Three studies showing the importance of quantitative methods in investigation of veterinary infectious disease

Hannah Trewby BVSc (Hons) MSc

Thesis submitted to fulfil the requirements for the degree of Master of Science (Veterinary Science) in the School of Veterinary Medicine, College of Medical Veterinary and Life Sciences, University of Glasgow

November 2011

True wisdom leaks from the joins between disciplines

Ian McDonald, The Dervish House

ABSTRACTS

Chapter 1 – EXAMINING THE EVOLUTIONARY HISTORY OF BOVINE PAPILLOMAVIRUS IN EQUINE SARCOIDS

The papillomavirus (PV) family consist of slowly evolving host-adapted DNA viruses. Bovine papillomaviruses (BPVs) -1 and -2 primarily cause warts in their natural host, the cow, but also lead to locally aggressive and invasive skin tumours in equids known as sarcoids. This chapter gives an account of the first phylogenetic analysis of BPV in equine sarcoids, undertaken in order to clarify the evolutionary history of the virus and its cross-species association with equine sarcoids. Phylogenetic trees were constructed for three different stretches of the BPV genome. Although two of these analyses used gene segments that proved too short to draw any firm conclusions, the phylogenetic analysis carried out on the BPV-1 transcriptional promoter region (LCR) from cattle and horse samples provided interesting insights into the evolution of the virus. The genetic diversity seen in the LCR variants was shown to be ancient, predating domestication of both equids and cattle. The phylogenetic tree shows clear geographic segregation, with an ancestral BPV-1 group consisting of African and Brazilian sequences and a more evolved European group of sequences. The distribution of the cattle samples within the phylogeny suggests the sequences originally evolved in ancestral cattle, and that the genetic diversity found in equine sarcoids is the result of multiple, relatively recent species jumps into horses from different seeding strains of the virus. In addition, a specific LCR sequence variant was isolated in equine samples from all countries sampled here, despite being absent from cattle samples, suggesting that viruses containing this sequence variant may have a selective advantage within the equine population.

Chapter 2 – SCOTTISH SHEEP MOVEMENTS AND THEIR POTENTIAL FOR DISEASE TRANSMISSION

Animal movements play a major role in the spread of livestock diseases. By identifying farms pivotal to the network of livestock movements, it may be possible to more efficiently curb the spread of disease. Diseases transmit over great ranges of timescale and infectiousness. Sheep are moved from premises to premises for a variety of different reasons and with widely varying residence times on the arrival premises, and different types of movement are important in the spread of different diseases. This report describes work identifying those sheep farms important in terms of the types of movements involved in both a fast-transmitting and a slowly-transmitting disease. In so doing it raises the possibility of achieving control of multiple infections by targeting just a single subset of farms. If this were possible it would provide a cost effective and efficient method to reduce the burden of disease in the national flock.

Chapter 3 – THE IMPLICATIONS OF POST-INFECTION IMMUNITY FOR THE EPIDEMIOLOGY AND CONTROL OF *Escherichia coli* O157 INFECTION OF CATTLE

This report describes the use of epidemiological modelling to investigate how a period of post-infection immunity impacts the transmission dynamics of *E. coli* O157. Shigatoxigenic strains of *E. coli*, including the O157 strain, cause severe disease in man despite being asymptomatic in cattle, their natural reservoir host. Previous work modelling the transmission dynamics of *E. coli* has assumed that an animal becomes immediately susceptible on recovery from an infection, but recent experimental evidence indicates this may not be the case. In this project, stochastic models were developed for *E. coli* in cattle, allowing comparison of the effects of a period of post-infection immunity with the previously used assumption of immediate return to susceptibility. The results show that post-infection immunity gives lower values for outbreak duration, and for mean and variance in prevalence, and that this is observed over a biologically plausible range of the basic reproduction number, R_o. This in turn indicates that *E. coli* infection is likely to be more difficult to control if post-infection immunity exists, especially at higher infection prevalences. This study also reveals that if the assumption of post-infection immunity is valid, an even higher degree of individual heterogeneity in transmission is needed to explain the degree of variance in *E. coli* O157 prevalence seen in the field, thus validating previous work which demonstrated the importance of supershedder animals and individual heterogeneity. This study provides the first steps in investigating how a period of immunity following *E. coli* infection of cattle affects conclusions drawn by previous work assuming an immediate return to susceptibility. Models allowing the incorporation of individual heterogeneity are needed to further investigate the subject.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I am extremely grateful to Roman Biek, Rowland Kao and Louise Matthews for the time and patience that they put into supervising the projects reported here. Many thanks are also due to the following: Lubna Nasir, who provided the sequence data for Chapter 1; Paul Bessell and Lisa Boden who calculated the hill codings used in Chapter 2; Fiona Houston, David Leggatt, David Logue and Elspeth Waugh for their help in understanding the British sheep industry; Richard Reeve who provided the original SIS model code used in Chapter 3; and Matt Denwood for his help with R. Thanks also go to the Biotechnology and Biological Sciences Research Council, the University of Glasgow and the staff that made the DTG programme (and this research) possible.

AUTHOR DECLARATION

I, Hannah Trewby, declare that the work in this thesis is original and was carried out solely by myself or with due acknowledgments. No part of this thesis has been submitted for any other degree.

LIST OF ABBREVIATIONS

AMLS	Animal Movements Licensing Scheme
BPV	Bovine Papillomavirus
Defra	Department for the environment, food and rural affairs
E	Early gene
FMD	Foot and Mouth Disease
H	Hill flock
L	Late gene
LCR	Long control region
Lo	Lowland flock
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
PV	Papillomavirus
R_0	Basic reproduction number for a disease outbreak
RGeS	Ram Genotyping Scheme
SAMS	Scottish Animal Movements Scheme
SIRS	Susceptible-Infectious-Recovered/Immune-Susceptible model
SIS	Susceptible-Infectious-Susceptible model
U	Upland flock
UL	Non-hill flock (i.e. upland or lowland flock)

OVERALL INTRODUCTION

This thesis presents three research projects in the broad field of veterinary disease, each contributing one chapter to the work. Although each project is independent and the topics covered are disparate in nature, when taken together they all involve the application of quantitative methods to investigate and understand veterinary infectious disease.

Three separate disciplines are commonly used in the investigation of infectious disease. Laboratory science allows determination of the molecular mechanisms of pathogen survival and host interaction in a controlled environment. Observational, field-based studies are the basis of traditional epidemiology and focus on disease within a population of interest. Finally, mathematical modelling is a growing area in which theoretical experiments enable *in silico* prediction of complex systems. The three research projects reported here each draw extensively from one of these three approaches to the study of infectious disease.

Chapter 1 describes the use of phylogenetic analysis to investigate the evolutionary history of a bovine papillomavirus and its cross species association with equine skin tumours. This project demonstrates application of quantitative methods to laboratory-generated data. Current phylogenetic programs make use of advanced statistical methodologies to draw inferences about the probable course of viral evolution from the available sequence data. In so doing, phylogenetics takes primarily a pathogen-based perspective, orientated towards the very small scale by its focus on molecular genetic events.

The project reported in Chapter 2 takes a completely different approach. Here, data on sheep movements, along with knowledge of the sheep industry, were used to infer the contact structure along which a disease may spread through the population. This study made use of field data on the movements of sheep collected at a national level and, although the main thrust of the project was a descriptive overview of sheep movement demographics, the use of concepts from network theory here illustrates the application of advanced numerical methodologies to the realm of traditional epidemiology. This project was concerned with events occurring on large (national) scales, and concentrates on the population perspective.

