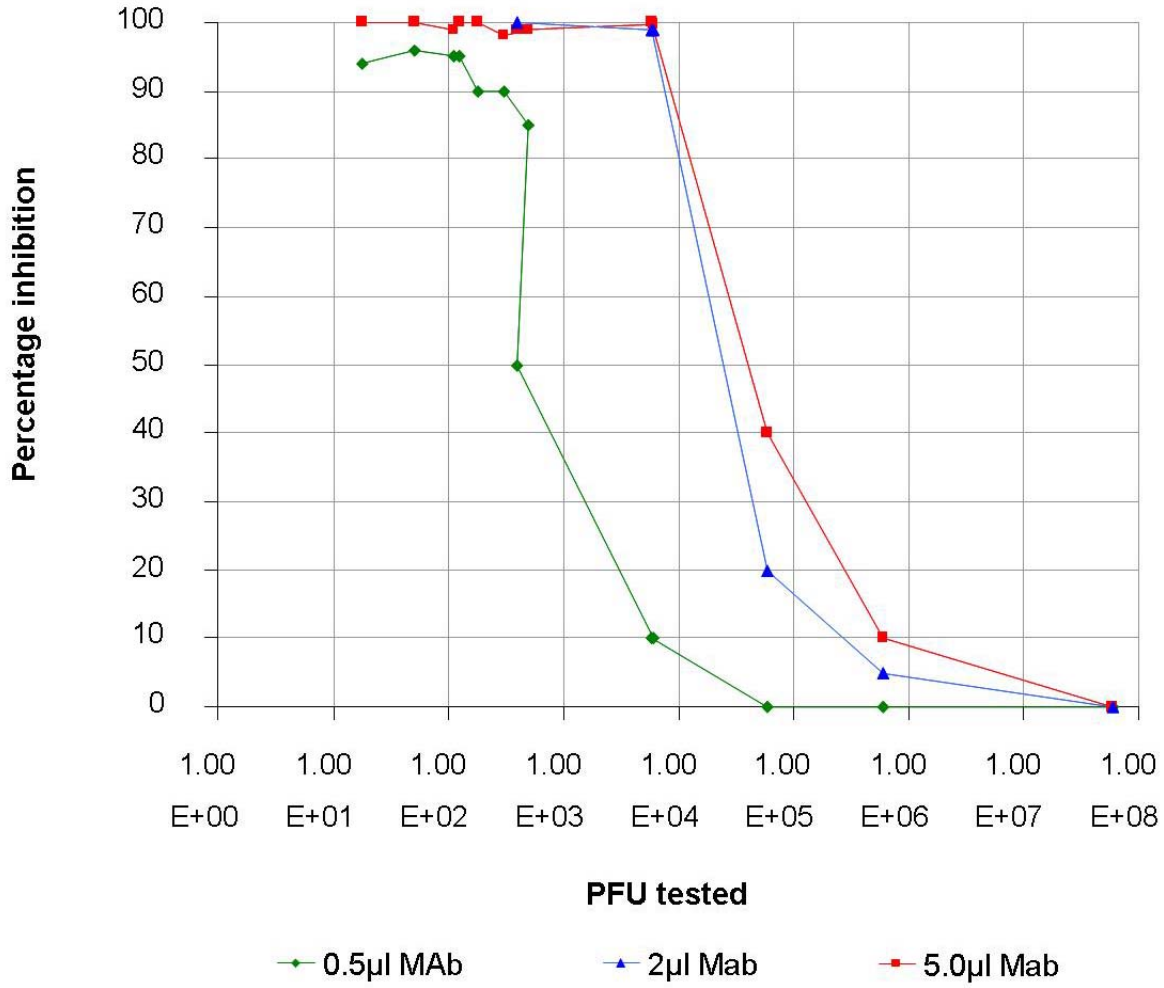


Figure 6.2

Inhibition of virus growth with varying concentrations of D9 MAb. The graph shows the percentage inhibition of O_1BFS virus plaque formation on BHK21 cells caused by different amounts of monoclonal antibody D9 (0.5µl in green, 2µl in blue and 5µl in red) on varying starting concentrations of virus (PFU). None of the volumes of MAb tested could substantially inhibit virus growth at a titre above 10^4 PFU.

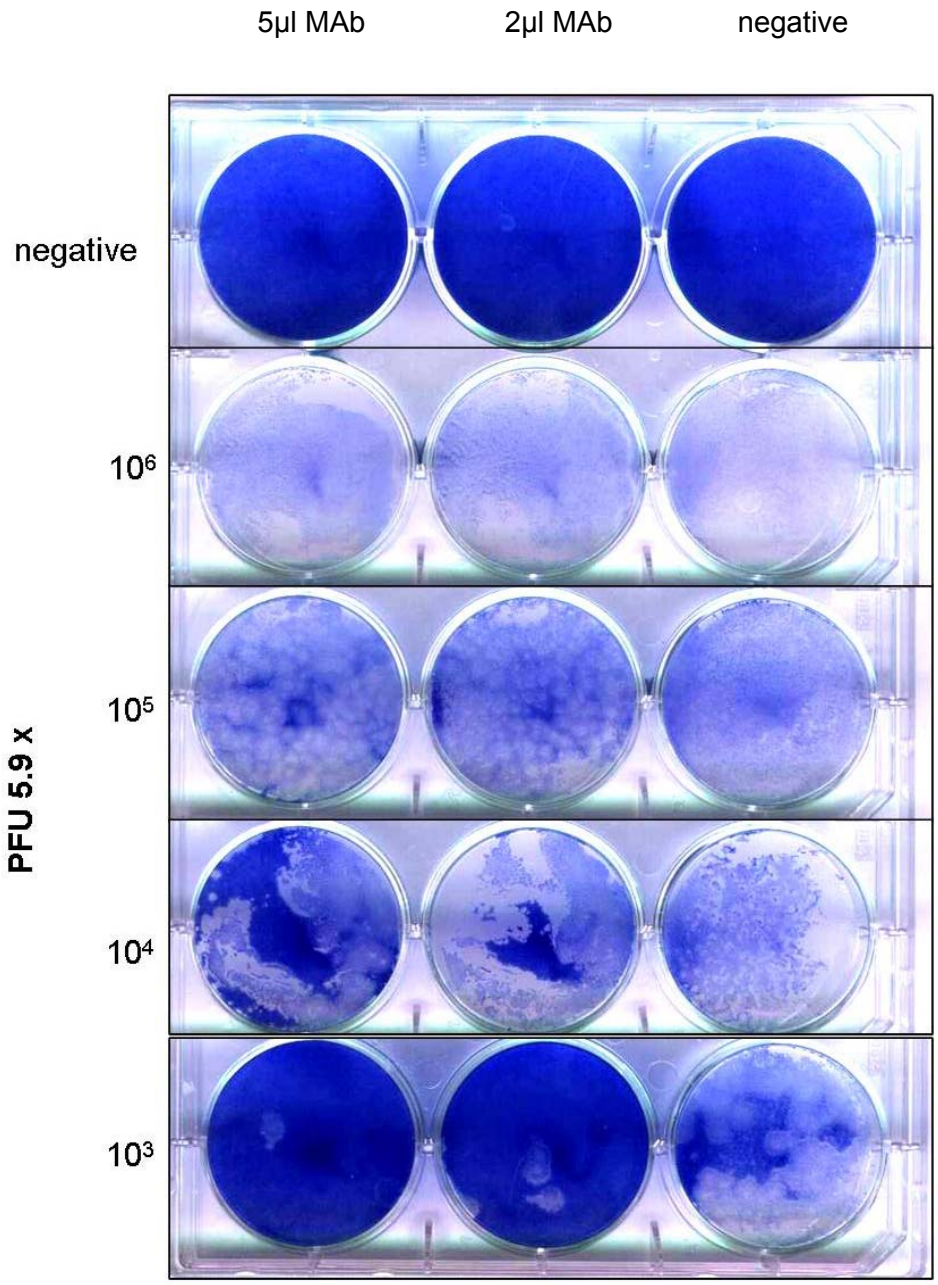


Figure 6.3

Inhibition of different quantities of O1BFS virus plaque formation by 2 and 5µl of D9 MAb. Three times plaque purified O_1BFS virus (5.9 x 10^6 PFU and 10 fold dilutions down to 5.9 x 10^3 PFU) was mixed with 5, 2 or 0µl of D9 MAb before inoculation of BHK21 cell monolayer in a 6 well plate. The cells were stained 48 hours post infection, and the extent of virus induced CPE visualised, as shown above, for the different amounts of virus and MAb D9. Even at virus PFU of 5.9 x 10^3 the MAb was unable to completely inhibit plaque formation as illustrated by the bottom left hand and bottom middle well.

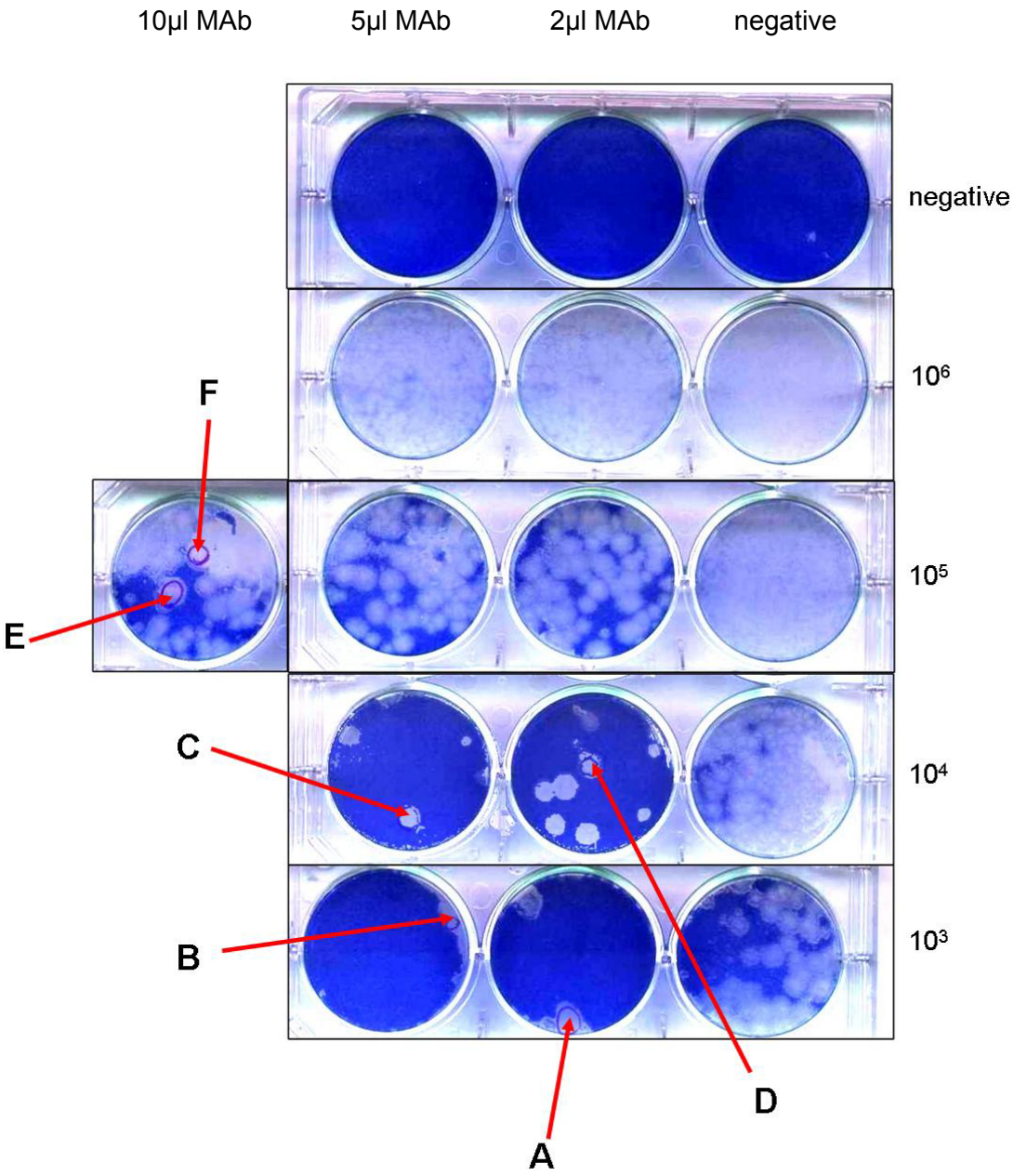


Figure 6.4

Inhibition of different quantities of O1BFS virus plaque formation by 2, 5 and 10µl of D9 MAb. Three times plaque purified O_1BFS virus (1.06 x 10^6 PFU and 10 fold dilutions down to 1.06 x 10^3 PFU) was mixed with 5, 2 or 0µl of D9 MAb (and in one case (1.06 x 10^5 PFU) 10µl of D9 MAb) before inoculation of BHK21 cell monolayer in a 6 well plate. The cells were stained 48 hours post infection, and the extent of virus induced CPE visualised, as shown above, for the different amounts of virus and MAb D9. Even with virus PFU of 1.06 x 10^3, the MAb was unable to completely inhibit plaque formation as illustrated by the bottom left hand and bottom middle well. A-F denote plaques that were picked to test to see if they were D9 MAb escape mutants.

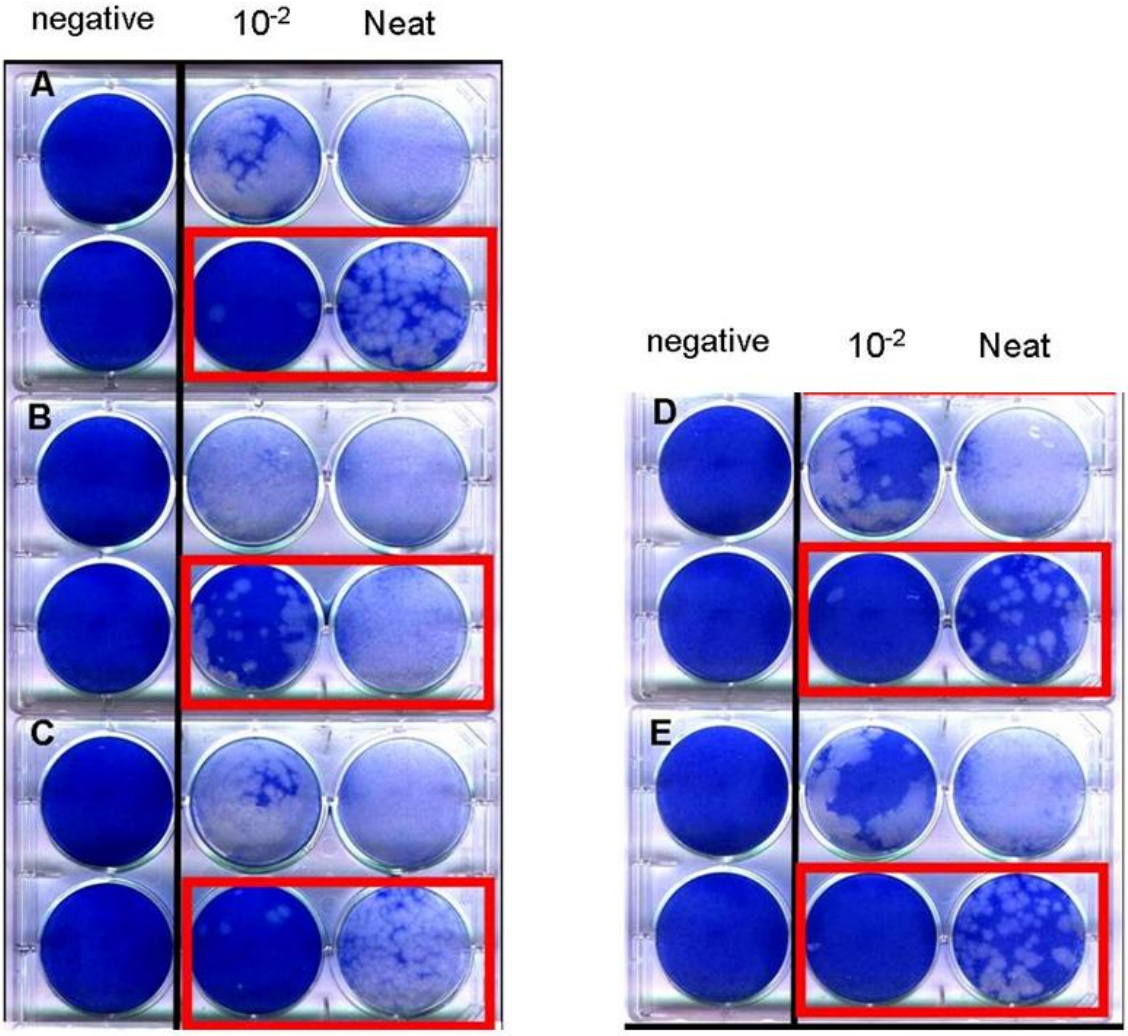


Figure 6.5

Testing whether picked virus populations A-F are D9 MAb escape mutants. Virus harvested from plaques A-F (Figure 6.4) was used neat and diluted 10^{-2} (as indicated) to infect BHK21 cell monolayers in 6 well tissue culture plates, either in the presence of 2µl of D9 MAb red boxes or not. All virus populations were inhibited by the MAb confirming that they were not MAb escape mutants.