Chapter 3 demonstrates the use of mathematical modelling, the third discipline described above, to explore how changing model assumptions can affect the predicted outcome of *E. coli* infection in cattle. The use of mathematical models allows extrapolation beyond the limits of the data, and although this project is purely theoretical, it relies heavily on both

observational studies and experimental work to provide its context. In an applied field such as veterinary infectious disease it is perhaps especially important to ensure that theoretical work is grounded in reality: using understanding and information drawn from the other two disciplines means that findings from theoretical models can be applied to real world problems.

These short projects by their very nature can only begin to explore the ever more complex array of mathematical, statistical and programming tools available to the study of infectious disease. However, taken together they illustrate the wide variety of ways in which quantitative methodologies can be used to increase the power and insight gained from more traditional approaches to the study of veterinary infectious disease.

Chapter 1

Examining the evolutionary history of Bovine Papillomavirus in equine sarcoids

Supervised by Dr Roman Biek
(roman.biek@glasgow.ac.uk)

Chapter 1 - EXAMINING THE EVOLUTIONARY HISTORY OF BOVINE PAPILLOMAVIRUS IN EQUINE SARCOIDS

INTRODUCTION

Papillomaviruses (PVs) are a large and diverse family of small, non-enveloped DNA viruses. They have a broad host range that includes mammals, birds and reptiles (Shah et al., 2010). PVs infect the epithelium and/or the dermis of their hosts to cause papillomas or fibropapillomas, commonly known as warts, which generally self-resolve over time. The PVs express several genes during infection divided into the early genes (E1-7) that control viral transcription, replication and also interfere with host-cell growth patterns, and the late genes (L1 and L2) which form the capsid of the mature virus particle and are expressed only in the fully differentiated keratinocytes at the surface of the papilloma (Borzacchiello and Roperto, 2008). The viral genome also contains an untranslated region, the long control region (LCR) which controls viral transcription in conjunction with host factors and the viral E2 protein (Nasir et al., 2007) and appears to mutate at a faster rate than other parts of the PV genome (Rector et al., 2007). As a consequence of the interference in the host-cell growth cycle by the early genes, some PVs are known to cause cancerous transformation of cells, most notably HPV -16 and -18 which cause cervical cancer in women (Anon, 2007). Other examples of PV-induced tumours are the gastric and urinary carcinomas in cattle caused by a combination of BPV infection and bracken fern toxins (Borzacchiello and Roperto, 2008), and equine sarcoids (Lancaster and Olson, 1982).

Double stranded DNA viruses like PVs evolve very slowly, generally only one order of magnitude faster than their hosts and up to four orders of magnitude slower than the fast-evolving RNA viruses (Tachezy et al., 2002). It was originally assumed that PVs co-speciated with their hosts and, although this does seems to be true in the majority of cases, Gottschling et al. found evidence for both viral divergence prior to host speciation events and viral transfer between different host species (‘host jumps’) in PVs (Gottschling et al., 2007). An example of this incongruence in host-pathogen evolution can be seen in the bovine papillomaviruses (BPVs) of cattle, in which a diverse range of PV lineages (δ, ε and ξ PVs) all infect a single host species. The two δ PVs of cattle, BPV-1 and -2, also give us our only existing example of cross-species transmission of a PV. The natural host of BPV-1 and -2 is the cow in which these strains primarily cause self-resolving warts but

they are also the likely cause of equine sarcoid (see below), an important disease of horses and other equids throughout the world.

Equine sarcoids are defined as locally invasive and aggressive fibroblastic tumours and are one of the most common skin tumours of equids, with a prevalence of between 1% - 8% worldwide (Knottenbelt, 2005). They affect all equids including horses, mules and donkeys. Sarcoids very rarely resolve without intervention and are difficult and often costly to treat. Although they do not metastasise to internal organs they can grow very large and a single horse may suffer from between one and several hundred sarcoids (Nasir and Campo, 2008). Sarcoids are not directly fatal but do cause loss of value, reduced performance, loss of use and welfare issues due to fly worry and secondary infection, all of which may lead to euthanasia (http://www.liv.ac.uk/sarcoids/, 2010). This can be especially important in developing countries where working equines are a key source of traction and transport in local communities.

When Jackson first described sarcoids in 1936 he hypothesised that they were caused by an infectious agent, linking them with warts of cattle caused by BPV (Jackson, 1936). Two decades later various transmission studies began to give support to the idea that BPV is the causal agent of sarcoids. Olson and Cook showed that sarcoids developed in one out of eleven horses inoculated with material from bovine warts (Olson and Cook, 1951), and cell free extracts from naturally occurring sarcoids have been shown to transmit the disease between horses (Voss, 1969, Gobeil et al., 2007). However it appears that equine sarcoids do not give rise to warts when inoculated back into cattle (Ragland et al., 1970). The means by which sarcoids are transmitted between equids has yet to be elucidated, although flies may play a role in spreading the disease.

Since the early studies, many experiments have demonstrated the presence of BPV-1 and -2 DNA in equine sarcoids (for a review see Chambers et al. (2003)) and BPV genes have been shown to be transcribed in sarcoid tissue (Nasir and Reid, 1999). Despite this, mature virions have never yet been identified in sarcoid tissue, which agrees with the situation in other species where tumorigenic PV infection is non-productive for virus particles.

Sarcoids are a widespread and common disease of equids with significant economic and welfare importance. The disease is reported to have different prevalence, clinical features and progression in different parts of the world (http://www.liv.ac.uk/sarcoids/, 2010). Although previous studies have looked into the molecular biology and field epidemiology

of equine sarcoids, no phylogenetic investigation of virus sequences has been published to date. This report presents the first phylogenetic analysis BPV-1 in equine sarcoids. In doing so our aim was to clarify the evolutionary history of the virus, and more specifically to answer the following questions:

- How genetically diverse are the BPV-1 isolates associated with equine sarcoids?
- How long ago did this diversity arise?
- Does the BPV phylogeny show geographic structure?
- Has the species jump to horses happened once or multiple times?

Table 1.1 - Origins of LCR sequence samples

Seq ID	Genbank no.	UK cattle samples	Equine samples							Total
			S. Africa	Italy	Ethiopia	UK	Swiss	Vienna	Brazil	
1			3							**3**
2				1	5				1	**7**
3					5					**5**
4				2						**2**
5				2						**2**
6					2					**2**
7					2					**2**
8					2					**2**
9					2					**2**
10					1					**1**
11				1						**1**
12				1						**1**
13									3	**3**
14		1		1				2		**4**
15	DQ855065						2	3		**5**
16	DQ855067					1				**1**
17	DQ855069						1			**1**
18	DQ855068	7		2		8				**17**
19		4		1				1		**6**
20	DQ855066		1	13	2	4	23	4	2	**49**
BPV-1 ref	X02346	3								**3**
Total		**15**	**4**	**24**	**21**	**13**	**26**	**10**	**6**	**119**

MATERIALS AND METHODS

LCR analysis

Materials

BPV-1 LCR sequences isolated from 119 tissue samples were kindly provided by L. Nasir. The samples had been collected over a period of 20 years from equine sarcoids (n= 104) and from cattle papillomas (n= 15), originating in three different continents and consisting

of 21 unique LCR sequence variants (see Table 1.1). One of these LCR sequence variants corresponded to the BPV-1 reference sample (Genbank accession number X02346) and the rest were assigned numbers from 1 – 20. Five of these numbered sequences had been previously described (Nasir et al., 2007) and their accession numbers are given in Table 1.1. The BPV-2 reference sequence (GenBank accession number M20219) was included as an outgroup.