6.4.4No MAb escapes detected

Plaques visualised under high MAb concentrations were tested for the ability to escape the MAb. The picked viruses were incubated with 2µl of MAb before inoculation. The results are shown in Figure 6.5, and clearly show that all 5 suspected mutant virus populations are inhibited by the MAb and are therefore not mutants.

6.4.5 O1BFS time course

Initially, a time course experiment was begun with <10 PFU virus inoculum per tissue culture well. At 24 hours, there was a large variation in PFU between the different wells (0 to 1x 10^6 PFU) and so it was concluded that it would not be possible to accurately determine the virus population size in all wells from representative wells. To obtain more consistent results the experiment was repeated starting with <100 PFU per well. The results for two different 3x plaque purified viruses, shown in Figure 6.6, are very similar. This suggests that virus populations could be reliably quantified from representative wells if the population was initiated with more than 10 but less than 100 PFU.

6.4.6 siRNA transfection

Following confocal analysis, it was found that 4µl of the TransIT-TKO transfection reagent gave the best transfection, shown in Figure 6.7. The Label IT RNAi was found to give a cytoplasmic distribution suggesting that it had been taken up into the BHK-21 cells. In this experiment, it appeared to show a close to 100% transfection rate. For the TransIT-siQuest reagent it was not possible to confirm that the RNA had been taken up into the cells because the fluorescence seen was very clumped and was not localised to the cells (data not shown).

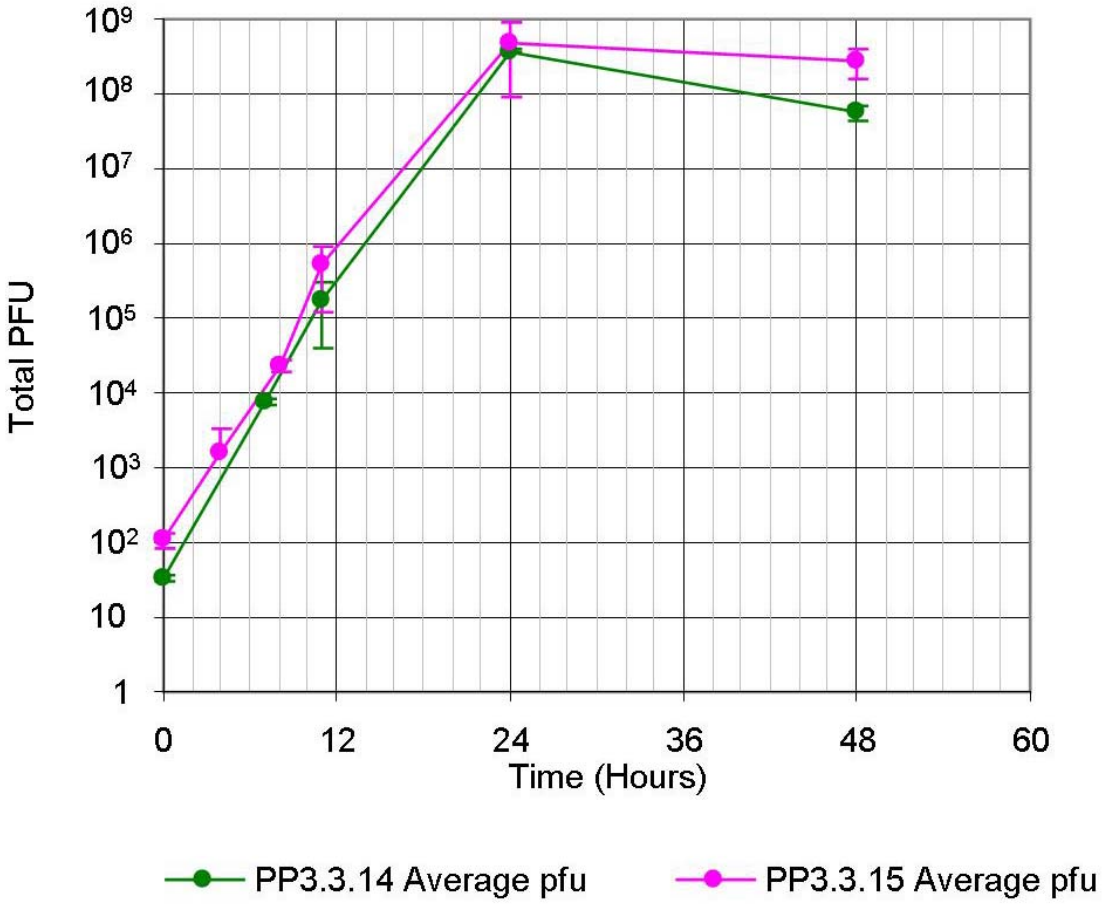


Figure 6.6

Time course of two 3 x plaque purified O1BFS viruses' growth on BHK21 cells in a 24 well tissue culture plate measured by duplicate plaque assay. Two separate three times plaque purified viruses pp3.3.14 (green) and pp3.3.15 (pink) were used to infect BHK21 cell monolayers in 24 well plates at approximately 100 PFU (dilution determined from previous plaque assay, data not shown). Total PFU present in the 1ml of cell and medium solution within two or more representative wells at each time point shown were assessed by duplicate plaque assays.

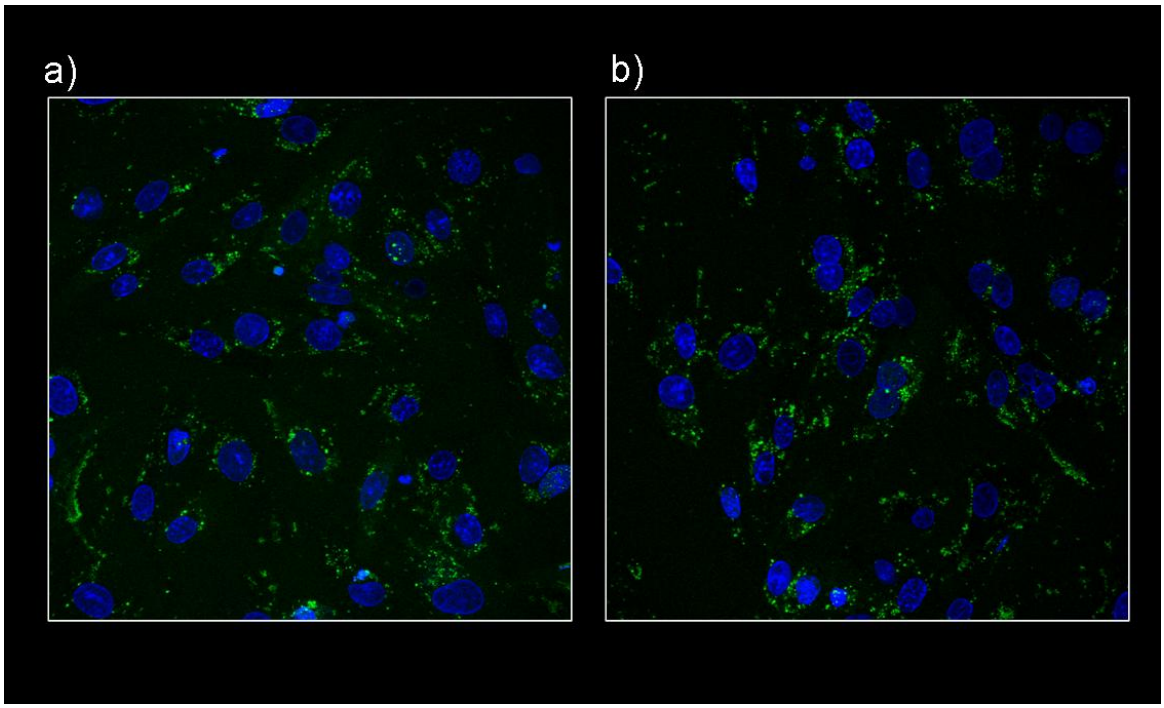


Figure 6.7

Transfection of BHK21 cells with Label IT RNAi Delivery Control fluorescin. The two confocal images (a and b) show Label IT RNAi Delivery control fluorescin (green) transfected into BHK21 cells using TransIT-TKO transfection reagent (Mirus). The nuclei of the cells are stained blue with DAPI.

6.5 Discussion

The O_1BFS virus capsid sequence of the starting stock was found to be comparable to that sequenced previously, AY593815 (Carrillo et al. 2005). The two sequences differed by two non-synonymous variations at two sites in the antigenically important site 4 in VP3. All of the variant amino acids have been seen in other isolates previously sequenced, as they represent changes that are indicative of adaptation to cell culture.

Although the MAb worked well at high concentrations it was not able to satisfactorily completely inhibit virus, nor to select true MAb escape mutants. Thus to complete the experiment a large amount of MAb would be required as any putative escape viruses would need to be picked and checked to see if they were real escape mutants. There was only 1 ml of D9 MAb ascetes fluid available which would not be enough to complete this experiment, as a minimum of 100 replicate populations would need to be screened, and then any putative escapes re-screened. If the mutation rate is 1 x 10^{-4} this would require more than 2ml of MAb which was not available. Thus, it would only be possible to complete the experiment if the mutation rate was in the order of x 10^{-3}, which is an unlikely scenario. To use up an irreplaceable commodity such as MAb to test to see if the mutation rate is equal to this would be a waste of a valuable resource. Previous studies to generate escape mutants have involved serial growth of virus with low levels of MAb, which would not be suitable approach for the fluctuation experiment (Kitson 1990).

For an ideal fluctuation experiment the starting population needs to be as low as possible, and so with this in mind it was initially attempted to grow replicate populations from less than 10 PFU. However, it was found that in order to obtain reliable estimates of population size at different time points closer to 100 PFU were required as a starting point as this reduced the inter population variation. However, if the mutation rate is

below 1 x 10^{-3} mutations per nucleotide site per genome replication (highly probable) then 100 PFU is sufficiently low in comparison with the end population to be able to accurately estimate the mutation rate by fluctuation analysis.

A possible way to overcome the limitation of the experiment due to low stocks of concentrated MAb, is to use an alternative method for detecting mutants. Such a screen could be carried out by using silencing RNA to silence the virus, and thus requiring a nucleotide mutation to arise to allow escape from the silencing agent. For this reason the feasibility of transfecting BHK21 cells with siRNA was investigated to assess the percentage transfection that might be possible. The results suggest that the TransIT-TKO reagent is a good candidate for efficient transfection of siRNA into BHK21 cells. A high transfection rate would be required if the virus is to be inhibited effectively. Looking for MAb mutants allows only the assessment of mutation rates for specific regions within the capsid sequence of the virus genome. siRNA however, offers the opportunity of assessing mutation rates in any location across the whole genome, which is a clear advantage.

This study, although it has not provided a formal conclusion, has been a pilot study for the assessment of the virus spontaneous mutation rate through fluctuation analysis. The data obtained concerning the growth rate of the virus and the method for siRNA transfection of BHK cells could facilitate future experiments to quantify this extremely important parameter.

Chapter 7

Conclusions and Discussion

7.2 Overview of Thesis

7.2.1 Overview

The potential of genetic data to enhance our understanding of FMDV evolution and transmission within a single epidemic has been investigated within this thesis through complete genome sequencing of FMDV. Figure 7.1 summarises the main findings of this thesis, and also illustrates how the separate chapters link together. The consensus genetic variation seen in the UK 2001 epidemic described in Chapter 3 was a result of nucleotide substitutions becoming fixed upon replication within, and transmission between, animals and farms. These mutation rates are described as μ_c and μ_d respectively in Figure 7.1. Although it was possible to investigate many farm-to-farm transmissions (described in Chapter 4), the data regarding underlying nucleotide changes upon animal-to-animal transmission are much more limited (mainly due to a lack of suitable samples). The nucleotide substitutions that become fixed into the consensus sequence of the virus were thought to arise from the within population diversity of virus genomes (d_b Figure 7.1), which has been studied in Chapter 5. The data obtained suggested that the within host diversity of virus needs to be investigated more thoroughly in the natural animal host if we are to understand the complexity of virus genetic evolution. Finally, as illustrated by Figure 7.1, the diversity of the virus population is dependent upon the spontaneous mutation rate of the virus (μ_a) which has not been precisely quantified to date. This problem is addressed in Chapter 6, however progress was hindered due to time constraints and recent events that prevented work on the virus strain O_1BFS.

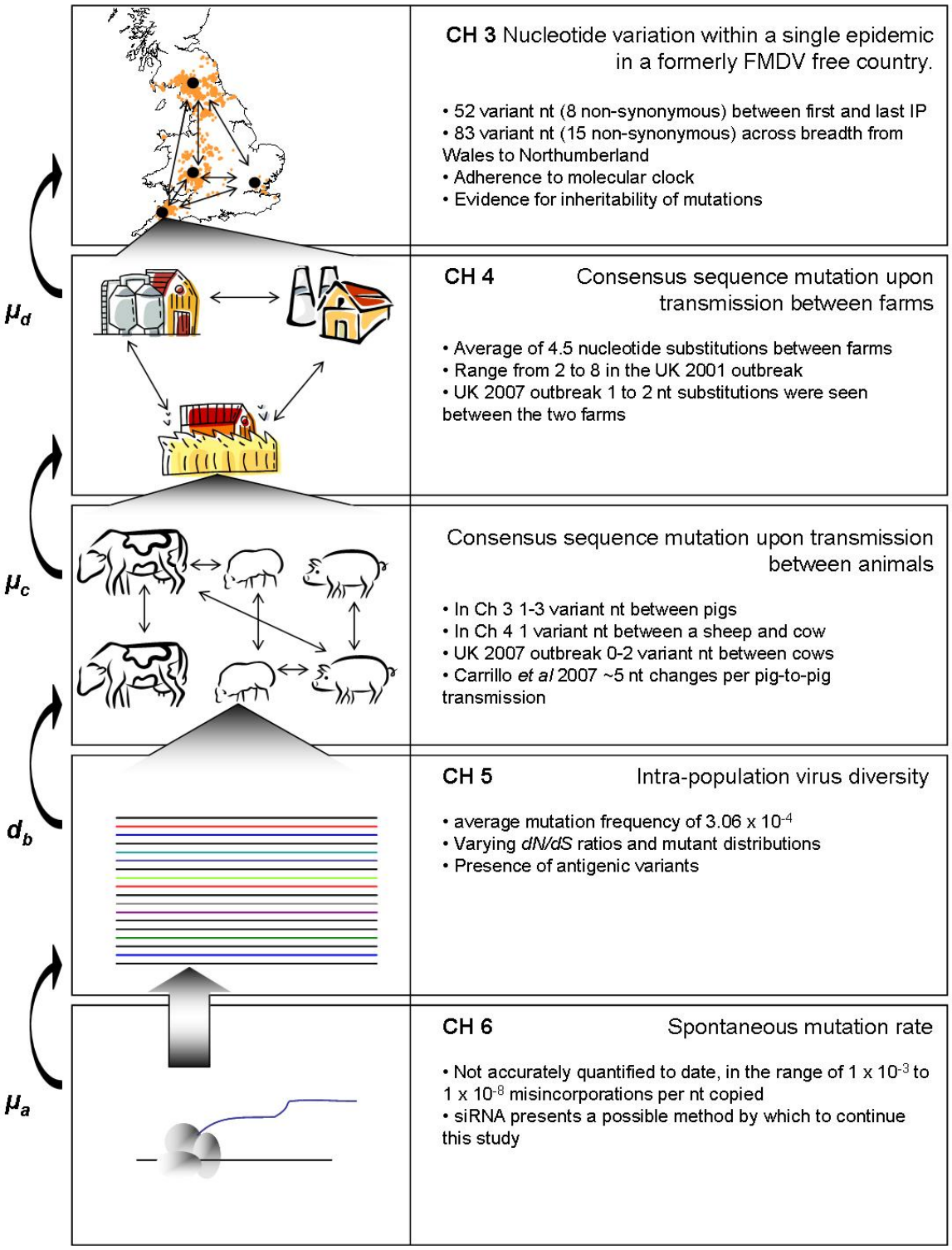


Figure 7.1

Conclusions of thesis. The main findings of each chapter are shown. On the left of the figure μ_a is the spontaneous mutation rate that determines the within population diversity (d_b), which in turn impacts upon the rate at which nucleotide substitutions occur upon transmission between animals (μ_c), by determining the extent of bottleneck required to enforce a sequence change within the consensus. The rate of nucleotide change upon

animal-to-animal transmission will influence the number of nucleotide substitutions that are apparent upon transmission between farms (μ_d), which resulted in the genetic variation observed across the whole UK 2001 FMDV outbreak.