Methods

The 21 individual LCR sequences in addition to the BPV-2 reference sequence were aligned in Geneious v5.1 (Drummond et al., 2010; available at www.geneious.com) using a global alignment with free end gaps. The entirety of the LCR sequence was then extracted (695bp, located between nucleotides 7252–7947 in BPV-1).

jModelTest (Posada, 2008, Guindon and Gascuel, 2003; available at http://darwin.uvigo.es) was used to identify the best-fitting model of nucleotide substitution for the LCR sequences. The optimum model under the Akaike inference criterion was the K80 model of nucleotide substitution (Kimura, 1980) with a proportion of invariable sites and a gamma-distributed rate variation (K80+I+Γ)

The K80+I+Γ model was used to inform both maximum likelihood (ML) and Bayesian methods to infer phylogenies for the 21 BPV-1 LCR sequences found in the samples plus the BPV-2 outgroup.

ML analysis was carried out in PhyML (Guindon and Gascuel, 2003; available at www.atgc-montpellier.fr/phyml) using the K80+I+Γ model with 1000 non-parametric bootstraps to evaluate statistical support for individual tree nodes.

Bayesian analysis was carried out with the K80+I+Γ model in MrBayes (Ronquist and Huelsenbeck, 2003, Huelsenbeck et al., 2001; available from www.mrbayes.net) using 1,000,000 generations of two simultaneous Markov chain Monte Carlo (MCMC) chains with a sampling frequency of 100 and a burn-in of 2500.

The Path-O-Gen program (available at http://tree.bio.ed.ac.uk/software/pathogen) was used to assess the temporal signal and “clocklikeness” of the LCR sequences. Because PVs evolve on a timescale of millennia, differences in sampling dates here can be considered irrelevant to sequence divergence and tip dates were assumed to be contemporaneous.

A literature search identified two previous studies quantifying the rate of PV evolution. Rector et al. estimated an average mutation rate of 1.95 x 10^{-8} nucleotide substitutions per site per year (95% CI: 1.32 x 10^{-8} to 2.47 x 10^{-8}) based on felid PVs (Rector et al., 2007). They also published evolutionary rates for the individual parts of the viral genome, with the LCR showing the fastest mutation rate at 2.69 x 10^{-8} nucleotide substitutions per site per year (95% CI: 1.75 to 3.69 x 10^{-8}). The second study, by Shah et al., estimated a considerably slower evolutionary rate for PVs. This was calculated using a wide range of host and viral lineages, although only two coding regions were used for the analysis: the E1 gene (with rate 7.10 x 10^{-9} nucleotide substitutions per site per year, SD 1.49 x 10^{-9}) and the L1 gene (9.57 x 10^{-9} nt subs/yr, SD 2.08 x 10^{-9}) (Shah et al., 2010).

Beast (Drummond and Rambaut, 2007; available at http://beast.bio.ed.ac.uk) was used in order to estimate the divergence times for the LCR phylogeny, which was constrained as monophyletic relative to the BPV-2 outgroup. The analysis assumed an Uncorrelated Relaxed Log-Normal clock (Drummond et al., 2006), a HKY nucleotide substitution model as well as a Bayesian skyline model as a flexible demographic prior (Drummond et al., 2005). The evolutionary rate estimated by Rector et al. (2007) was used primarily to impose a normal distribution on the UCLD mean prior, with a mean of 2.69 x 10^{-8} and standard deviation of 5.1 x 10^{-9}. The MCMC simulation was run for a chain length of 10,000,000 with a sampling frequency of 1,000 and a burn-in of 100. The same analysis was then also re-run, using instead the estimate of the evolutionary rate of PVs given by Shah et al. to impose a uniform distribution on the UCLD prior mean with a range of between 7.10 x 10^{-9} and 9.57 x 10^{-9} (Shah et al., 2010).

Tajima's D statistic (Tajima, 1989) was calculated at wwwabi.snv.jussieu.fr/achaz/neutralitytest.html. We calculated the D statistic three times: firstly using only the 21 unique LCR sequences, secondly including all the duplicate sequences and thirdly including all the duplicates with the exception of the Sequence 20 samples. For all three calculations the BPV-2 reference sequence was used as an outgroup. This was carried out in order to assess whether the LCR sequences evolved by random ("neutral") processes or whether there were signs of selective pressures acting on them.

Table 1.2 - Species distribution of L2 samples

	a	b	c	d	e	f	g	**Total**
Bovine	1		3				8	**12**
Equine	2	4	1	1	1	1		**10**
Total	**3**	**4**	**4**	**1**	**1**	**1**	**8**	**22**

L2 analysis

Materials

Seven L2 gene sequences (Sequences a-g) of 352bp were used in this section of the analysis, kindly provided by L. Nasir. These sequences had been identified in both horses and cattle as shown in Table 1.2, and for five of the L2 samples, information on the LCR sequence was also available (details of these are given in Table 1.3). The BPV-1 reference sequence and another partial BPV-1 sequence (Genbank accession number J02045) were included in the analysis, along with the BPV-2 reference sequence which was used as an outgroup.

Methods

The phylogenetic analysis of the L2 sequences was conducted using the programs described above. The sequences were aligned and a 355bp stretch was extracted and the TIM2 model (Posada, 2003) was chosen as best fitting the data. This model was then applied using Bayesian and ML tree building methods as described for the LCR sequences to obtain a phylogenetic tree for the L2 sequences.

E2/E5 analysis

Materials

Genbank was searched for archival BPV-1 and -2 E2/E5 gene sequences. The E2 and E5 genes are adjacent in the BPV-1 and -2 genomes and several studies have sequenced parts of both genes. The accession numbers and origins of these sequences are given in Table 1.4. They consist of seven BPV-2 sequences and seventeen BPV-1 sequences from Canada, UK and South African samples in various equid species.

Methods

Phylogenetic analysis of the E2/E5 sequences was conducted using the programs described above. The sequences were aligned and the coding regions extracted, giving a concatenated stretch containing 299bp. HKY+Γ (Hasegawa et al., 1985) was identified as the best fitting model for the combined E2/E5 sequences and used to construct Bayesian and ML trees as before.

Table 1.3 - L2 and LCR variants for samples where both regions were sequenced

L2 variant	LCR variant
e	5
d	11
b	18
b	20
f	20
b	12

Table 1.4 - Origins of the sequences used in the E2/E5 analysis

Accession no	BPV-1 or -2	Species	Country	Reference
X02346 (BPV-1 ref)	1	?	USA	[42]
FJ648519	1	Zebra	South Africa	[22]
FJ648520	1	Zebra	South Africa	[22]
FJ648521	1	Zebra	South Africa	[22]
FJ648522	1	Zebra	South Africa	[22]
FJ648523	1	Zebra	South Africa	[22]
FJ648524	1	Zebra	South Africa	[22]
FJ648525	1	Zebra	South Africa	[22]
FJ895875	1	Horse	Canada	[43]
FJ895876	1	Horse	Canada	[43]
AY232257	1	Horse	Switzerland	[44]
AY232258	1	Horse	Switzerland	[44]
AY232259	1	Horse	Switzerland	[44]
AY232260	1	Horse	Switzerland	[44]
AY232261	1	Horse	UK	[44]
AY232262	1	Cow	UK	[44]
AY232263	1	Horse	UK	[44]
M20219 (BPV-2 ref)	2	?	USA	Unpub
AY232264	2	Horse	Switzerland	[44]
FJ648526	2	Zebra	South Africa	[22]
FJ648527	2	Zebra	South Africa	[22]
FJ648528	2	Zebra	South Africa	[22]
FJ895874	2	Horse	Canada	[43]
FJ895877	2	Horse	Canada	[43]

The sequences were then separated into BPV-1 and BPV-2 viruses, realigned in Geneious and checked to ensure they were correctly aligned with respect to codons. The Mega program (Tamura et al., 2007) was then used to conduct a codon based Z-test for neutral vs. selective evolution, with 500 bootstraps and H_A specified as $d_n \neq d_s$.

RESULTS

LCR analysis

The Bayesian and the ML phylogenies for the 21 BPV-1 and the BPV-2 reference LCR sequences are shown in Figures 1.1 and 1.2 respectively.

Divergence dates and the molecular clock

The divergence dates of important nodes, calculated using the LCR evolutionary rate from Rector et al. (2007) in Beast, are also shown in Figures 1.1 and 1.2. The confidence intervals for these estimates are large, reflecting both the comparatively low nucleotide diversity of the viral sequences and also the uncertainty in the estimated viral evolutionary rate used. Despite this it can be seen that the node ages in this phylogeny are very old - the most recent divergence date being 53,300 years ago (95% CI: 22,300-209,000 yrs) whereas the root of the tree, the most recent common ancestor of BPV-1 and -2, was estimated to diverge 1,100,000 years ago (390,000-2,560,000 years). Divergence dates were also calculated using the estimate from Shah et al. (2010). These are not shown in the figures, but are considerably deeper: the root of the tree is estimated to have diverged 3,700,000 years ago (95% CI: 1,320,000-9,180,000 years), and the most recent divergence date as 188,000 years ago (95% CI: 32,400-450,000yrs).

The validity of assuming a molecular clock for the sequence evolution was checked initially using the Beast program. Here, the UCLD parameter describes the mean branch lengths of the tree: under a strict molecular clock the variance of the UCLD parameter should be zero. For our sequences the posterior distribution of the UCLD variance included zero, indicating that the molecular clock assumption could not be rejected. The Path-O-Gen program was also used in order to determine the validity of the molecular clock assumption. Because PVs evolve over a timescale of millennia, differences in sampling dates here can be considered irrelevant to sequence divergence. Therefore the variance of the root-to-tip distances can be seen as giving an indication of how "clocklike" the data is. For the LCR sequences the mean root-to-tip distance was 0.083 substitutions per site with a high variance of 0.073. This suggests violation of the assumption of a strict molecular clock and was taken into account in the analysis by using the relaxed clock prior in the Beast.

African and European groups

Both trees consist of a group of predominantly African sequences closer to the root of the tree (consisting of sequences 1, 3, 13, 10, 8, 6, 9 and 2) and a more recently diverged European group (sequences 11, 14, 19, 5, 12, 18, 15, 16, 20 and 4). There is one Italian sample that clusters with the African group, as do the four Brazilian samples. Barring Sequence 20 (discussed below) the European group consists of purely European samples. The clade containing the BPV-1 reference sequence, which consists of both European (Sequence 17) and African (Sequence 7) sequences, takes different positions in the ML compared with the Bayesian phylogeny. Excluding this difficult-to-categorise clade, the diversity of sequence types is significantly different between the two groups ($p<0.001$, $\chi^2=83.2$)

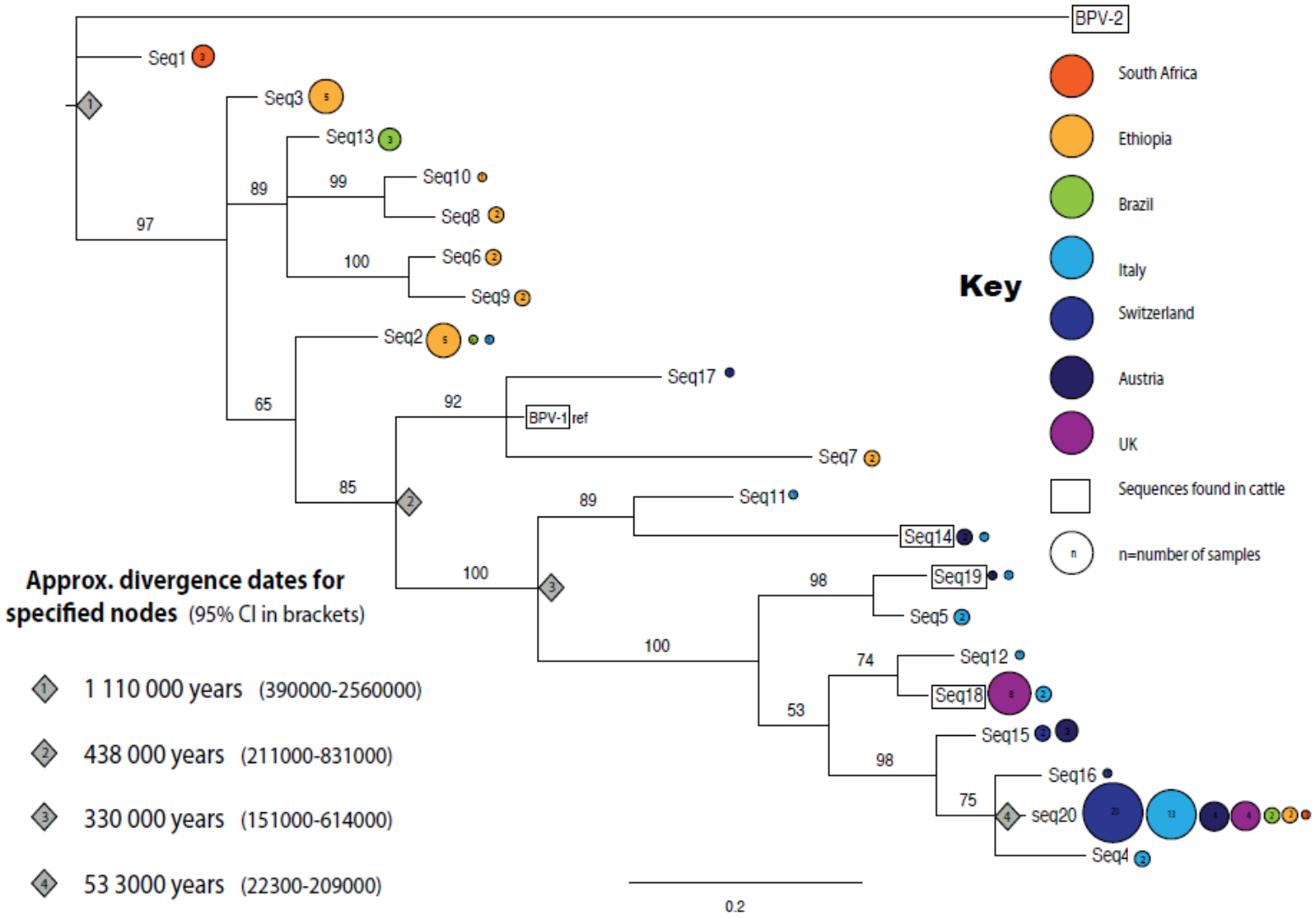


Figure 1.1 – Bayesian phylogenetic tree for LCR sequence variants

Posterior probabilities (%) are given to indicate branch support for the corresponding branches. The scale bar shows genetic distance between the sequences (nucleotide substitutions per site).