7.3 FMDV evolution in the field

7.3.1 Observations

A key finding of this thesis was the extent of genetic nucleotide evolution of FMDV that occurred during the 2001 FMD outbreak in the UK. This was previously unknown, and not quantified across complete genomes, for any other such epidemic. Virus sequenced from the last infected premises in the outbreak differed by 52 nucleotide changes from the primary case, and 83 variant nucleotides were noted between an isolate from Wales and one from Northumberland. This extent of nucleotide change enabled further in depth study of the evolution of the virus during this epidemic. This genetic variation is illustrated in Figure 7.2, a Bayesian majority rule consensus tree of all the UKG 2001 FMDV complete genomes sequenced during this study.

Nucleotide substitutions were shown to accrue throughout the entire genome, and the VP1 gene was not found to be more variable than other genes in this situation. A higher number of synonymous changes than non-synonymous were observed which suggested that the virus has evolved mainly by genetic drift and purifying selection. However, these findings do not exclude positive selection since a single amino acid in the VP1 protein (positioned at $VP1^{42}$ adjacent to neutralising antigenic site 3), was mutated twice through different nucleotide point mutations on independent lineages (Chapter 3). Additionally antigenic variants were noted when studying intra-population sequence differences in Chapter 5.

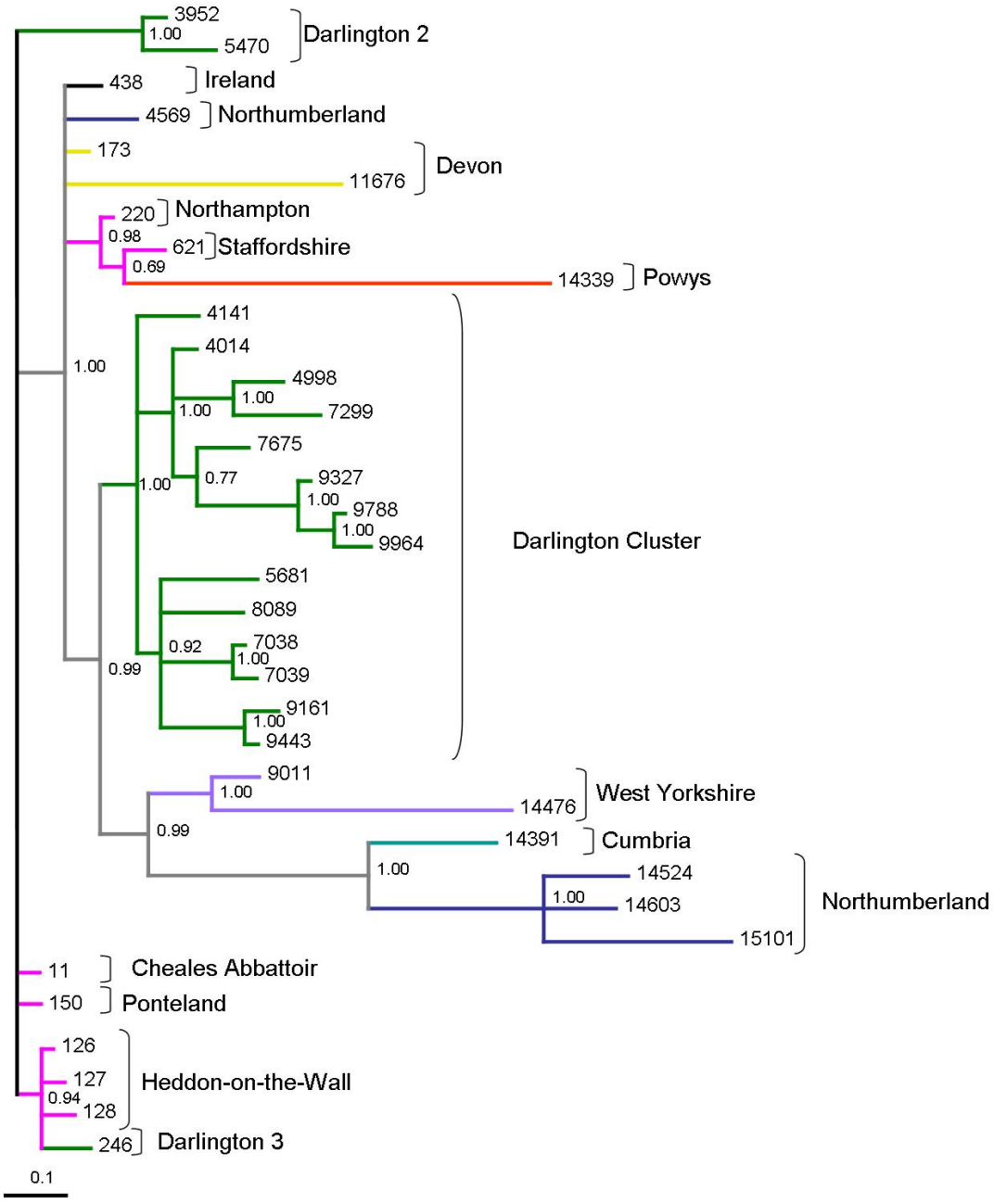


Figure 7.2

A Bayesian majority rule consensus tree of all UK 2001 complete genome sequences studied to date. The majority rule consensus tree (based on 1000 trees sampled from 1 million generations) was estimated using the software package Mr Bayes (Huelsenbeck & Ronquist 2001), assuming a Generalised Time Reversible model with gamma-distributed rate variation across sites and a proportion of invariable sites. Posterior probabilities are indicated on the tree, and each sequence is labelled by its UKG WRL number, and geographic location of the source. Dark green shows those isolates studied in Chapter 4 from Darlington.

The rate at which nucleotide substitutions accumulated within the consensus genome was shown to adhere to a molecular clock, with a rate of 2.26 x 10^{-5} per nucleotide site per day. It was also found that in the majority of cases nucleotide substitutions that were fixed into the consensus sequence were inherited, and represented in virus sampled at later dates from the same outbreak. The virus was shown to acquire changes such that 2 to 8 substitutions were apparent between closely related farms, with a mean of 4.2 upon transmission.

Although an average of 2.5 x 10^{-5} nucleotide substitutions/site/day was found over the whole outbreak (Chapter 3), for some consecutive farm-to-farm transmissions a much higher rate was seen, for example 3.1 x 10^{-4} between IP6 and IP7, and 1.2 x 10^{-4} between IP16 and IP38. Thus a high number of nucleotide substitutions occurred for the expanse of time that elapsed. This lends argument to the theory that nucleotide changes are fixed upon transmission of the virus. In consideration of this, the adherence of the rate of FMDV genetic evolution to a molecular clock is also a result of the environmental conditions that determine the rate at which virus is transmitted. Interestingly on going experiments at IAH suggest that transmission is almost always around onset of clinical signs and accordingly, the main variable in terms of inter-transmission period would be expected to be the incubation period which can be short or long, but is perhaps also most often around 2-5 days. Thus it is likely that the incubation period is a parameter that contributes to the adherence of the virus evolutionary rate to a molecular clock.

7.3.2 Implications and epidemiological tools

At the start of this PhD project it was acknowledged that an increased understanding of FMDV genetic evolution could potentially provide a framework with which to develop tools applicable to epidemiology and disease control. As a result of this the genetic data generated as part of this thesis have been applied in different epidemiological contexts.

As noted in Chapter 3, the accumulation of mutations into the UK Pan Asia O virus genome correlated with time. This molecular clock allowed us to extrapolate back in time to determine the date on which the virus started replicating in the UK, using coalescent methods. This date was shown to be identical to that estimated from lesion dating. This has established a new, molecular genetic method for dating the start of a FMD outbreak; a useful epidemiological tool.

The use of the genetic data for disease transmission tracing is discussed in depth in Chapter 4, and it is clear that it was possible to trace the transmission of the UK Pan Asia O serotype virus in the UK in 2001. Although genetic data provides high resolution data concerning the relationships between farms, it can be limited in determining exact transmission links between farms. However, integrating epidemiological data concerning the timing of infection supplements the conclusions that can be drawn from the data. The importance of the genetic tracing technique was demonstrated during the 2007 FMDV outbreak of O1BFS virus in Surrey, where it was used to assist in identifying the source of the outbreak and also the order in which farms were infected.

To date tracing of FMDV transmission has been limited to the level of premises-to-premises. Within this thesis it has been shown that animals on a single premises can differ from 1 to 3 nucleotides from each other. A recent paper (Carrillo et al. 2007) shows that nucleotide changes do occur on transmission of virus between pigs in an experimental setting at a rate of 6.4×10^{-4} substitutions per nucleotide per transmission. Although this implies that deciphering animal-to-animal transmission networks will be possible, an important factor to take into consideration is the mode by which animals upon a single farm or epidemiological unit become infected. For example, do many become infected simultaneously or is there a chain of sequential infections? In the case of the former tracing between the animals would not be possible because the virus was not necessarily spread between them. Interestingly one of the IP diagnosed in 2007 in

the UK was shown to have a large proportion of the herd PCR positive for virus (18 out of 48), even though no animal was showing clinical signs of disease. It is usually suggested that FMDV is transmitted from cattle when they are showing clinical signs and not before, and thus these data would suggest that the animals were indeed infected simultaneously. It is likely that both infection scenarios are possible in the field during an outbreak. Another important consideration is whether or not nucleotide substitutions are inherited and therefore apparent within the virus consensus sequence of the recipient animal. A recent study (Carrillo et al. 2007) suggests that not all substitutions are inherited into the consensus virus sequence within the recipient following transmission, although they can re-occur upon subsequent transmission. This suggests that during close contact direct pig-to-pig transmission, the plasticity of the consensus genetic sequence is apparent, fluctuating between animals. However, such fluctuation within the genetic consensus sequence was not seen within the studies on the UK 2001 outbreak reported in this thesis, where nucleotide substitutions were fixed into the population and there was no evidence of reversion. This difference could possibly be a consequence of the mode of transmission; it is possible that herd-to-herd transmission across different premises during the 2001 epidemic represented a more extreme bottleneck situation than direct contact animal-to-animal transmission as studied by Carrillo et al 2007. Due to little data concerning the mode in which a herd of animals are exposed to FMDV and the limited data available concerning the inheritability of nucleotide substitutions, the potential utility of genetic tracing from animal-to-animal remains currently unclear.

The investigation of genetic tracing for FMDV in this study has concentrated on a single cluster of outbreaks in Darlington, County Durham. More studies in different areas and scenarios (i.e. airborne spread) should be done to validate this procedure and there are many more cases from the UK 2001 outbreak that could be investigated. To carry this project further, a new DEFRA funded project has begun, sequencing more viruses from the UK 2001 epidemic.

7.3.3 Further work

Within this thesis, investigation of genetic tracing of FMDV transmission has been limited to studying the behaviour of virus in a naïve country formerly FMDV free. Although beyond the scope of this project, it would be interesting to explore the potential of these techniques in situations where FMDV is endemic. This could alter the way in which the virus evolves, possibly enforcing increased selection pressures as not all animals are naïve, and also allows recombination between viruses to play a greater role. Additionally, it would be interesting to study genetic transmission tracing in FMD viruses of different serotypes, the substitution rate and extents of which are reported to differ. For example, serotype A viruses have been shown to change genetically within the VP1 coding region at a faster rate than serotype O viruses (Haydon et al. 2001).

Finally, it was speculated that missing farms within a network of transmissions could be located if the number of changes expected upon transmission to a new farm, and the rate at which changes accumulate upon a farm, were known. This was investigated in Chapter 4 of this thesis. The main difficulty encountered was in determining how many nucleotide changes should occur on transmission, as a wide variation was found possibly arising from uncertainty surrounding whether or not *all* premises involved had been identified, or possibly due to the varying nature of each farm and its own infection and transmission scenario. Further work on the model is required that incorporates the farm type and characteristics when considering the rate at which mutations are fixed into the consensus sequence of the population. Also characterising further at which point during the virus life cycle nucleotide changes are most likely to be fixed, i.e. upon replication and selection or upon transmission through a bottleneck, would assist in the development of this method. This could include more in depth study of within herd and also within host virus variation.

7.4 FMDV within-host evolution

7.4.1 Observations

With the intention of furthering our understanding of the evolution of FMDV and the mode by which mutations are fixed into the consensus sequence, the virus variability within a population was investigated in Chapter 5 and the spontaneous mutation rate of FMDV was investigated in Chapter 6. The spontaneous mutation rate is an extremely important parameter to quantify accurately as discussed in Chapter 6 section 6.2, and should be investigated further. Although a final result for the spontaneous mutation rate of O1BFS was not achieved, this study represented a pilot study. A considerable proportion of the preliminary work required for the completion of a fluctuation analysis experiment to determine the spontaneous mutation rate was completed.

The extent of genetic variability that surrounds the consensus sequence of the virus genome could potentially impact upon the rate at which mutations are fixed within the consensus sequence. Within the study described in Chapter 5, the genetic variability of the virus within the host animal did not differ from previous *in vitro* estimates of mutant frequency in cell culture (Airaksinen et al. 2003; Arias et al. 2001; Gu et al. 2006; Pariente et al. 2001; Sierra et al. 2000). Evidence was found to suggest that different populations of virus can exhibit different characteristics such as evolutionary rates and *dN/dS* ratios. These differences could possibly be due to the age and size of populations influencing the mutant spectrum characteristics, or host environment.

7.4.2 Implications and prospects for further study

During the study of within population genetic diversity, it was noted that in one particular population investigated that the sequence of an ancestral virus consensus was present as a minor population. This suggests that the change in the consensus sequence apparent in the virus population studied might not be due solely to a bottleneck event on

transmission causing a founder effect, but could also be due to a mutation occurring early on upon replication within the new host. This highlights the requirement to define in greater detail precisely how nucleotide substitutions become fixed into the virus consensus population. Interestingly a UKG virus strain sequenced (using the protocol developed within this thesis, data not shown) from a secondary lesion from a needle challenged cow inoculated with virus that was derived from a pig UKG 34/2001 from Cheale's Abattoir infected in 2001, (epithelium from UKG 34/2001 was used to needle infect 2 cattle, the epithelium from which was ground and combined to form the inoculum for the needle challenged cow) was found to be identical to virus sequenced from another pig from the same abattoir UKG 35/2001 that was passaged twice in pigs prior to sequencing (Mason et al. 2003b). The fact that no nucleotide changes were fixed within the consensus through any of these needle challenge passages would corroborate the view that bottleneck events or selection play an important role in the fixation of mutations into the virus consensus sequence such that they are inherited.

Further study of the nature in which the virus population changes and evolves (for example, detailed clonal sequence analysis of virus populations within the natural host and also the down stream recipient) could provide the opportunity to apply coalescent methodology and determine the size of bottleneck that occurs upon transmission. As discussed earlier the size of bottleneck appears to be an important factor in the rate at which nucleotide substitutions are fixed into the consensus population. An interesting study (Edwards et al. 2006) applied coalescent theory to HIV-1 and determined that >99% of diversity is lost upon transmission. If the extent of bottlenecking is known, then it will assist estimates of virus intra-population diversity from data concerning the rate of consensus sequence nucleotide evolution. Additionally it may enable estimates of the degree of virus genetic heterogeneity that is compatible with defined amounts of consensus sequence evolution.

Many studies on other viruses have found correlations between nucleotide diversity within the virus population and specific disease pathogenicities such as virus persistence (see Chapter 5: section 5.1). It is feasible that this could also be important for FMDV, and the extent of genetic diversity could play a role in virus persistence in some hosts and not others, for example. On finding within this thesis that virus populations within hosts do differ in characteristics and diversity (perhaps due to age or size of population or animal host), it is essential to take these studies further to compliment our current understanding of FMDV disease dynamics.