The names of sequences that are found in cattle are surrounded by a rectangle. The circles next to each sequence indicate the distribution of the equine samples for that sequence, the colour corresponding with the country of origin and the number within the circle equating to the number of samples containing that sequence variant in that particular country.

The divergence dates for specified nodes (indicated by numbered diamonds) and their 95% confidence intervals as calculated using the evolutionary rate figure from Rector et al. (2007) are also shown

Distribution of cattle samples

Fifteen cattle samples were present in the dataset, all of which originated in the UK. These cattle samples comprise four sequence variants, and all are found within the European group. They appear to be spread throughout these European sequences rather than clustering together as a clade.

Sequence 20

Sequence 20 was the predominant sequence present in the LCR samples. It was found in 49 equids (47% of the equine samples) but not in any of the cattle samples. The absence of Sequence 20 was statistically significant under the null expectation that the frequency distribution of sequence types should be similar in both host species ($p<0.01$, $\chi^2=12.0$). Sequence 20 was part of three-taxa clade within the European group (consisting of sequences 16/20/4) that showed the most divergence from the root of the phylogeny. It is also notable that Sequence 20 was present in equids from all the countries sampled here.

Discrepancy between the Bayesian and ML phylogenies

Although the Bayesian and ML phylogenetic trees are broadly similar, there are two main discrepancies between them.

The first concerns the position of the BPV-1 ref/Seq17/Seq7 clade. In the Bayesian tree it forms a sister clade to the European BPV-1 group, whereas in the ML tree it falls into a more basal position within BPV-1, closer to the BPV-1/BPV-2 divergence.

The second discrepancy is the difference in branch times, the Bayesian branch times being significantly higher than those produced by the ML analysis. Brown et al. describe this as a common problem with the MrBayes program, used here to produce the Bayesian tree (Brown et al., 2010). They suggest a method of adjusting for this whereby the correct branch times (here these would be the ML times) are averaged over the tree, and this value is then used to scale the MrBayes branch times accordingly. This correction was attempted for our analysis, although it still failed to give us the appropriate branch times for the Bayesian tree.

Neutral evolution vs. selection

Tajima's D statistic was calculated to ascertain whether the sequences showed any evidence of selective processes acting during their evolution. The D statistics for the three

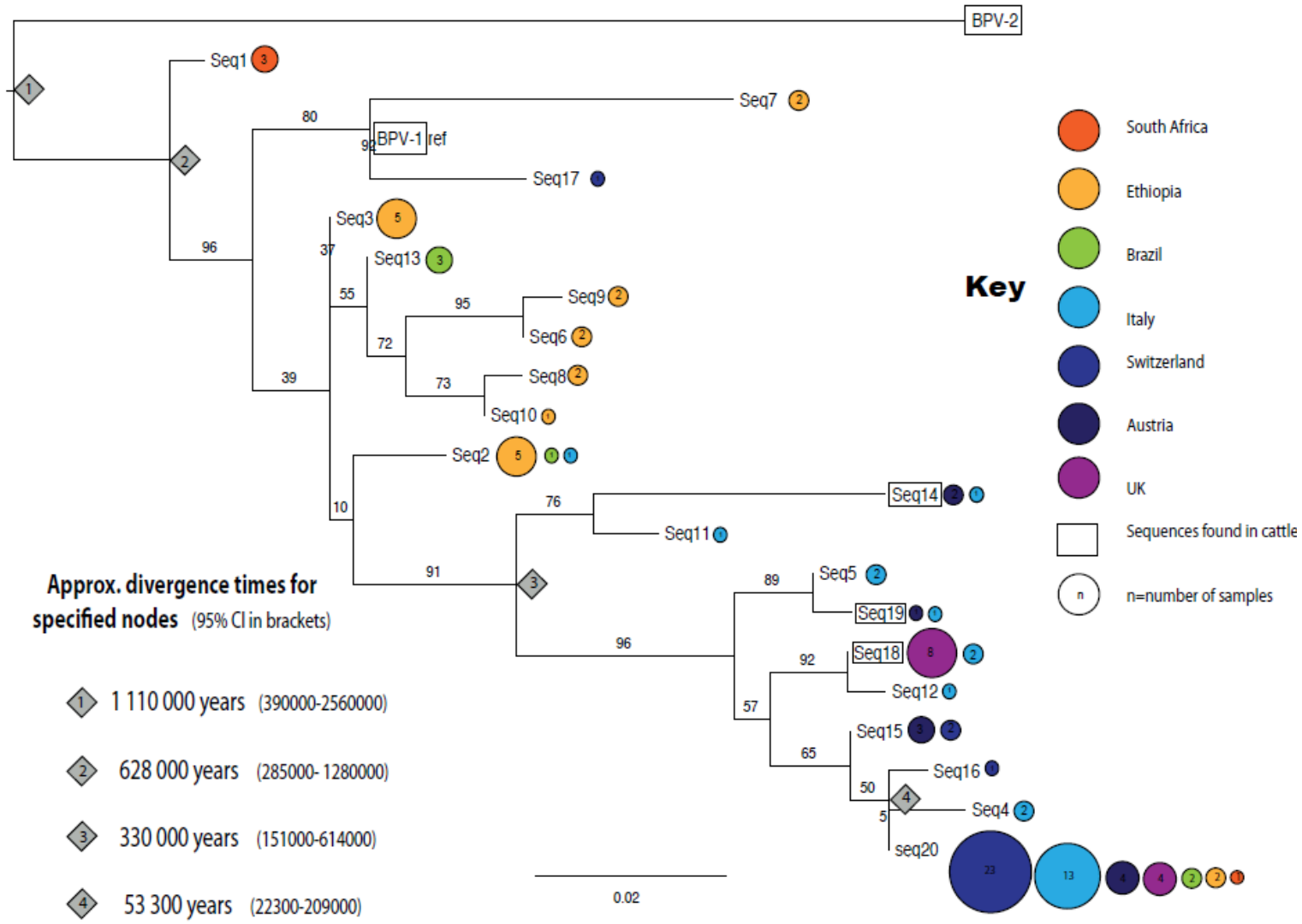


Figure 1.2 – ML phylogenetic tree for LCR sequence variants

Bootstrap values (%) are given to indicate branch support for the corresponding branches. The scale bar shows genetic distance between the sequences (nucleotide substitutions per site).

The names of sequences that are found in cattle are surrounded by a rectangle. The circles next to each sequence indicate the distribution of the equine samples for that sequence, the colour corresponding with the country of origin and the number within the circle equating to the number of samples containing that sequence variant in that particular country.

The divergence dates for specified nodes (indicated by numbered diamonds) and their 95% confidence intervals as calculated using the evolutionary rate figure from Rector et al. (2007) are also shown

different groups of sequences used are shown in Table 1.5. For our sequences the D values were below 2 and p >0.1, therefore the null hypothesis of neutral evolution could not be rejected.