7.5 Conclusion

As demonstrated in previous studies, and within this thesis, viral genetic diversity can be measured at defined levels; within host cells or tissues, between host groups (herds and farms), between regional populations, and also at an international scale. Although, in order to understand transmission, epidemiologists are primarily concerned with the larger scale distribution of virus within and between host populations, it is important to be aware that the processes that give rise to the virus diversity seen at higher levels inevitably occur at far smaller scales within a single host. At each distinct level at which viral genetic diversity can and has been studied, various processes operate to determine the distribution and abundance of viral genotypes. Remarkably little is known about the interplay of these processes that result in the patterns of genetic diversity observed at a larger scale, and it is within this area that future research should be focused. Modelling the population genetic processes characteristic of viral transmission systems (supported by empirical data that is increasingly available) at all levels, from within cells to international spread of disease (both previously studied in isolation), could potentially inform epidemiologists of virus genetic variation and transmission and as such improve the design and implementation of methods for disease control.

Chapter 8

The use of complete genome genetic tracing of FMDV in real-time during the 2007 O_1BFS outbreaks in south-east England

This work has been submitted to Science;

Transmission pathways of the UK 2007 foot-and-mouth disease virus outbreaks

Eleanor M. Cottam[1], Jemma Wadsworth[1], Andrew E. Shaw[1], Rebecca J. Rowlands[1], Lynnette Goatley[1], Sushila Maan[1], Narender S. Maan[1], Katja Ebert[1], Eoin D. Ryan[1], Nicholas Juleff[1], Nigel P. Ferris[1], John W. Wilesmith[2], Daniel T. Haydon[3], Donald P. King[1], David J. Paton[1] and Nick J. Knowles[1]

1. Institute for Animal Health, Pirbright Laboratory, Ash Road, Pirbright, Woking, Surrey, GU24 0NF, United Kingdom.
2. Animal Health and Welfare, Defra, 1a Page Street, London SW1P 4PQ, United Kingdom.
3. Division of Environmental and Evolutionary Biology, University of Glasgow, Glasgow G12 8QQ, UK.

8.1 Summary

Foot-and-mouth disease virus (FMDV) (family *Picornaviridae*, genus *Aphthovirus*) causes an acute vesicular disease of domesticated and wild ruminants and pigs. Identifying sources of FMD outbreaks is often confounded by incomplete epidemiological evidence and the numerous routes by which virus can spread (movements of infected animals or their products, contaminated persons, objects and aerosols). Here, we show that the outbreaks of FMD in the UK in August 2007 were caused by a derivative of FMDV O_1 BFS 1860, a virus strain handled at two FMD laboratories located on a single site at Pirbright. Genetic analysis of complete viral genomes generated in real-time reveals a chain of transmission events, predicting undisclosed infected premises, and connecting the second cluster of outbreaks in September to those in August.

8.2 Introduction

Foot-and-mouth disease (FMD) outbreaks in the UK during August and September 2007 have caused severe disruption to the farming sector and cost hundreds of millions of pounds. Investigating and determining the source of these outbreaks is imperative for their effective management and future prevention. Foot-and-mouth disease virus (FMDV) has a high mutation rate resulting in rapid genetic evolution (Domingo et al. 2002) and genome sequencing has been used to retrospectively trace virus spread (Cottam et al. 2006). We now show how complete genome sequences (acquired within 24-48 h of sample receipt) can be used to track FMDV movement from farm-to-farm in real-time. This helped to determine the most likely source of the outbreak, assisted ongoing epidemiological investigations as to whether these field cases were linked to single or multiple releases from the source, and predicted the existence of undetected intermediate infected premises.

The UK 2007 FMD outbreaks have been characterised by the emergence of two temporally and spatially distinct clusters. Eight infected premises (IP1-8: designation of IP numbering is according to The Department for Environment, Food and Rural Affairs [Defra], UK) have been identified (Figure 8.1 and Table 8.1), two in August and six in September. The first case (IP1b) was recognised in beef cattle in a field in Westwood Lane, Normandy, Surrey, UK. Samples collected on 3rd August 2007 from animals exhibiting suspect clinical signs were submitted to the World Reference Laboratory for FMD located at the Institute for Animal Health (IAH), Pirbright, Surrey. Within 24 hours, FMDV sequence data obtained from the first IP (holding IP1b) revealed a VP1 gene-identity of 99.84% to FMDV O_1 British Field Sample 1860 (O_1 BFS 1860) (Knowles & Samuel 2003). O_1BFS 1860 is a widely used reference and vaccine strain, originally derived from bovine tongue epithelium received at the World Reference Laboratory for FMD at Pirbright in 1967 from a farm near Wrexham, England. The Pirbright site,

comprising the laboratories of the IAH and Merial Animal Health Limited (Merial), is situated 4.4 km from the first IP. Both laboratories were working with the O_1 BFS 1860 virus strain, making this site a likely source of the outbreak. Three days after the case at IP1b, a second infected premises (IP2b) was identified at Willey Green, approximately 1.5 Km from IP1b. Cattle at a further holding (IP2c) near to and under the same ownership as IP2b were found to be incubating disease at the time of slaughter. Animals on both the affected farms were destroyed and the premises were disinfected. Subsequent clinical and serological surveillance within a 10 km control zone found no evidence of further dissemination of FMD. However, on 12th September 2007, five weeks after the IP1 and IP2 cattle had been culled, FMD was confirmed on the holding of a new IP (IP3b) situated outside the 10 km control zone surrounding IP1 and IP2 (Figure 8.1). FMD outbreaks were subsequently reported on an additional holding (IP3c) and five more premises (IP4, 5, 6, 7 and 8) all located close to IP3b and outside the original surveillance area (Figure 8.1).

These outbreaks of FMDV in the UK during August and September 2007 have caused severe disruption to the farming sector and cost hundreds of millions of pounds. Investigating and determining the source of these outbreaks has been imperative for their effective management and is vital for future prevention. The aim of this study was to trace FMDV movement from farm-to-farm by comparing complete genome sequences acquired during the course of the epidemic. These “real-time” analyses helped to determine the most likely source of the outbreak, assisted ongoing epidemiological investigations as to whether these field cases were linked to single or multiple releases from the source, and predicted the existence of undetected intermediate infected premises that were subsequently identified.

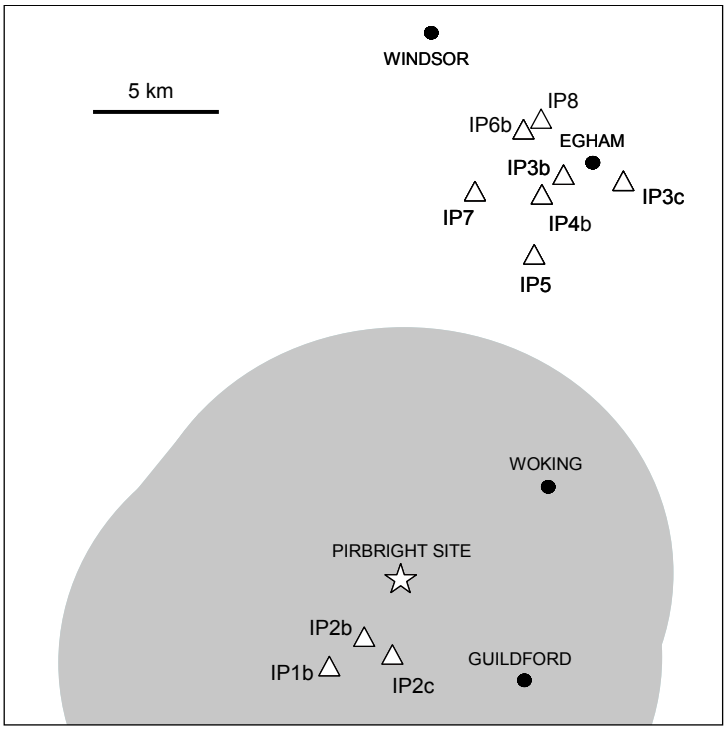


Figure 8.1

The geographical location of FMDV infected premises and holdings (Δ with official numbering) in August and September 2007 in Surrey (UK). The grey shaded area denotes the extent of the 10 km control and surveillance zone established for the first phase of the outbreak. Cases that occurred in September (IP3b, IP3c, IP4b, IP5, IP6b and IP7) were outside of this zone. The location of the Pirbright site is indicated by the star.

Table 8.1 Details of FMDV infected premises and the virus isolates that were sequenced

IP/holding designation	Date collected	WRLFMD Ref. No.	GenBank accession no.	Date animals examined	Species sampled	Sample	Estimated lesion age of individual (days)	Oldest lesion on farm (days)	Affected/total stock
IP1b	03/08/2007	UKG/7/2007		04/08/2007	bovine	epithelium	6	10	38/38 cattle
	04/08/2007	UKG/7B/2007			bovine	epithelium	3		
IP2b	06/08/2007	UKG/93/2007		07/08/2007	bovine	epithelium	6	7	44/49 cattle
IP2c	07/08/2007	UKG/150/2007		07/08/2007	bovine	blood	none	none	0/58 cattle
IP3b	12/09/2007	UKG/643/2007		12/09/2007	bovine	epithelium	4-5	5	36/47 cattle
IP3c	15/09/2007	UKG/1153/2007		15/09/2007	bovine	blood?	not known	5	9/15 cattle
IP4b	13/09/2007	UKG/800/2007		13/09/2007	bovine	epithelium	5-7	10	54/54 cattle
IP5	17/09/2007	UKG/1421/2007		17/09/2007	ovine	oesophageal/ pharyngeal scrapings	not known	21*	10/16 sheep; 16/24 cattle; 0/2 pigs
IP6b	21/09/2007	UKG/1484/2007		21/09/2007	bovine	epithelium	2-4	4	2/32 cattle
IP7	24/09/2007	UKG/1679/2007		24/09/2007	bovine	epithelium	1-2	5	14/16 cattle
IP8	29/09/2007	UKG/2366/2007		29/09/2007	bovine	epithelium	2	2	1/54

The Pirbright site, comprising the laboratories of the Institute for Animal Health (IAH) and Merial Animal Health Limited (Merial), is situated 4.4 km from the first IP. Both laboratories were working with the O_1 BFS 1860 virus strain, making this site a likely source of the outbreak. Complete genome sequences of viruses from the Pirbright laboratories and those from the outbreak were studied in real time to assist with epidemiological investigation and control of the outbreak.

8.3 Materials and Methods

Total RNA was extracted directly from a 10% epithelial suspension using the RNeasy Mini Kit (Qiagen, Crawley, West Sussex), or from blood or oesophageal/pharyngeal scrapings using TRIzol (Invitrogen, Paisley, UK). Reverse transcription of the RNA was performed using Superscript III reverse transcriptase (Invitrogen) and an oligo-dT primer (see table of primers, supplementary information). Twenty four PCR reactions per genome were performed with Platinum Taq Hi-Fidelity (Invitrogen), using 23 primer sets tagged with forward and reverse M13 universal primer sequences, and one primer set with a oligo-dT reverse primer to obtain the very 3' end genomic sequence (Table 8.2). The PCR products overlap such that each nucleotide is covered by two products. The reactions were run on a thermal cycling programme of 94ºC for 2 min, followed by 40 cycles of 94ºC for 30 s, 55ºC for 30 s, 72ºC for 1 min, with final step at 72ºC for 7 min. Sequencing reactions were performed using the Beckman DTCS kit, with M13 universal forward and reverse primers and specific forward and reverse primers for each PCR product. This resulted in an average of 7.4 times coverage of each base. The raw data was assembled using the Lasergene® 7 Software package (DNASTAR, Madison, WI) and all further sequence manipulations were performed using BioEdit (version 7.0.1 (Hall 1999)) and DNAsp (version 3.52 (Rozas & Rozas 1999)). The data were analysed by statistical parsimony methods (Templeton et al. 1992) incorporated in the TCS freeware (Clement et al. 2000). A Bayesian majority rule consensus tree (based on 10,000 trees

sampled from 10 million generations) was estimated in MrBayes (Huelsenbeck & Ronquist 2001) assuming a General Time Reversible model of nucleotide substitution with invariant sites (the model most strongly supported by more extensive genome data from the UK 2001 outbreak).

Table 8.2 Oligonucleotide primers used for the amplification and sequencing of the FMDV genomes studied (designed by NJ Knowles).

PCR set	Primer name*	Primer sequence (5' to 3')†	Location on genome‡	Amplicon size (bp)§
1	BFS-1F	TTGAAAGGGGGCRCTAG	1-17	413
	BFS-379R	GGGGGGGGGGGGTGAAA	363-379	
2	BFS-370F	CCCCCCCCCCCCCTAAG	370-386	735
	BFS-1070R	AGGTGCCGGCCTCCGGT	1054-1070	
3	BFS-720F	TTTGACTCCACGCTCGA	720-736	735
	BFS-1420R	GTGCTTGATGTTCCAGATC	1402-1420	
4	BFS-1068F	CCTTTCCTTTTATAACCACTGA	1068-1089	735
	BFS-1768R	GTTGTTTATTATGCTGCCAG	1749-1768	
5	BFS-1418F	CACTTGCTCCACACCGGC	1418-1435	735
	BFS-2118R	ACAACTCTGGTCTCGAGA	2101-2118	
6	BFS-1773F	ACATGCAGCAGTATCAAAAC	1773-1792	735
	BFS-2473R	GTGAACCTTGTACTGGTCG	2455-2473	
7	BFS-2123F	GCAGAACGGTTTTTCAAAAC	2123-2142	735
	BFS-2823R	ACCCTGTCCGAGTCTGTT	2806-2823	
8	BFS-2476F	GCCTTGGACCCTTGTGGTT	2476-2494	735
	BFS-3176R	GTGTAATTTGAAACAAGCAG	3157-3176	
9	BFS-2826F	TTGCTCAGTTTGATATGTCT	2826-2845	735
	BFS-3526R	GGTGAGGTCTCCCTCGTG	3509-3526	
10	BFS-3176F	CATGGCAAGGCCGACGGC	3176-3193	735
	BFS-3876R	TTCTGTTTGTGTCTGGCTTCA	3856-3876	
11	BFS-3526F	CTGGGTTCCAAATGGAGC	3526-3543	741
	BFS-4232R	CGGTGTCGGCCAGCATGAT	4214-4232	
12	BFS-3876F	AAATTGTGGCACCGGTGA	3876-3893	735
	BFS-4576R	CTGCTTTTCAAGGATGCCAG	4557-4576	
13	BFS-4226F	GACACCGGTCTCGAGATT	4226-4243	735
	BFS-4926R	AAGTCCTTGCCGTCAGGGT	4908-4926	
14	BFS-4576F	GCGGGATCTGAACGACC	4576-4592	734
	BFS-5275R	TTGAGGCTTGAACATGTCTTG	5255-5275	
15	BFS-4929F	AATACTTTGCCCAAATGGTCT	4929-4949	732
	BFS-5626R	CACCATTTTCTGCCTCTTG	5608-5626	
16	BFS-5276F	CCACCCCTCCAGAATGTG	5276-5293	725
	BFS-5966R	CCTTCACGACCGGGGCT	5950-5966	
17	BFS-5626F	GGATGATGCAGTGAATGAG	5626-5644	735
	BFS-6326R	CGCGGTTCCCACGGTGGA	6309-6326	
18	BFS-5976F	ACGAGGGACCGGTGAAGA	5976-5993	735
	BFS-6676R	GTGATGTGGTTCGGGGTC	6659-6676	
19	BFS-6328F	GAGGGACATCACGAAGCAC	6328-6346	732
	BFS-7025R	CTGTRTCTGGTTCCATGGC	7007-7025	
20	BFS-6676F	CGAGGGGTTGATTGTGGA	6676-6693	735

	BFS-7376R	ACTGGGCGAAGTGTGTGC	7359-7376	
21	BFS-7026F	CGCCCGGCCTCCCCTGG	7026-7042	742
	BFS-7733R	CCTCGAAGTCCAAGTCATAAT	7713-7733	
22	BFS-7376F	TACAGAAACGTGTGGGATGT	7376-7395	624
	BFS-7965R	CCTGCCACGGAGATCAACTTCT	7944-7965	
23	BFS-7685F	TACGGAGACGACATCGTG	7685-7702	543
	BFS-8193R	TTTTTTTTTTTTTTGATTAAGG	8172-8193	
24	BFS-7685F	TACGGAGACGACATCGTG	7685-7702	536
	UKFMD/Rev6	GGCGGCCGCTTTTTTTTTTTTTT	poly(A)	
-	Uni-1F	GTAAAACGACGGCCAGT	-	
-	Uni-1R	CAGGAAACAGCTATGAC	-	

*, the last letter indicates a Forward or Reverse primer.

†, a duplicate set of primers (except for UKFMD/Rev6) were synthesized which were preceeded by M13 forward (Uni-1F) or reverse (Uni-1R) primer sequences.

‡, numbering according to GenBank sequence AY593815.

§, amplicon size includes 34 bp derived from the M13 sequences at the 5' end of each primer.