Table 1.5 - Tajima's D statistic calculated for: a) the 21 unique LCR sequences only, b) all of the duplicate sequences, c) all duplicate sequences barring Sequence 20 samples

	D statistic	p value
a) 21 unique sequences	-0.71	0.26
b) All duplicates	-0.46	0.38
c) Duplicates without Seq 20	-0.07	0.55

L2 and E2/E5 analysis

The ML and Bayesian methods of tree-building gave identical morphologies for both the L2 and the E2/E5 gene segments, and these are shown in Figures 1.3 and 1.4 respectively, although the issue of conflicting branch lengths was again encountered. Unfortunately, due to the slow evolution of the virus and the fact that we were only using a relatively short region for these analyses, little

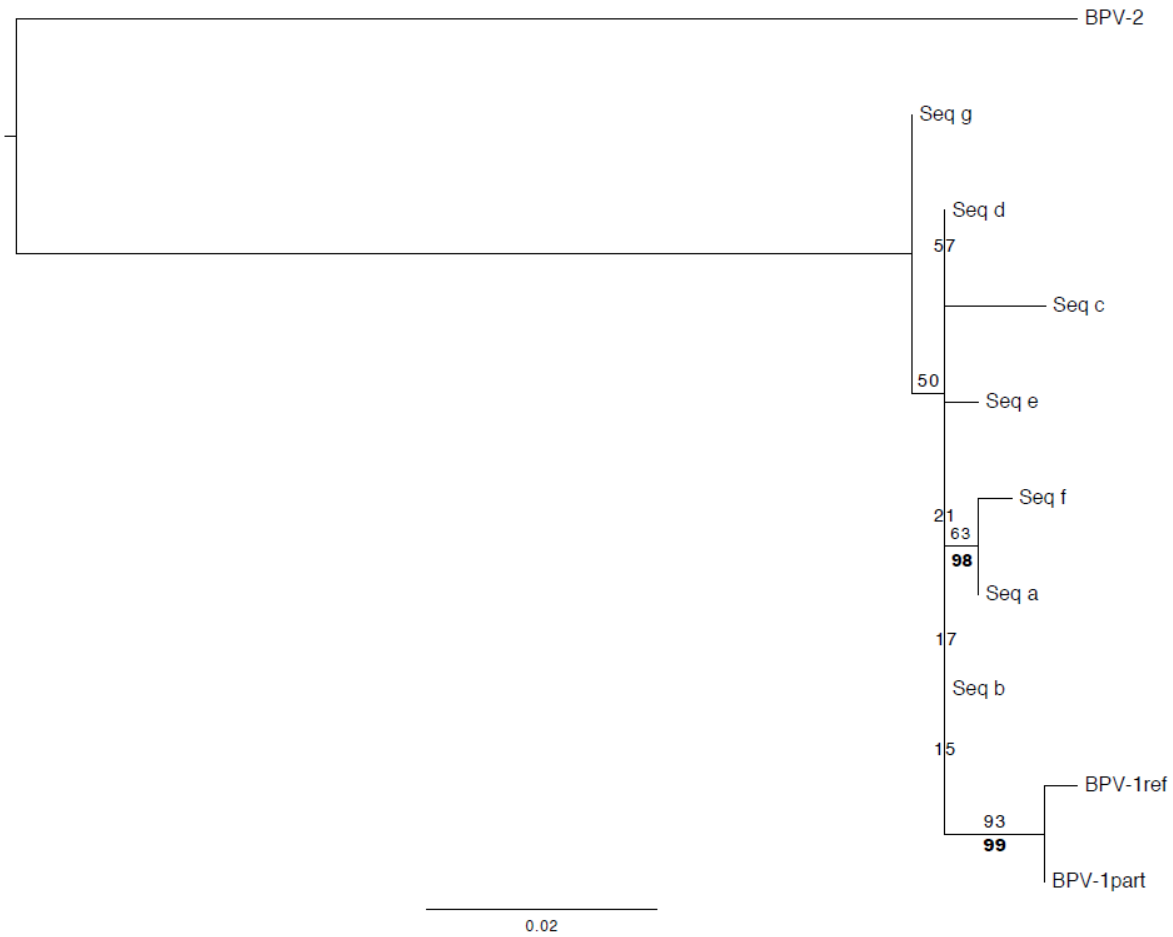


Figure 1.3 - Phylogenetic tree for L2 sequence variants

Bayesian posterior probabilities (bold) and ML bootstrap values are given as percentages at the corresponding branches to indicate statistical support. The scale bar shows genetic distance between the sequences in nucleotide substitutions per site.

variation was seen in the L2 or E2/E5 phylogenies. However neither tree obviously contradicted the more resolved LCR phylogeny.

The codon based Z-test was conducted on the E2/E5 gene segments for the BPV-1 sequences in order to ascertain whether the genes have evolved under selective pressures. This gave a Z value of -1.75 and a p value of 0.083, and for the BPV-2 sequences Z=0.00 and p=1.00. At a p value of >0.05 these results are not statistically significant and no firm conclusion can be drawn with respect selective pressures in these genes.

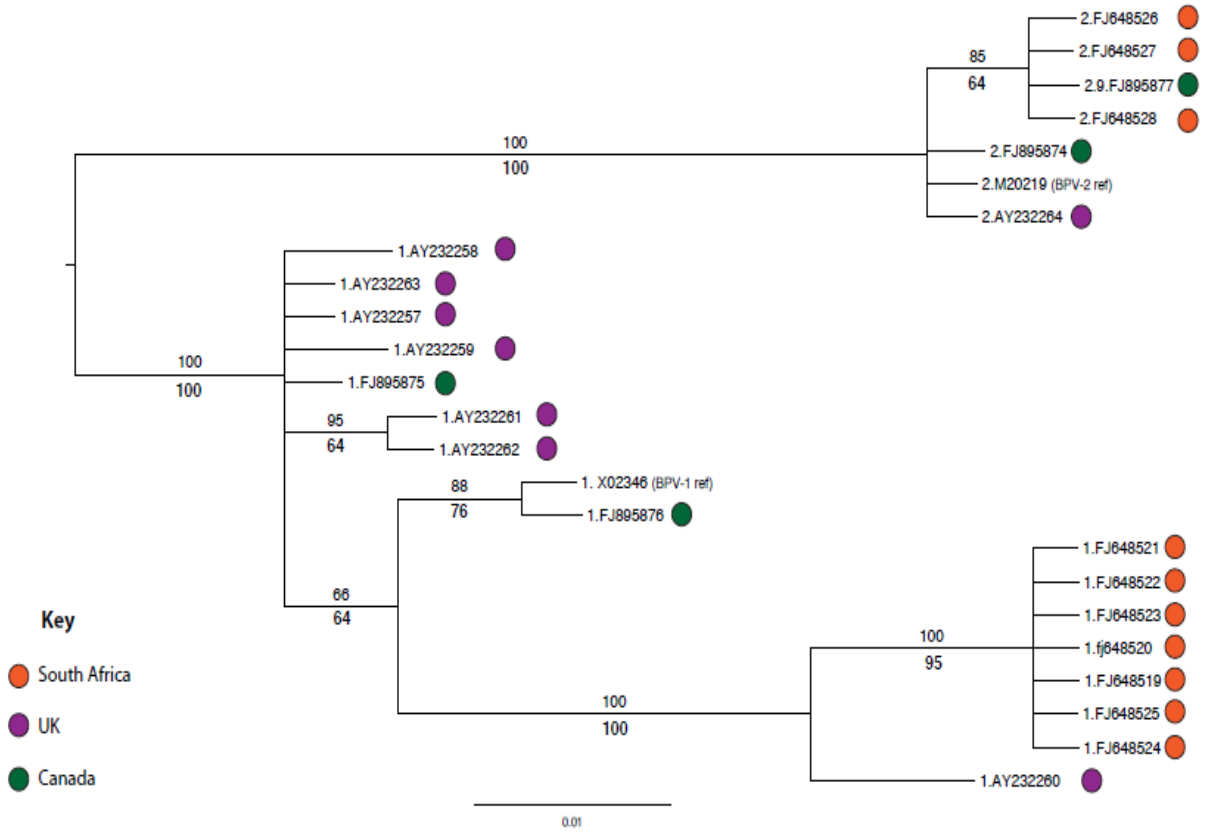


Fig 1.4 - Phylogenetic tree constructed for the combined E2 and E5 sequences

Bayesian posterior probabilities (bold) and ML bootstrap values are given as percentages at the correspondin branches to indicate statistical support. The scale bar shows genetic distance between the sequences in nucleotid substitutions per site.