8.4 Results and Discussion

The genetic relationships of FMDV present in eleven field samples from the 2007 outbreak, three cell culture derived laboratory viruses (see Table 8.3) used at the Pirbright site during July 2007 (designated IAH1, IAH2 and MAH) and a published sequence of O_1 BFS 1860 (AY593815) are illustrated in Figure 8.2a . Whereas IAH1 and the virus from which the published sequence was derived are believed to have been passaged no more than ten times in cell cultures, the IAH2 and MAH viruses had been extensively adapted to grow in a baby hamster kidney cell line (Table 8.3). In natural hosts, FMDV attaches to integrin receptors on the cell surface (Jackson et al. 2000). However, when grown in cell cultures, the virus may adapt to attach to heparan sulphate (HS), through acquisition of positively charged amino acid residues on the virus coat at positions $VP2^{134}$ and $VP3^{56}$ (Fry et al. 1999; Sa-Carvalho et al. 1997). An additional change from a negatively charged amino acid residue at $VP3^{60}$ to a neutral residue often occurs but may not be essential for HS binding (Sa-Carvalho et al. 1997). IAH1 and the previously sequenced isolate of O_1 BFS 1860 have none of the changes associated with HS binding, whereas the changes at $VP3^{56}$ and $VP3^{60}$ are present in MAH and IAH2, consistent with their history of extensive culture passage (Table 8.3). The presence of the HS binding residue at $VP3^{60}$ in all but one of the field viruses provides evidence that a cell culture adapted virus is an ancestor of the outbreak. The wild type configurations at $VP3^{56}$ in all of the outbreak viruses and at $VP3^{60}$ in the IP5 virus most likely reflect reversions that have been selected upon replication within the animal host. It is known that there is a strong selection pressure for the reversion at $VP3^{56}$ when FMDV replicates in cattle (Sa-Carvalho et al. 1997).

Table 8.3 Passage histories of the reference viruses studied.

Identifier	Passage history	Genbank accession no.	Origin
O_1 BFS 1860	Presumed CP1, BTy3, LK1	AY593815	Carrillo et al., 2005
O_1 BFS 1860 (IAH1)	BTy4, BHK2		In use at the IAH.
O_1 BFS 1860 (IAH2)	Between 18 and 24 passages in BHK cells including at least one passage in suspension cells.		In use at the IAH; derived from the Wellcome vaccine strain.
O_1 BFS 1860 (Merial)	Unknown		Vaccine seed virus used by Merial Animal Health Ltd.

CP, cattle passage

BTy, primary bovine thyroid cells

BHK, BHK-21 (baby hamster kidney) cells

LK, lamb kidney cells.

The viruses from the outbreaks differ by at least five unique synonymous substitutions from the laboratory viruses examined (Table 8.4, Figure 8.2A). In terms of nucleotide substitutions, two very closely related laboratory viruses (MAH and IAH2) are closest to the sequence of the virus from IP1b (6 and 7 substitutions respectively) compared with IAH1 (12 substitutions). Viruses IAH2 and MAH differ by only one non-synonymous change at amino acid residue 2 of the Leader-b (Lb) polypeptide (a papain-like cysteine proteinase) (Table 8.4). Since FMDV is known to exist as variant populations of genetically related viruses (Domingo et al. 2002), it is possible that virus containing the MAH consensus sequence was present as a minority component within the virus population of IAH2. It is also possible that a reversion of the amino acid change at residue 2 of Lb could be selected when the virus goes back into the natural host. Consequently, either of these viruses could be the source of the 2007 outbreak.

Table 8.4 Nucleotide and amino acid substitutions observed in the genomes of the FMD viruses studied

Position*	AY593815	IAH1	IAH2	MAH	UKG/7B/2007 (IP1b)	UKG/7/2007 (IP1b)	UKG/93/2007 (IP2b)	UIKG/150/2007 (IP2c)	UKG/643/2007 (IP3b)	UKG/1153/2007 (IP3c)	UKG/800/2007 (IP4b)	UKG/1421/2007 (IP5)	UKG/1484/2007 (IP6)	UKG/1679/2007 (IP7)	UKG/2366/2007 (IP8)	Gene	Codon no. in gene†	aa / codon in AY593815	Substituted aa / codon
159	C	C	C	C	C	C	C	C	C	C	C	C	C	C	A	5' UTR	-	-	-
170	C	C	C	C	T	T	T	T	T	T	T	T	T	T	T	5' UTR	-	-	-
262	T	T	T	T	T	T	T	T	T	T	T	T	T	C	T	5' UTR	-	-	-
792	G	G	G	G	G	G	G	A	G	G	G	G	G	G	G	5' UTR	-	-	-
913	A	A	A	A	A	A	A	A	T	T	T	A	T	T	T	5' UTR	-	-	-
1083	C	C	A	A	A	A	A	A	A	A	A	A	A	A	A	5' UTR	-	-	-
1181	G	G	A	G	G	G	G	G	G	G	G	G	G	G	G	Lb	2	E	K (aaa)
1951	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	VP4	-	-	-
2184	G	A	G	G	G	G	G	G	G	G	G	G	G	G	G	VP2	78	C (tgc)	Y (tac)
2377	T	T	T	T	C	C	C	C	C	C	C	C	C	C	C	VP2	-	-	-
2446	C	C	C	C	C	C	C	C	T	T	T	T	T	T	T	VP2	-	-	-
2558	A	A	A	A	A	A	G	A	A	A	A	A	A	A	A	VP2	203	I (att)	V (gtt)
2772	A	A	G	G	A	A	A	A	A	A	A	A	A	A	A	VP3	56	H	R (cgc)
2784	A	A	G	G	G	G	G	G	G	G	G	A	G	G	G	VP3	60	D	G (ggc)
3043	A	A	A	A	A	A	A	G	A	A	A	A	A	A	A	VP3	-	-	-
3157	C	C	C	C	C	C	C	C	C	T	C	C	C	C	C	VP3	-	-	-
3433	T	T	T	T	C	C	C	C	C	C	C	C	C	C	C	VP1	-	-	-
3862	C	C	C	C	C	C	C	C	C	C	C	C	C	C	T	VP1	-	-	-
3994	G	G	G	G	G	G	G	G	G	G	G	G	G	A	A	2B	-	-	-
4045	T	T	T	T	T	T	Y	T	T	T	T	T	T	T	T	2B	-	-	-
4407	A	A	A	A	A	A	A	A	A	A	A	A	A	A	G	2B	152	E	G (ggg)
4639	A	A	A	A	A	A	A	A	G	A	A	A	G	G	G	2C	-	-	-
4927	C	C	C	C	C	C	C	C	C	C	C	T	C	C	C	2C	-	-	-
5152	C	C	C	C	C	C	C	C	C	C	C	Y	C	C	C	2C	-	-	-
5335	C	C	C	C	T	T	T	T	T	T	T	T	T	T	T	2C	-	-	-
5462	G	G	G	G	G	G	G	G	G	G	G	G	G	A	A	3A	32	D	N (aac)
5473	A	A	A	A	A	A	A	A	G	G	G	G	G	G	G	3A	-	-	-
5492	A	A	C	C	C	C	C	C	C	C	C	C	C	C	C	3A	42	I (atc)	L (ctc)
5592	C	T	T	T	T	T	T	T	T	T	T	T	T	T	T	3A	75	T(	M (atg)
5599	T	C	T	T	T	T	T	T	T	T	T	T	T	T	T	3A	-	-	-
5655	A	A	A	A	A	A	A	A	A	A	A	A	G	A	A	3A	96	K	R (aga)
5808	A	A	A	A	A	A	A	A	A	A	A	A	G	A	A	3A	147	E	G (ggg)
6175	T	T	T	T	T	T	T	T	C	T	T	T	C	C	C	3C	-	-	-
6616	A	A	A	A	A	A	A	A	A	A	A	A	A	A	G	3C	-	-	-
6673	T	T	T	T	T	T	T	T	C	C	C	T	C	C	C	3C	-	-	-
7021	C	C	T	T	T	T	T	T	T	T	T	T	T	T	T	3D	-	-	-
7162	C	C	C	C	C	C	C	C	T	C	C	C	T	T	T	3D	-	-	-
7729	C	C	C	C	C	C	C	C	T	T	T	C	T	T	T	3D	-	-	-
7750	T	T	T	T	T	T	T	T	C	C	C	T	C	C	C	3D	-	-	-
8149	A	A	G	G	G	G	G	G	G	G	G	G	G	G	G	3' UTR	-	-	-
8152	C	C	C	C	T	T	T	T	T	T	T	T	T	T	T	3' UTR	-	-	-
8176	A	A	A	A	A	A	A	A	A	A	A	G	A	A	A	3' UTR	-	-	-
8180	A	A	A	A	A	A	T	A	T	T	T	T	T	T	T	poly(A	-	-	-

*, numbered according to AY593815; the following positions were not sequenced: 1-17and 363-466

†, only listed where an amino acid substitution has occurred.

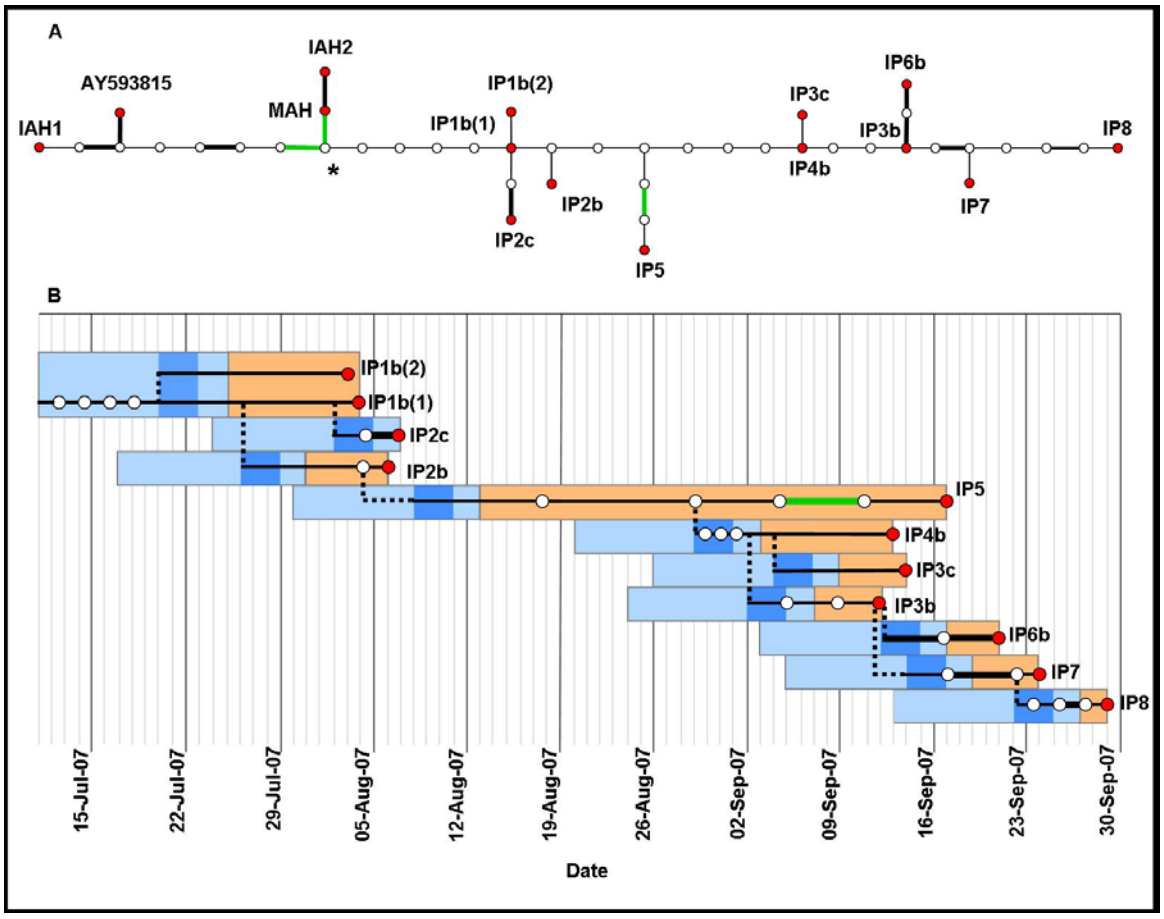


Figure 8.2

A) Statistical parsimony analysis by TCS (Clement et al. 2000) of complete genome sequences of four laboratory and 10 outbreak FMDV; connecting lines represent a nucleotide substitution, thicker lines represent non-synonymous substitutions, with substitutions indicative of adaptation to cell culture coloured green. Sequenced haplotypes (●), and putative ancestral virus haplotypes (○) are shown. AY593815 is a previously published sequence (Carrillo et al. 2005) of FMDV O_1 BFS 1860.

B) Lesion age derived infection profiles of holdings overlaid with the outbreak virus geneology. The orange shading estimates the time when animals with lesions were present from the oldest lesion age at post-mortem (Anon 1986). For IP2c, there were no clinical signs of disease. The light blue shading represents incubation periods for each holding, estimated to begin no more than 14 days prior to appearance of lesions (Garland & Donaldson 1990). The dark blue shading is the most likely infection date based on the average incubation time for this strain of 2-5 days (Hugh-Jones & Tinline 1976). Each UK 2007 outbreak virus haplotype is plotted according to the time it was isolated (x axis). The dashed lines link the TCS tree together but do not denote any genetic change.

Sequence analysis of virus from the first affected holding identified in the second cluster of outbreaks (IP3b) demonstrated that it had evolved from virus from the first cluster of outbreaks (Figure 8.2A and B). The sequence data are not consistent with a second escape of virus from the Pirbright site, as the virus from IP3b shares five common nucleotide changes with IP1b and IP2c and six in common with IP2b. A Bayesian majority rule consensus tree, Figure 8.3, estimated in MrBayes (Huelsenbeck & Ronquist 2001) indicated that the group linking the second cluster of outbreaks to the first is strongly supported with a posterior probability of 1.00.

The virus from IP3b was nine nucleotides different from the virus from IP1b (Table 1, Figure 8.2a). This is a high number of changes for a single farm-to-farm transmission [a retrospective study of virus genomes acquired from sequentially infected farms during the UK 2001 outbreak in Darlington, County Durham, found a mean of 4.5 (SD 2.1) nucleotide changes, see Chapter 4], and indicated that there were likely to be intermediate undetected infected premises between the first outbreaks in August and IP3b. Subsequent field investigations discovered IP4b and IP3c, which differed by one nucleotide from each other. IP4b was three nucleotides closer to virus from the first outbreaks, and IP3c also branched off the tree at this point. However, there were still six nucleotide differences between FMDV sourced from IP4b and FMDV sourced from the August outbreaks. Serosurveillance of all sheep within 3 km of the September outbreaks revealed the existence of IP5, on which it was estimated that disease had been present for at least two, and possibly up to five weeks. As Figure 8.2b shows, IP5 is a likely link between the August and September outbreaks.

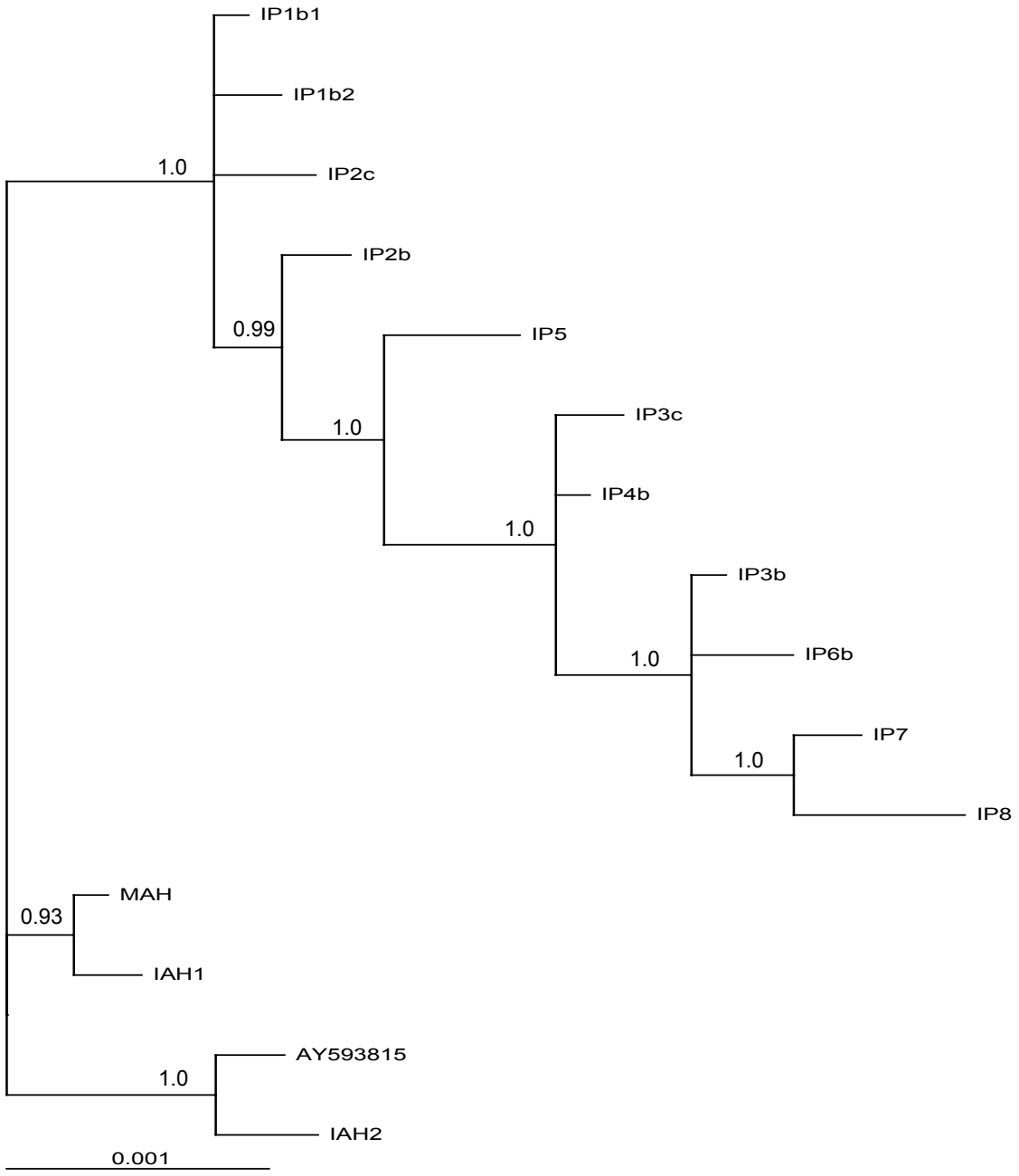


Figure 8.3

A Bayesian majority rule consensus tree of eleven viruses from the UK2007 outbreaks and four laboratory viruses, labelled as described in figure legend for figure 8.2.

The genetic relationships between viruses from individual animals shown in Figure 8.2A and B follows an identical topology to the Bayesian majority rule consensus tree (Figure 8.3) and ingroup relationships are strongly supported by posterior probabilities on

genome groupings that were never less than 99%. Although a more confident resolution of the IP-to-IP transmission pathways might be achieved by characterising additional virus haplotypes present on individual holdings, previous sequencing of virus from different animals from the same farm conducted following the UK 2001 outbreaks indicated very limited intra-farm sequence variability. Furthermore, the relationships presented here reveal a transmission pathway between outbreaks that is consistent with the estimates of when holdings became infected and infectious (Figure 8.2B). The small number of nucleotide substitutions observed between viruses from source and recipient IPs suggests that there has been direct transmission without the involvement of other susceptible species, e.g. sheep or deer.

References

Acharya, R., Fry, E., Stuart, D., Fox, G., Rowlands, D. & Brown, F. 1989 The three-dimensional structure of foot-and-mouth disease virus at 2.9 A resolution. *Nature* **337**, 709-16.

Airaksinen, A., Pariente, N., Menendez-Arias, L. & Domingo, E. 2003 Curing of foot-and-mouth disease virus from persistently infected cells by ribavirin involves enhanced mutagenesis. *Virology* **311**, 339-49.

Albert, J., Wahlberg, J., Leitner, T., Escanilla, D. & Uhlen, M. 1994 Analysis of a rape case by direct sequencing of the human immunodeficiency virus type 1 pol and gag genes. *J Virol* **68**, 5918-24.

Alexandersen, S., Kitching, P., Mansley, L. M. & Donaldson, A. I. 2003a Clinical and laboratory investigations of five outbreaks of foot-and mouth disease during the 2001 epidemic in the United Kingdom. *Vet Rec* **152**, 489-496.

Alexandersen, S., Quan, M., Murphy, C., Knight, J. & Zhang, Z. 2003b 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**, 268-82.

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

Alexandersen, S., Zhang, Z., Reid, S. M., Hutchings, G. H. & Donaldson, A. I. 2002 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**, 1915-1923.

Alfonso, V., Mbayed, V. A., Sookoian, S. & Campos, R. H. 2005 Intra-host evolutionary dynamics of hepatitis C virus E2 in treated patients. *J Gen Virol* **86**, 2781-6.

Anderson, I. 2002 *Foot and Mouth Disease 2001: Lessons to be Learned Inquiry*. London: The Stationary Office.

Anon. 1986 Foot-and-mouth disease ageing of lesions, vol. Reference book 400 (ed. MAFF): Her Majesty's Stationary Office.

Arezi, B. & Hogrefe, H. H. 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**, 84-91.

Arias, A., Lazaro, E., Escarmis, C. & Domingo, E. 2001 Molecular intermediates of fitness gain of an RNA virus: characterization of a mutant spectrum by biological and molecular cloning. *J Gen Virol* **82**, 1049-60.

Arias, A., Ruiz-Jarabo, C. M., Escarmis, C. & Domingo, E. 2004 Fitness increase of memory genomes in a viral quasispecies. *J Mol Biol* **339**, 405-12.

Back, N. K., Nijhuis, M., Keulen, W., Boucher, C. A., Oude Essink, B. O., van Kuilenburg, A. B., van Gennip, A. H. & Berkhout, B. 1996 Reduced replication of 3TC-resistant HIV-1 variants in primary cells due to a processivity defect of the reverse transcriptase enzyme. *EMBO J* **15**, 4040-9.

Banks, J., Speidel, E. S., Moore, E., Plowright, L., Piccirillo, A., Capua, I., Cordioli, P., Fioretti, A. & Alexander, D. J. 2001 Changes in the haemagglutinin and the neuraminidase genes prior to the emergence of highly pathogenic H7N1 avian influenza viruses in Italy. *Arch Virol* **146**, 963-73.

Barrero, P. R. & Mistchenko, A. S. 2004 Complete genome sequencing of dengue virus type 1 isolated in Buenos Aires, Argentina. *Virus Res* **101**, 135-145.

Bastos, A. D., Haydon, D. T., Sangare, O., Boshoff, C. I., Edrich, J. L. & Thomson, G. R. 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**, 1595-606.

Bedard, K. M. & Semler, B. L. 2004 Regulation of picornavirus gene expression. *Microbes and Infection* **6**, 702-713.

Bernardi, F., Nadin-Davis, S. A., Wandeler, A. I., Armstrong, J., Gomes, A. A., Lima, F. S., Nogueira, F. R. & Ito, F. H. 2005 Antigenic and genetic characterization of rabies viruses isolated from domestic and wild animals of Brazil identifies the hoary fox as a rabies reservoir. *J Gen Virol* **86**, 3153-62.

Borrego, B., Novella, I. S., Giralt, E., Andreu, D. & Domingo, E. 1993 Distinct repertoire of antigenic variants of foot-and-mouth disease virus in the presence or absence of immune selection. *J Virol* **67**, 6071-9.

Bracho, M. A., Gosalbes, M. J., Blasco, D., Moya, A. & Gonzalez-Candelas, F. 2005 Molecular epidemiology of a hepatitis C virus outbreak in a hemodialysis unit. *J Clin Microbiol* **43**, 2750-5.

Bronsvoort, B. M., Radford, A. D., Tanya, V. N., Nfon, C., Kitching, R. P. & Morgan, K. L. 2004 Molecular epidemiology of foot-and-mouth disease viruses in the Adamawa province of Cameroon. *J Clin Microbiol* **42**, 2186-96.

Carrillo, C., Borca, M., Moore, D., Morgan, D. & Sobrino, F. 1998 In vivo analysis of the stability and fitness of variants recovered from foot-and-mouth disease virus quasispecies. *J Gen Virol* **79**, 1699-1706.

Carrillo, C., Lu, Z., Borca, M. V., Vagnozzi, A., Kutish, G. F. & Rock, D. L. 2007 Genetic and Phenotypic Variation of FMDV During Serial Passages in a Natural Host. *J Virol*.

Carrillo, C., Plana, J., Mascarella, R., Bergada, J. & Sobrino, F. 1990 Genetic and phenotypic variability during replication of foot-and-mouth disease virus in swine. *Virology* **179**, 890-2.

Carrillo, C., Tulman, E. R., Delhon, G., Lu, Z., Carreno, A., Vagnozzi, A., Kutish, G. F. & Rock, D. L. 2005 Comparative Genomics of Foot-and-Mouth Disaease Virus. *J Virol* **79**, 6487-6504.

Charpentier, N., Davila, M., Domingo, E. & Escarmis, C. 1996 Long-term, large-population passage of aphthovirus can generate and amplify defective noninterfering particles deleted in the leader protease gene. *Virology* **223**, 10-8.

Chen, W., Yan, W., Du, Q., Fei, L., Liu, M., Ni, Z., Sheng, Z. & Zheng, Z. 2004 RNA interference targeting VP1 inhibits foot-and-mouth disease virus replication in BHK-21 cells and suckling mice. *J Virol* **78**, 6900-7.

Christensen, L. S., Normann, P., Thykier-Nielsen, S., Sorensen, J. H., de Stricker, K. & Rosenorn, S. 2005 Analysis of the epidemiological dynamics during the 1982-1983 epidemic of foot-and-mouth disease in Denmark based on molecular high-resolution strain identification. *J Gen Virol* **86**, 2577-84.

Clement, M., Posada, D. & Crandall, K. A. 2000 TCS: a computer program to estimate gene genealogies. *Mol Ecol* **9**, 1657-9.

Cottam, E. M., Haydon, D. T., Paton, D. J., Gloster, J., Wilesmith, J. W., Ferris, N. P., Hutchings, G. H. & King, D. P. 2006 Molecular Epidemiology of the Foot-and-Mouth Disease Virus Outbreak in the United Kingdom in 2001. *J Virol* **80**, 11274-82.

Dallas, A. & Vlassov, A. V. 2006 RNAi: a novel antisense technology and its therapeutic potential. *Med Sci Monit* **12**, RA67-74.

de la Torre, J. C. & Holland, J. J. 1990 RNA virus quasispecies populations can suppress vastly superior mutant progeny. *J Virol* **64**, 6278-81.

de la Torre, J. C., Martinez-Salas, E., Diez, J., Villaverde, A., Gebauer, F., Rocha, E., Davila, M. & Domingo, E. 1988 Coevolution of cells and viruses in a persistent infection of foot-and-mouth disease virus in cell culture. *J Virol* **62**, 2050-8.

Diez, J., Mateu, M. G. & Domingo, E. 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)**, 3281-9.

Domingo, E. 2002 Quasispecies Theory in Virology. *J Virol* **76**, 463-465.

Domingo, E., Baranowski, E., Escarmis, C. & Sobrino, F. 2002 Foot-and-mouth disease virus. *Comparative Immunology, Microbiology and Infectious Diseases* **25**, 297-308.

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

Domingo, E., Pariente, N., Airaksinen, A., Gonzalez-Lopez, C., Sierra, S., Herrera, M., Grande-Perez, A., Lowenstein, P. R., Manrubia, S. C., Lazaro, E. & Escarmis, C. 2005 Foot-and-mouth disease virus evolution: exploring pathways towards virus extinction. *Curr Top Microbiol Immunol* **288**, 149-73.

Domingo, E., Ruiz-Jarabo, C. M., Sierra, S., Arias, A., Pariente, N., Baranowski, E. & Escarmis, C. 2001 Emergence and selection of RNA virus variants: memory and extinction. *Virus Res* **82**, 39-44.

Domingo, E., Sabo, D., Taniguchi, T. & Weissmann, C. 1978 Nucleotide sequence heterogeneity of an RNA phage population. *Cell* **13**, 735-44.

Don, R. H., Cox, P. T., Wainwright, B. J., Baker, K. & Mattick, J. S. 1991 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Res* **19**, 4008.

Dowell, S. F., Groves, C., Kirkland, K. B., Cicirello, H. G., Ando, T., Jin, Q., Gentsch, J. R., Monroe, S. S., Humphrey, C. D., Slemp, C. & et al. 1995 A multistate outbreak of oyster-associated gastroenteritis: implications for interstate tracing of contaminated shellfish. *J Infect Dis* **171**, 1497-503.

Drake, J. W. 1993 Rates of spontaneous mutation among RNA viruses. *Proc Natl Acad Sci U S A* **90**, 4171-5.

Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**, e88.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G. & Solomon, W. 2002 Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161**, 1307-20.

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

Eckerle, L. D., Lu, X., Sperry, S. M., Choi, L. & Denison, M. R. 2007 High Fidelity of Murine Hepatitis Virus Replication Is Decreased in nsp14-Exoribonuclease Mutants. *J Virol*.

Edwards, C. T., Holmes, E. C., Wilson, D. J., Viscidi, R. P., Abrams, E. J., Phillips, R. E. & Drummond, A. J. 2006 Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. *BMC Evol Biol* **6**, 28.

Escarmis, C., Davila, M., Charpentier, N., Bracho, A., Moya, A. & Domingo, E. 1996 Genetic lesions associated with Muller's ratchet in an RNA virus. *J Mol Biol* **264**, 255-67.

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

Farci, P., Shimoda, A., Coiana, A., Diaz, G., Peddis, G., Melpolder, J. C., Strazzera, A., Chien, D. Y., Munoz, S. J., Balestrieri, A., Purcell, R. H. & Alter, H. J. 2000 The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. *Science* **288**, 339-44.

Felsenstein, J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368-76.

Felsenstein, J. 2004 PHYLIP (Phylogeny Inference Package) version 3.66. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Follett, E. A., Pringle, C. R. & Pennington, T. H. 1975 Virus development in enucleate cells: echovirus, poliovirus, pseudorabies virus, reovirus, respiratory syncytial virus and Semliki Forest virus. *J Gen Virol* **26**, 183-96.

Forss, S., Strebel, K., Beck, E. & Schaller, H. 1984 Nucleotide sequence and genome organization of foot-and-mouth disease virus. *Nucleic Acids Res* **12**, 6587-601.

Freistadt, M. S., Vaccaro, J. A. & Eberle, K. E. 2007 Biochemical characterization of the fidelity of poliovirus RNA-dependent RNA polymerase. *Virol J* **4**, 44.

Fry, E. E., Lea, S. M., Jackson, T., Newman, J. W., Ellard, F. M., Blakemore, W. E., Abu-Ghazaleh, R., Samuel, A., King, A. M. & Stuart, D. I. 1999 The structure and function of a foot-and-mouth disease virus-oligosaccharide receptor complex. *EMBO J* **18**, 543-54.

Garcia-Arriaza, J., Manrubia, S. C., Toja, M., Domingo, E. & Escarmis, C. 2004 Evolutionary Transition toward Defective RNAs That Are Infectious by Complementation. *J Virol* **78**, 11678-11685.

Garland, A. J. M. & Donaldson, A. I. 1990 *Surveillance* **17**, 6-8.

Gebauer, F., de la Torre, J. C., Gomes, I., Mateu, M. G., Barahona, H., Tiraboschi, B., Bergmann, I., de Mello, P. A. & Domingo, E. 1988 Rapid selection of genetic and antigenic variants of foot-and-mouth disease virus during persistence in cattle. *J Virol* **62**, 2041-9.

Gibbens, J. C., Sharpe, C. E., Wilesmith, J. W., Mansley, L. M., Michalopoulou, E., Ryan, J. B. M. & Hudson, M. 2001 Descriptive epidemiology of the 2001 foot-and-mouth disease epidemic in Great Britain: the first five months. *Vet Rec* **149**, 729-743.

Gibbens, J. C. & Wilesmith, J. W. 2002 Temporal and geographical distribution of cases of foot-and-mouth disease during the early weeks of the 2001 epidemic in Great Britain. *Vet Rec* **151**, 407-412.

Gloster, J. & Alexandersen, S. 2004 New Directions: Airborne Transmission of Foot-and-Mouth Disease Virus. *Atmospheric Environment* **38**, 503-505.

Gloster, J., Champion, H. J., Sorensen, J. H., Mikkelsen, T., Ryall, D. B., Astrup, P., Alexandersen, S. & Donaldson, A. I. 2003 Airborne transmission of foot-and-mouth disease virus from Burnside Farm, Heddon-on-the-Wall, Northumberland, during the 2001 epidemic in the United Kingdom. *Vet Rec* **152**, 525-533.