The circles give information on the country of origin of the sequences and the name of each sequence consists of th BPV strain and the sequence accession number (see Table 1.4)

DISCUSSION

LCR analysis

African and European groups

The phylogeny produced for the BPV-1 LCR region shows clear geographic segregation (Figures 1.1 and 1.2). It can be seen that LCR sequences from Africa and Brazil form a separate, more basal group compared to those LCR sequences originating in Europe, which are derived from the African group. It is interesting to note that whereas African cattle are variable hybrids of the two subspecies of domestic cattle (*Bos taurus* and *B. indicus*) and Brazilian cattle are predominantly *B. indicus*, European breeds consist of pure *B. taurus* stock (Ajmone-Marsan et al., 2010). Therefore the geographical structure of the two groups seen in this phylogeny appears to be broadly mirrored in the geographical distribution of the subspecies of domestic cattle. However, if one considers the large-scale cattle migrations known to have occurred after domestication (and therefore well after the date of the most recent divergence in this phylogeny) (Ajmone-Marsan et al., 2010) it seems unlikely that the two geographic groups in this phylogeny can be explained by a close co-evolution of the virus with the different sub-species of cattle. Another, more plausible explanation might be that the LCR sequence variants isolated from African samples are more fit within *B. indicus* and hybrid cattle at the expense of European variants.

Divergence dates and the molecular clock

Divergence dates calculated from the Rector et al. evolutionary rate are shown in Figures 1.1 and 1.2. It can be seen that the diversity of the LCR sequences is ancient, and despite wide confidence intervals even the most recent divergence (the Sequence 16/20/4 clade, estimated to have diverged 53,000 years ago) substantially predates domestication (horses and donkeys were domesticated around 5000 years ago (Vila et al., 2001, Kimura et al., 2011) and the two subspecies of cattle around 10,000 years ago (Ajmone-Marsan et al., 2010). The date estimated for the divergence of the European from the African group of sequence variants could overlap with the dates estimated for the split between the ancestors of the *B. indicus* and *B. taurus* subspecies of cattle, thought to have diverged somewhere between 33,000 and 2 million years ago.

The divergence dates estimated for any phylogeny rely heavily on the assumption of a constant rate of evolutionary change over time for the sequences involved (the “molecular clock”). As has been mentioned above, it appears that the European sequences are further

from the root of the phylogeny. It is reasonable to assume that for this very slowly evolving virus sampling dates ranging over as little as twenty years (as here) should not affect the expected amount of divergence from the root of the tree. If this is the case it implies that the European group has evolved at faster rate than the African group, thus suggesting that the assumption of a strict molecular clock is not appropriate here, and this is supported by the results of the Path-O-Gen analysis for "clocklikeness". The Beast program allows for such deviations from a strict molecular clock by giving the option of the relaxed clock prior as used here. This deviation from a clocklike mutation rate suggests that selective pressures have been acting on BPV-1 evolution, either though positive selection in the European group, or negative (purifying) selection in the African sequences. Tajima's D statistic was calculated in an attempt to identify whether selective pressures were acting on the sequences,

The accuracy of the divergence date estimates is also greatly dependent on the value used to define the rate of the molecular clock. Here we had a choice of two evolutionary rate estimates for PVs. Rector et al. (2007) calculated the evolutionary rate of felid PVs based on co-speciation with their hosts. This assumption of co-speciation was supported by congruence between viral and host trees and by the geographical isolation of the species involved. The paper gave separate rate estimates for each part of the PV genome, putting the rate for the LCR region at 2.69×10^{-8} nucleotide substitutions per site per year (95% CI: $1.75 - 3.69 \times 10^{-8}$). Shah et al. (2010) similarly used host speciation dates to calculate an estimate for the evolutionary rate of PVs of $7.1\text{-}9.7 \times 10^{-9}$ substitutions per site per year, but over a much wider range of viruses and host species. Their analysis is potentially based on more solid foundations than Rector et al. as their calculated rate was averaged over a much wider range of viral lineages, and they did not fully constrain the phylogeny to host-virus co-speciation. However, their estimates of evolutionary rate were calculated for coding genes only with no reference to the faster-evolving LCR.

Both values were used here to calculate divergence dates for the LCR phylogeny. The divergence dates using the Rector et al. value are shown in Figures 1.1 and 1.2 while results using the Shah et al. figure are given in the Results section (the latter, slower, rate estimate predictably gives considerably older node ages, ranging between 3,700,000 and 188,000 years ago). We feel the Rector et al. figure is more appropriate to this phylogeny for two reasons. Firstly it provides a specific rate estimate for the LCR region. Secondly, Ho and Larson suggested that the evolutionary rate of any one organism rapidly declines with increasing time away from the present (Ho and Larson, 2006). For this analysis into

relatively recent viral diversity, an evolutionary rate estimated from a more recent phylogeny is more suitable. Even then, Ho and Larson suggest that it is possible that the divergence dates may be overestimated by up to a factor of ten (Ho and Larson, 2006), although this still puts the majority of the BPV-1 divergence dates before the domestication of cattle.

Distribution of cattle samples

All the cattle samples here originated in the UK, and although there are comparatively few cattle samples present in this dataset (15 cattle out of a total of 119 samples) it can be seen that they are spread through the European group of sequences within the LCR phylogeny. If cattle-horse transmission only happened once the cattle samples would be expected to group together, with purely equine sequence variants being derived from one of the cattle sequences - this is not the pattern seen here. Rather than clustering within one clade at the base of the phylogeny, they appear to be mixed randomly amongst the equine samples in the European group, and sequence variants found in cattle are also often also present in the equine samples. This suggests that cross-species transfer of BPV-1 between cattle and equids has occurred multiple times over the course of its evolutionary history. This may mean that BPV-1 has been transferred multiple times from a variety of cattle strains into horses in the course of its evolutionary history, or alternatively that the virus has diverged within equids followed by transfer back into cattle. However experimental evidence suggests horse to cattle transfer is unlikely (Ragland et al., 1970) therefore the hypothesis of multiple cattle to equid transfer events is more plausible (see below).

Based on the above, we can theorise that if one were to search for BPV in African cattle they should contain LCR sequence variants similar to the African sequences in this phylogeny.

Host-species transfer

The distribution of the cattle samples within the LCR phylogeny is not consistent with the theory that BPV-1 was transferred from cattle to equids only once, and that following this, the genetic diversity seen in these equine samples subsequently evolved solely within equids. Instead we hypothesise that, based on the pattern of host species within the phylogeny and the known biological behaviour of the virus, the nucleotide diversity seen among the LCR sequences probably evolved within the cattle population and the virus was later transferred into equids multiple times from a variety of different seeding variants.