Gonzalez-Candelas, F., Bracho, M. A. & Moya, A. 2003 Molecular epidemiology and forensic genetics: application to a hepatitis C virus transmission event at a hemodialysis unit. *J Infect Dis* **187**, 352-8.

Goodfellow, I. G., Kerrigan, D. & Evans, D. J. 2003 Structure and function analysis of the poliovirus cis-acting replication element (CRE). *Rna-a Publication of the Rna Society* **9**, 124-137.

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

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

Hasegawa, M., Kishino, H. & Yano, T. 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160-74.

Haydon, D. T., Chase-Topping, M., Shaw, D. J., Matthews, L., Friar, J. K., Wilesmith, J. & Woolhouse, M. E. 2003 The construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. *Proc Biol Sci* **270**, 121-7.

Haydon, D. T., Samuel, A. R. & Knowles, N. J. 2001 The generation and persistence of genetic variation in foot-and-mouth disease virus. *Preventive Veterinary Medicine* **51**, 111-124.

Heath, L., van der Walt, E., Varsani, A. & Martin, D. P. 2006 Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J Virol* **80**, 11827-32.

Ho, S. Y., Phillips, M. J., Drummond, A. J. & Cooper, A. 2005 Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation. *Mol Biol Evol* **22**, 1355-63.

Holguin, A., Hernandez, J., Martinez, M. A., Mateu, M. G. & Domingo, E. 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)**, 601-9.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S. & VandePol, S. 1982 Rapid evolution of RNA genomes. *Science* **215**, 1577-85.

Holmes, E. C. 2003 Error thresholds and the constraints to RNA virus evolution. *Trends in Microbiology* **11**, 543-546.

Honhold, N., Taylor, N. M., Mansley, L. M. & Paterson, A. D. 2004 Relationship of speed of slaughter on infected premises and intensity of culling of other premises to the rate of spread of the foot-and-mouth disease epidemic in Great Britain, 2001. *Vet Rec* **155**, 287-94.

Horsington, J. & Zhang, Z. 2007 Consistent change in the B-C loop of VP2 observed in foot-and-mouth disease virus from persistently infected cattle: implications for association with persistence. *Virus Res* **125**, 114-8.

Huelsenbeck, J. P. & Ronquist, F. 2001 MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754-5.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. 2001 Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-4.

Hugh-Jones, M. E. & Tinline, R. R. 1976 Studies on the 1967--68 foot and mouth disease epidemic: incubation period and herd serial interval. *J Hyg (Lond)* **77**, 141-53.

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

Jackson, A. L., O'Neill, H., Maree, F., Blignaut, B., Carrillo, C., Rodriguez, L. & Haydon, D. T. 2007 Mosaic structure of foot-and-mouth disease virus genomes. *J Gen Virol* **88**, 487-92.

Jackson, T., King, A. M. Q., Stuart, D. I. & Fry, E. 2003 Structure and receptor binding. *Virus Res* **91**, 33-46.

Jackson, T., Sheppard, D., Denyer, M., Blakemore, W. & King, A. M. 2000 The epithelial integrin alphavbeta6 is a receptor for foot-and-mouth disease virus. *J Virol* **74**, 4949-56.

Jerzak, G. V., Bernard, K., Kramer, L. D., Shi, P. Y. & Ebel, G. D. 2007 The West Nile virus mutant spectrum is host-dependant and a determinant of mortality in mice. *Virology* **360**, 469-76.

Jinding, C., Mingqiu, Z., Hui, K. H. & Leung, F. C. 2006 Molecular Characterization of Foot-and-mouth Disease Virus in Hong Kong During 2001-2002. *Virus Genes* **32**, 139-43.

Jones, M. E., Thomas, S. M. & Rogers, A. 1994 Luria-Delbruck fluctuation experiments: design and analysis. *Genetics* **136**, 1209-16.

Kahana, R., Kuznetzova, L., Rogel, A., Shemesh, M., Hai, D., Yadin, H. & Stram, Y. 2004 Inhibition of foot-and-mouth disease virus replication by small interfering RNA. *J Gen Virol* **85**, 3213-7.

Kanno, T., Yamakawa, M., Yoshida, K. & Sakamoto, K. 2002 The complete nucleotide sequence of the PanAsia strain of foot-and-mouth disease virus isolated in Japan. *Virus Genes* **25**, 119-25.

Keeling, M. J., Woolhouse, M. E., Shaw, D. J., Matthews, L., Chase-Topping, M., Haydon, D. T., Cornell, S. J., Kappey, J., Wilesmith, J. & Grenfell, B. T. 2001 Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* **294**, 813-7.

Keulen, W., van Wijk, A., Schuurman, R., Berkhout, B. & Boucher, C. A. 1999 Increased polymerase fidelity of lamivudine-resistant HIV-1 variants does not limit their evolutionary potential. *AIDS* **13**, 1343-9.

Kitson, J. D. A. 1990 Analysis of the antigenic sites of O1Kaufburen foot-and-mouth disease virus. In *Department of Microbiology*, vol. PhD Thesis: University of Reading.

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

Kobayashi, Y., Sato, G., Shoji, Y., Sato, T., Itou, T., Cunha, E. M., Samara, S. I., Carvalho, A. A., Nociti, D. P., Ito, F. H. & Sakai, T. 2005 Molecular epidemiological analysis of bat rabies viruses in Brazil. *J Vet Med Sci* **67**, 647-52.

Lanave, C., Preparata, G., Saccone, C. & Serio, G. 1984 A new method for calculating evolutionary substitution rates. *J Mol Evol* **20**, 86-93.

Lazaro, E., Escarmis, C., Perez-Mercader, J., Manrubia, S. C. & Domingo, E. 2003 Resistance of virus to extinction on bottleneck passages: study of a decaying and fluctuating pattern of fitness loss. *Proc Natl Acad Sci U S A* **100**, 10830-5.

Lee, C. H., Gilbertson, D. L., Novella, I. S., Huerta, R., Domingo, E. & Holland, J. J. 1997 Negative effects of chemical mutagenesis on the adaptive behavior of vesicular stomatitis virus. *J Virol* **71**, 3636-40.

Leitner, T. & Albert, J. 1999 The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci U S A* **96**, 10752-7.

Li, D., Shang, Y. J., Liu, Z. X., Liu, X. T. & Cai, X. P. 2007 Molecular relationships between type Asia 1 new strain from China and type O Panasia strains of foot-and-mouth-disease virus. *Virus Genes*, 273-9.

Liu, J., Lim, S. L., Ruan, Y., Ling, A. E., Ng, L. F., Drosten, C., Liu, E. T., Stanton, L. W. & Hibberd, M. L. 2005a SARS transmission pattern in Singapore reassessed by viral sequence variation analysis. *PLoS Med* **2**, e43.

Liu, M., Chen, W., Ni, Z., Yan, W., Fei, L., Jiao, Y., Zhang, J., Du, Q., Wei, X., Chen, J., Liu, Y. & Zheng, Z. 2005b Cross-inhibition to heterologous foot-and-mouth disease virus infection induced by RNA interference targeting the conserved regions of viral genome. *Virology* **336**, 51-9.

Lopman, B. A., Gallimore, C., Gray, J. J., Vipond, I. B., Andrews, N., Sarangi, J., Reacher, M. H. & Brown, D. W. 2006 Linking healthcare associated norovirus outbreaks: a molecular epidemiologic method for investigating transmission. *BMC Infect Dis* **6**, 108.

Lowings, J. P., Paton, D. J., Sands, J. J., De Mia, G. M. & Rutili, D. 1994 Classical swine fever: genetic detection and analysis of differences between virus isolates. *J Gen Virol* **75 (Pt 12)**, 3461-8.

Malet, I., Belnard, M., Agut, H. & Cahour, A. 2003 From RNA to quasispecies: a DNA polymerase with proofreading activity is highly recommended for accurate assessment of viral diversity. *J Virol Methods* **109**, 161-70.

Malpica, J. M., Fraile, A., Moreno, I., Obies, C. I., Drake, J. W. & Garcia-Arenal, F. 2002 The rate and character of spontaneous mutation in an RNA virus. *Genetics* **162**, 1505-11.

Manrubia, S. C., Escarmis, C., Domingo, E. & Lazaro, E. 2005 High mutation rates, bottlenecks, and robustness of RNA viral quasispecies. *Gene* **347**, 273-82.

Martinez, M. A., Dopazo, J., Hernandez, J., Mateu, M. G., Sobrino, F., Domingo, E. & Knowles, N. J. 1992 Evolution of the capsid protein genes of foot-and-mouth disease virus: antigenic variation without accumulation of amino acid substitutions over six decades. *J Virol* **66**, 3557-65.

Mason, P. W., Bezborodova, S. V. & Henry, T. M. 2002 Identification and characterization of a cis-acting replication element (cre) adjacent to the internal ribosome entry site of foot-and-mouth disease virus. *J Virol* **76**, 9686-94.

Mason, P. W., Grubman, M. J. & Baxt, B. 2003a Molecular basis of pathogenesis of FMDV. *Virus Res* **91**, 9-32.

Mason, P. W., Pacheco, J. M., Zhao, Q.-Z. & Knowles, N. J. 2003b Comparisons of the complete genomes of Asian, African and European isolates of a recent foot-and-mouth disease virus type O pandemic strain (PanAsia). *J Gen Virol* **84**, 1583-1593.

Mason, P. W., Rieder, E. & Baxt, B. 1994 RGD sequence of foot-and-mouth disease virus is essential for infecting cells via the natural receptor but can be bypassed by an antibody-dependent enhancement pathway. *Proc Natl Acad Sci U S A* **91**, 1932-6.

Medina, M., Domingo, E., Brangwyn, J. K. & Belsham, G. J. 1993 The two species of the foot-and-mouth disease virus leader protein, expressed individually, exhibit the same activities. *Virology* **194**, 355-9.

Moffat, K., Knox, C., Howell, G., Clark, S. J., Yang, H., Belsham, G. J., Ryan, M. & Wileman, T. 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**, 1129-39.
Mohapatra, J. K., Sanyal, A., Hemadri, D., Tosh, C., Sabarinath, G. P. & Venkataramanan, R. 2002 Sequence and phylogenetic analysis of the L and VP1 genes of foot-and-mouth disease virus serotype Asia1. *Virus Res* **87**, 107-18.
Moya, A., Holmes, E. C. & Gonzalez-Candelas, F. 2004 The population genetics and evolutionary epidemiology of RNA viruses. *Nat Rev Microbiol* **2**, 279-88.
Muse, S. V. & Gaut, B. S. 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* **11**, 715-24.
Nei, M. & Li, W. H. 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci U S A* **76**, 5269-73.
Nobiron, I., Remond, M., Kaiser, C., Lebreton, F., Zientara, S. & Delmas, B. 2005 The nucleotide sequence of foot-and-mouth disease virus O/FRA/1/2001 and comparison with its British parental strain O/UKG/35/2001. *Virus Res* **108**, 225-9.
Novak, J. E. & Kirkegaard, K. 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**, 3384-3387.
Oem, J. K., Lee, K. N., Cho, I. S., Kye, S. J., Park, J. H. & Joo, Y. S. 2004 Comparison and analysis of the complete nucleotide sequence of foot-and-mouth disease viruses from animals in Korea and other PanAsia strains. *Virus Genes* **29**, 63-71.
Pariente, N., Airaksinen, A. & Domingo, E. 2003 Mutagenesis versus inhibition in the efficiency of extinction of foot-and-mouth disease virus. *J Virol* **77**, 7131-8.
Pariente, N., Sierra, S., Lowenstein, P. R. & Domingo, E. 2001 Efficient virus extinction by combinations of a mutagen and antiviral inhibitors. *J Virol* **75**, 9723-30.
Parvin, J. D., Moscona, A., Pan, W. T., Leider, J. M. & Palese, P. 1986 Measurement of the mutation rates of animal viruses: influenza A virus and poliovirus type 1. *J Virol* **59**, 377-83.
Pfeiffer, J. K. & Kirkegaard, K. 2005 Increased fidelity reduces poliovirus fitness and virulence under selective pressure in mice. *PLoS Pathog* **1**, e11.
Pfeiffer, J. K. & Kirkegaard, K. 2006 Bottleneck-mediated quasispecies restriction during spread of an RNA virus from inoculation site to brain. *Proc Natl Acad Sci U S A* **103**, 5520-5.
Pond, S. L., Frost, S. D. & Muse, S. V. 2005 HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676-9.
Poon, A. F., Kosakovsky Pond, S. L., Bennett, P., Richman, D. D., Leigh Brown, A. J. & Frost, S. D. 2007 Adaptation to human populations is revealed by within-host polymorphisms in HIV-1 and hepatitis C virus. *PLoS Pathog* **3**, e45.
Portner, A., Webster, R. G. & Bean, W. J. 1980 Similar frequencies of antigenic variants in Sendai, vesicular stomatitis, and influenza A viruses. *Virology* **104**, 235-8.
Posada, D. & Crandall, K. A. 1998 MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817-8.
Quer, J., Esteban, J. I., Cos, J., Sauleda, S., Ocana, L., Martell, M., Otero, T., Cubero, M., Palou, E., Murillo, P., Esteban, R. & Guardia, J. 2005 Effect of bottlenecking on evolution of the nonstructural protein 3 gene of hepatitis C virus during sexually transmitted acute resolving infection. *J Virol* **79**, 15131-41.
Racaniello, V. R. 1996 Picornaviridae: The Viruses and Their Replication. In *Virology* (ed. Fields), pp. 685-722.
Randrup, A. 1954 On the stability of bovine foot-and-mouth disease virus dependent on pH; investigations on the complement fixing and the immunizing antigen as well as on the infective agent. *Acta Pathol Microbiol Scand* **35**, 388-95.
Reid, S. M., Forsyth, M. A., Hutchings, G. H. & Ferris, N. P. 1998 Comparison of reverse transcription polymerase chain reaction, enzyme linked immunosorbent assay and virus isolation for the routine diagnosis of foot-and-mouth disease. *J Virol Methods* **70**, 213-7.

Rhodes, T., Wargo, H. & Hu, W. S. 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**, 11193-200.

Rieder, E., Bunch, T., Brown, F. & Mason, P. W. 1993 Genetically engineered foot-and-mouth disease viruses with poly(C) tracts of two nucleotides are virulent in mice. *J Virol* **67**, 5139-45.

Rouzine, I. M., Rodrigo, A. & Coffin, J. M. 2001 Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. *Microbiol Mol Biol Rev* **65**, 151-85.

Rozas, J. & Rozas, R. 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174-5.

Ruiz-Jarabo, C. M., Arias, A., Baranowski, E., Escarmis, C. & Domingo, E. 2000 Memory in viral quasispecies. *J Virol* **74**, 3543-7.

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

Sa-Carvalho, D., Rieder, E., Baxt, B., Rodarte, R., Tanuri, A. & Mason, P. W. 1997 Tissue culture adaptation of foot-and-mouth disease virus selects viruses that bind to heparin and are attenuated in cattle. *J Virol* **71**, 5115-23.

Salzberger, B., Wiesel, W., Schwingel, E. & Hauber, I. 2000 Occupational HIV infection documented by sequence analysis of source and index virus strains. *HIV Med* **1**, 256-8.

Sanchez, A. & Rollin, P. E. 2005 Complete genome sequence of an Ebola virus (Sudan species) responsible for a 2000 outbreak of human disease in Uganda. *Virus Res* **113**, 16-25.

Sangare, O., Bastos, A. D. S., Venter, E. H. & Vosloo, W. 2003 Retrospective genetic analysis of SAT-1 type foot-and-mouth disease outbreaks in West Africa (1975-1981). *Vet Microbiol* **93**, 279-289.

Savill, N. J., Shaw, D. J., Deardon, R., Tildesley, M. J., Keeling, M. J., Woolhouse, M. E., Brooks, S. P. & Grenfell, B. T. 2007 Effect of data quality on estimates of farm infectiousness trends in the UK 2001 foot-and-mouth disease epidemic. *J R Soc Interface* **4**, 235-41.

Savolainen, C., Mulders, M. N. & Hovi, T. 2002 Phylogenetic analysis of rhinovirus isolates collected during successive epidemic seasons. *Virus Res* **85**, 41-6.

Schneider, W. L. & Roossinck, M. J. 2000 Evolutionarily related Sindbis-like plant viruses maintain different levels of population diversity in a common host. *J Virol* **74**, 3130-4.

Schneider, W. L. & Roossinck, M. J. 2001 Genetic Diversity in RNA Virus Quasispecies Is Controlled by Host-Virus Interactions. *J Virol* **75**, 6566-6571.

Schrag, S. J., Rota, P. A. & Bellini, W. J. 1999 Spontaneous mutation rate of measles virus: direct estimation based on mutations conferring monoclonal antibody resistance. *J Virol* **73**, 51-4.

Scudamore, J. M. 2002 Origin of the UK Foot-and-mouth disease epidemic in 2001: Department of Environment, Food and Rural Affairs.

Sedivy, J. M., Capone, J. P., RajBhandary, U. L. & Sharp, P. A. 1987 An inducible mammalian amber suppressor: propagation of a poliovirus mutant. *Cell* **50**, 379-89.

Serrano, P., Pulido, M. R., Saiz, M. & Martinez-Salas, E. 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**, 3013-22.

Sevilla, N., Ruiz-Jarabo, C. M., Gomez-Mariano, G., Baranowski, E. & Domingo, E. 1998 An RNA virus can adapt to the multiplicity of infection. *J Gen Virol* **79 (Pt 12)**, 2971-80.

Sevilla, N., Verdaguer, N. & Domingo, E. 1996 Antigenically profound amino acid substitutions occur during large population passages of foot-and-mouth disease virus. *Virology* **225**, 400-5.

Sierra, S., Davila, M., Lowenstein, P. R. & Domingo, E. 2000 Response of foot-and-mouth disease virus to increased mutagenesis: influence of viral load and fitness in loss of infectivity. *J Virol* **74**, 8316-23.

Simmonds, P., Tuplin, A. & Evans, D. J. 2004 Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *Rna-a Publication of the Rna Society* **10**, 1337-1351.

Sobrino, F., Davila, M., Ortin, J. & Domingo, E. 1983 Multiple genetic variants arise in the course of replication of foot-and-mouth disease virus in cell culture. *Virology* **128**, 310-8.

Sobrino, F., Palma, E. L., Beck, E., Davila, M., de la Torre, J. C., Negro, P., Villanueva, N., Ortin, J. & Domingo, E. 1986 Fixation of mutations in the viral genome during an outbreak of foot-and-mouth disease: heterogeneity and rate variations. *Gene* **50**, 149-59.

Spada, E., Sagliocca, L., Sourdis, J., Garbuglia, A. R., Poggi, V., De Fusco, C. & Mele, A. 2004 Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. *J Clin Microbiol* **42**, 4230-6.

Stene-Johansen, K., Jenum, P. A., Hoel, T., Blystad, H., Sunde, H. & Skaug, K. 2002 An outbreak of hepatitis A among homosexuals linked to a family outbreak. *Epidemiol Infect* **129**, 113-7.

Susetya, H., Sugiyama, M., Inagaki, A., Ito, N., Oraveerakul, K., Traiwanatham, N. & Minamoto, N. 2003 Genetic characterization of rabies field isolates from Thailand. *Microbiol Immunol* **47**, 653-9.

Swofford, D. L. 2001 PAUP*: phylogenetic analysis using parsimony (*and other methods), beta version 4.0b2. *Sinauer Associates, Sunderland, Massachusetts*.

Taylor, N. M., Honhold, N., Paterson, A. D. & Mansley, L. M. 2004 Risk of foot-and-mouth disease associated with proximity in space and time to infected premises and the implications for control policy during the 2001 epidemic in Cumbria. *Vet Rec* **154**, 617-26.

Templeton, A. R., Crandall, K. A. & Sing, C. F. 1992 A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **132**, 619-33.

Thompson, D., Muriel, P., Russell, D., Osborne, P., Bromley, A., Rowland, M., Creigh-Tyte, S. & Brown, C. 2002 Economic costs of the foot and mouth disease outbreak in the United Kingdom in 2001. *Rev Sci Tech* **21**, 675-87.

Thrusfield, M., Mansley, L., Dunlop, P., Taylor, J., Pawson, A. & Stringer, L. 2005 The foot-and-mouth disease epidemic in Dumfries and Galloway, 2001. 1: Characteristics and control. *Vet Rec* **156**, 229-52.

Toja, M., Escarmis, C. & Domingo, E. 1999 Genomic nucleotide sequence of a foot-and-mouth disease virus clone and its persistent derivatives: Implications for the evolution of viral quasispecies during a persistent infection. *Virus Res* **64**, 161-171.

Tsai, C.-P., Pan, C.-H., Liu, M.-Y., Lin, Y.-l., Chen, C.-M., Huang, T.-S., Cheng, I.-C., Jong, M.-H. & Yang, P.-C. 2000 Molecular epidemiological studies on foot-and-mouth disease type O Taiwan viruses from the 1997 epidemic. *Vet Microbiol* **74**, 207-216.

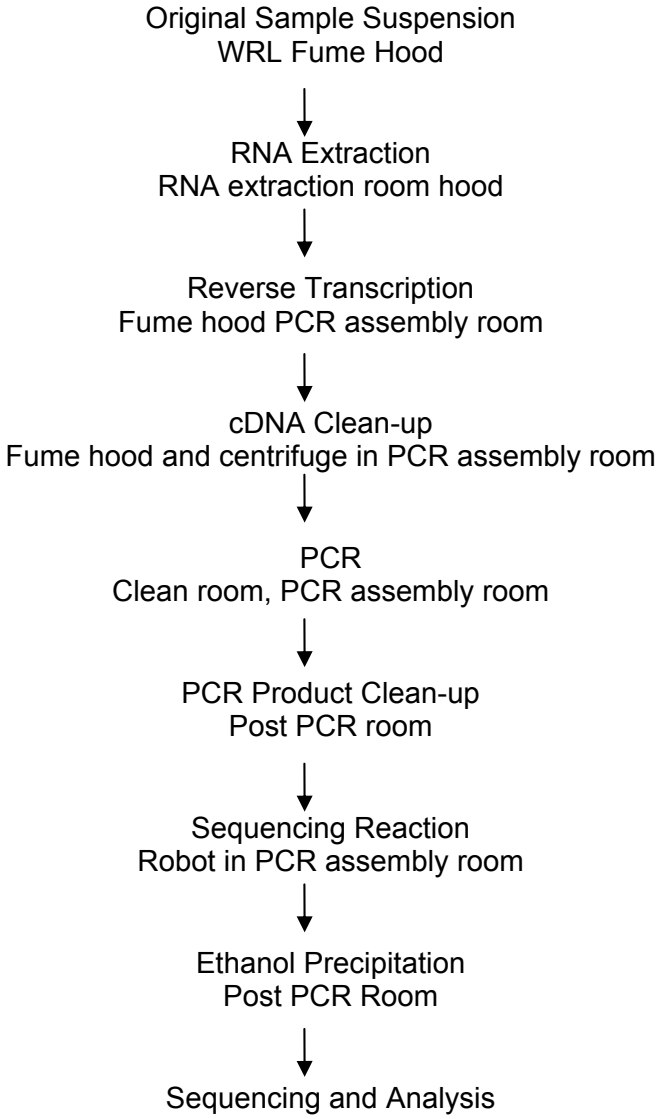
Vignuzzi, M., Stone, J. K., Arnold, J. J., Cameron, C. E. & Andino, R. 2006 Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* **439**, 344-8.

Vosloo, W., Bastos, A. D., Kirkbride, E., Esterhuysen, J. J., van Rensburg, D. J., Bengis, R. G., Keet, D. W. & Thomson, G. R. 1996 Persistent infection of African buffalo (Syncerus caffer) with SAT-type foot-and-mouth disease viruses: rate of fixation of mutations, antigenic change and interspecies transmission. *J Gen Virol* **77 (Pt 7)**, 1457-67.

Wallace, R. G., Hodac, H., Lathrop, R. H. & Fitch, W. M. 2007 A statistical phylogeography of influenza A H5N1. *Proc Natl Acad Sci U S A* **104**, 4473-8.
Walsh, P. D., Biek, R. & Real, L. A. 2005 Wave-like spread of Ebola Zaire. *PLoS Biol* **3**, e371.
Ward, C. D. & Flanegan, J. B. 1992 Determination of the poliovirus RNA polymerase error frequency at eight sites in the viral genome. *J Virol* **66**, 3784-93.
Ward, C. D., Stokes, M. A. & Flanegan, J. B. 1988 Direct measurement of the poliovirus RNA polymerase error frequency in vitro. *J Virol* **62**, 558-62.
Wells, V. R., Plotch, S. J. & DeStefano, J. J. 2001 Determination of the mutation rate of poliovirus RNA-dependent RNA polymerase. *Virus Res* **74**, 119-32.
Widjojoatmodjo, M. N., van Gennip, H. G., de Smit, A. J. & Moormann, R. J. 1999 Comparative sequence analysis of classical swine fever virus isolates from the epizootic in The Netherlands in 1997-1998. *Vet Microbiol* **66**, 291-9.
Wild, T. F., Burroughs, J. N. & Brown, F. 1969 Surface structure of foot-and-mouth disease virus. *J Gen Virol* **4**, 313-20.
Witwer, C., Rauscher, S., Hofacker, I. L. & Stadler, P. F. 2001 Conserved RNA secondary structures in Picornaviridae genomes. *Nucleic Acids Res* **29**, 5079-89.
Wong, C. W., Albert, T. J., Vega, V. B., Norton, J. E., Cutler, D. J., Richmond, T. A., Stanton, L. W., Liu, E. T. & Miller, L. D. 2004 Tracking the evolution of the SARS coronavirus using high-throughput, high-density resequencing arrays. *Genome Res* **14**, 398-405.
Yang, Z. 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-6.
Yang, Z. & Rannala, B. 1997 Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol* **14**, 717-24.
Yu, W., Rusterholtz, K. J., Krummel, A. T. & Lehman, N. 2006 Detection of high levels of recombination generated during PCR amplification of RNA templates. *Biotechniques* **40**, 499-507.
Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. 2000 RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25-33.
Zhang, L., Diaz, R. S., Ho, D. D., Mosley, J. W., Busch, M. P. & Mayer, A. 1997 Host-specific driving force in human immunodeficiency virus type 1 evolution in vivo. *J Virol* **71**, 2555-61.
Zhao, Z., Li, H., Wu, X., Zhong, Y., Zhang, K., Zhang, Y. P., Boerwinkle, E. & Fu, Y. X. 2004 Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol* **4**, 21.
Zhou, D., He, Q. S., Wang, C., Zhang, J. & Wong-Staal, F. 2006 RNA interference and potential applications. *Curr Top Med Chem* **6**, 901-11.

Appendix 1

Complete genome sequencing overview



Original Epithelium Sample Suspension

Class II Safety Cabinet

Method

Fill fish kettle with water (put to boil after)
Weigh Bijou
Blot epithelium sample on tissue to remove glycerol and place in bijou
Re-weigh bijou and calculate weight of epithelium sample
Require a 1/10 suspension – multiply weight by 9 and add that many mls (*) of buffer solution
Use scissors and forceps to cut tissue up into small a bits as possible
Add approximately same amount of sand as epithelium sample and grind into a paste with pestle and mortar
Add *ml of buffer solution (kept under centrifuge)
Pour into centrifuge universal
Centrifuge for 10 minutes
Remove supernatant and pour into a fresh bijou
Aliquot into tubes and store at -70◦C
Add 0.2ml to trizol

RNA Extraction in TRIZOL

Class II Safety Cabinet

Method

Ensure any samples frozen and previously stored in TRIZOL are thawed. Or add 200µl of sample to 1ml of TRIZOL. Incubate at room temperature for 5 minutes
Add 200µl of chloroform direct to the tube.
Vortex mix the tube for 10-15 seconds, then leave at ~22oC for 3 minutes.
Centrifuge for 15 minutes at 13000 rpm at 2-8oC
Label a fresh 1.5ml eppendorf and add 1µl of glycogen.
Remove as much of the top phase as possible (650-700µl) from centrifuged tubes and add to corresponding glycogen filled tubes
Add an equivalent volume (650-700µl) of Isopropyl alcohol to each tube
Vortex mix each tube for a few seconds
Incubate at ~22oC for 10 minutes
Centrifuge for 15 minutes at 13000 rpm at 2-8oC
Discard the supernatant and add 1000µl of 70% ethanol to each one. Gently mix by flicking the tube
Centrifuge for 15minutes at 13000 rpm at 2-8oC
Remove by pipetting as much supernatant as possible and air drying
Re-suspend the RNA pellet by adding 12µl nuclease free ddH2O
Maintain on ice for reverse transcription or store at -70oC

Reverse Transcription

Class II Safety Cabinet

Method

Add 4µl Primer UK/FMD Rev6, 1µl Random hexamers (50 ng/µl), 3µl dNTP to 12ul RNA.
Place at 70oC for 3 minutes
Place on ice for 3 minutes
Add 18µl of RT mix made fresh (4µl 10x RT buffer, 8µl 25mM MgCl, 4µl 0.1M DTT, 2µl ddH2O)
Add 2µl Superscript III Reverse Transcriptase
Incubate at 42oC for 4 hours followed by termination of the reaction at 85oC for 5 minutes. Thermocycler programme Revtran1

cDNA Clean-up

Class II Safety Cabinet

Method

Add 200µl of buffer PB to the 40µl reverse transcription product
Transfer to a spin column tube and centrifuge at 13000 rpm for 1 minute
Discard flow through
Add 0.7ml of buffer PE and centrifuge for 1 minute at 13000 rpm to wash
Discard the flow through
Centrifuge again for 1 minute at 13000 rpm
Add 30µl of Elution Buffer, leave for 1 minute
Place into a fresh labelled 1.5ml eppendorf and centrifuge for 1 minute at 13000 rpm to elute the DNA

PCR set up

Class II Safety Cabinet

Method

Set up PCR master mix in 96 well PCR plates in clean room for the 5 primer sets (F1/R1, F2/R2, F3/R7,F4/R4, F5/R6) as described in Table A1.
Add 2.5 µl cDNA to each reaction in hood
Run primer sets 1-4 on programme FMD-EC1, and primer set 5 on FMD-EC2 programmed onto thermocycler.

Table A1

Reagent	*x1*
10x buffer	5 µl
MgSO4	2 µl
10mM dNTP mix	1 µl
UK/FMD For?	1 µl
UK/FMD Rev?	1 µl
Platinum TAQ HiFi	0.2 µl
ddH2O	37.8 µl
	48 µl

Table A2

	FMD-EC1	
Step	Temp oC	Time
1	94	2 min
2	94	30sec
3	68	30sec
4	72	3min
5	Go to step 2 x 39	
6	72	7min
7	4	∞
8	END	

Table A3

	FMD-EC2	
Step	Temp oC	Time
1	94	2min
2	94	45sec
3	60	45sec
4	72	3min
5	Go to step 2 x 39	
6	72	7min
7	4	∞
8	END	

PCR Clean-up

Method

Run 2µl of PCR product on 1.2% agarose gel at 105v for 30 minutes to check reaction has worked
Add 240µl of buffer PB to the 48µl PCR product
Transfer to a spin column tube and centrifuge at 13000 rpm for 1 minute
Discard flow through
Add 0.7ml of buffer PE and centrifuge for 1 minute at 13000 rpm to wash
Discard the flow through
Centrifuge again for 1 minute at 13000 rpm
Add 50µl of Elution Buffer, leave for 1 minute
Place into a fresh labelled 1.5ml eppendorf and centrifuge for 1 minute at 13000 rpm to elute the DNA
Run 1µl product on gel to quantitate DNA concentration.
Dilute products to give the following concentrations and volumes for sequencing (Table A4)

Table A4

Primer Set	*DNA concentration*	*Volume*
1	20ng/8µl	40µl
2	70ng/8µl	112µl
3	100ng/8µl	184µl
4	100ng/µl	152µl
5	100ng/µl	216µl

Sequencing reaction

Method

Make up primer plate of 1mM Forec 1-42 and Revec 43-84 primers in clean room (Fig 1). Use the robot programme FMD genome sequencing to add 8µl of Mastermix to 84 wells of a 96 well Beckman sequencing plate, 4µl of Forec#/Revec# Primer (from pre-made plate) and 8µl of PCR product.
Spin plate in centrifuge briefly to accumulate reagents in bottom of well.
Run on programme Cycleseq on Thermocycler.

Figure A1

1	9	17	25	33	41	49	57	65	73	81	
2	10	18	26	34	42	50	58	66	74	82	
3	11	19	27	35	43	51	59	67	75	83	
4	12	20	28	36	44	52	60	68	76	84	
5	13	21	29	37	45	53	61	69	77		
6	14	22	30	38	46	54	62	70	78		
7	15	23	31	39	47	55	63	71	79		
8	16	24	32	40	48	56	64	72	80		

X = Fragment 1
X = Fragment 2
X = Fragment 3
X = Fragment 4
X = Fragment 5

Ethanol Precipitation

Method

Turn on centrifuge to be at 4oC
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