The molecular biology of the virus within cattle and equids appears consistent with this theory of host-species transfer. It is known that in cattle BPV papillomas are productive, generating infectious viral particles. No virons have ever been demonstrated in equine sarcoids and it is not known how sarcoids are transmitted between equids in the natural environment. It appears reasonable to suppose that BPV-1 is more transmissible and more likely to be maintained (i.e. has a higher basic reproduction number) in cattle populations than it is in equids. Olson and Cook were able to produce sarcoid in horses from cattle papilloma material in one out of eleven attempts (Olson and Cook, 1951), suggesting that the barrier to cross species transmission of BPV from cattle to horses is not high. In light of the above it is plausible that BPV-1 is preferentially maintained in the cattle population, with a relatively high occurrence of cross-species transmission to equids but with relatively poor transmissibility once the cross-species jump has occurred.

Sequence 20

Sequence 20 may be the exception to this theorised low transmissibility within equids. Sequence 20 is present in 47% of the equine samples in this study but not in any cattle samples, and was found in equids from all countries sampled. This suggests that this sequence variant may have a particular selective advantage within equids. Nasir et al. have already demonstrated that the Sequence 20 LCR (previously named Variant II) has significantly higher transcriptional activity in equine fibroblasts than the BPV-1 reference LCR (Nasir et al., 2007), and natural equine transmission has also been demonstrated for Sequence 20-containing virus (Nasir and Campo, 2008).

As Sequence 20 is one of the most recently evolved variants of BPV-1 it is tempting to postulate that it has evolved adaptively to the equine environment. However, given BPV is a very slowly evolving virus and that according to the divergence dates calculated here any variation in the LCR region occurred tens of thousands of years before the domestic horse or donkey existed, any advantage this variant has within modern equids is likely to be coincidental rather than adaptive.

Further work

- More samples of BPV-1 from cattle are needed to confirm host species transfer theories, ideally sampling cattle LCR variants originating outside Europe. This would enable investigation of whether BPV-1 in cattle segregates with other cattle variants, or with geographic region regardless of host species

- BPV-1 samples from other continents would help refine our understanding of the geographical structure in the LCR phylogeny.
- In a very slowly evolving virus such as BPV, lack of temporal resolution is a problem (Wirth et al., 2005). This is demonstrated here in the low significance of the results gained from the Tajima's D analysis and the wide confidence intervals around the divergence date estimates. If another gene were to be sequenced for the LCR variants used here, and if there were no signs of intergenic recombination, the two regions could be combined and the phylogenetic analysis repeated to create stronger statistical support for the results presented here.
- Further molecular characterisation of Sequence 20 should give a better understanding of the molecular biology of BPV-1 in equine sarcoids, and the of specific determinants that make this variant so successful in equids.

L2 and E2/E5 analysis

The phylogenetic trees derived for the L2 and E2/E5 gene segments are shown in Figures 1.3 and 1.4 respectively. Unsurprisingly due to the slowly evolving nature of the virus and the short segments used for these analyses, the conclusions that can be drawn from them are limited. Branch supports derived for the L2 tree are variable and the genetic diversity seen in both analyses is too limited to show any definite patterns. However, neither analysis appears to contradict any of the conclusions drawn from the LCR tree detailed above. In fact, Table 1.2 shows that several of the sequences used in the L2 phylogeny were found in both cattle and horse samples and this supports the theory that the BPV-1 virus has crossed the host species barrier multiple times from several different seeding variants.

Chapter 2

Scottish sheep movements and their potential for disease transmission

Supervised by Prof Rowland Kao
(rowland.kao@glasgow.ac.uk)

Chapter 2 - SCOTTISH SHEEP MOVEMENTS AND THEIR POTENTIAL FOR DISEASE TRANSMISSION

INTRODUCTION

Controlling livestock disease is an expensive business. Defra quote the cost of an exotic disease outbreak as between £2 million for a minor outbreak to £3 billion for a major outbreak, and predict that a "major unknown infectious disease" will occur around once every eight years at an annual cost of £64 million (www.defra.gov.uk/food-farm/animals/diseases). The control of endemic disease is also expensive, for example £28 million was spent on the National Scrapie Plan in 2004-2005 (www.farmersguardian.com/national-scrapie-plan%92s-future-is-under-review/5757.article)

If it were possible to integrate the control of livestock infections such that a single control measure could target multiple diseases, the result would be an economical and efficient way to ease the burden of farm animal disease. In this project, a preliminary investigation was undertaken using the principles of contact network theory to evaluate the feasibility of targeting one subset of sheep farms for the control and prevention of multiple diseases.

Contact networks and their application to livestock disease

A network consists of a group of points ("nodes") and the connections between them. The UK livestock movement database, in which movements of sheep and other livestock have been compulsorily recorded since 2002, provides ideal data for the construction of a contact network in which the nodes are individual farms or animal holdings, and the connections between them consist of directional movements of livestock. Livestock movements are a major mechanism by which disease can be transmitted through a population (Fevre et al., 2006), and the importance of animal movements in propagating a disease outbreak is demonstrated by the role of sheep movements during the initial stages of the 2001 UK foot and mouth disease outbreak (Gibbens et al., 2001)

Mathematical modelling of the spread of disease within a population has important applications to the prevention and control of infectious disease. Traditionally, mathematical models of disease spread assumed a homogeneously mixing population, however the availability of more detailed data allows for the application of network theory to disease modelling. This approach explicitly incorporates the contact structure of individuals or farms within a population into the analysis of the spread of a disease. The

Table 2.1 - Table showing infectious diseases of sheep, assigned to one of three categories: fast transmitting diseases where transmission occurs in days to weeks, medium transmitting diseases where transmission occurs over months, and slow transmitting diseases where transmission occurs over years. Diseases highlighted in red are notifiable in Scotland

Constructed with the help of D. Logue

		Fast (wks/days)	**Medium** (mths)	**Slow** (yrs)
Direct	Foot and mouth disease	X		
	Pasteurellosis	X		
	Watery mouth	X		
	Lamb dysentry	X		
	Rotavirus	X		
	Salmonella diarrhoea	X		
	Contagious foot rot	X		
	Erisipelothrix	X		
	Orf	X		
	Coccidia	X		
	Cryptosporidium	X		
	Caseous lymphadenitis		X	
	Scrapie			X
	Jaagsiektie			X
	Maedi-visna			X
	Johnes			X
Breeding-related	Enzootic abortion		X	
	Toxoplasmosis		X	
	Q-fever		X	
	Border disease		X	
	Salmonella abortion		X	
Vector borne	Bluetongue virus		X	
	Louping ill		X	
	Tick pyaemia			X
Macroparasites	Sheep scab	X		
	Lice	X		
	Parasitic Gasteroenteritis		X	
	Fasciola			X

underlying contact structure (here represented by movements between farms) provides a framework along which a disease can potentially transmit, and along which it is possible to simulate its spread. Various papers have investigated the contact network of British sheep movements, characterising the properties of the network itself and/or modelling disease spread through the network (Kao et al., 2006, Kao et al., 2007, Kiss et al., 2006, Webb, 2005, Webb, 2006, Volkova et al., 2010). Kiss et al. (2006) showed significant heterogeneities in the number of movements associated with different farms within the British sheep industry. Most farms are involved in few movements, while there is a small

proportion of farms that have high numbers of movements and which are therefore disproportionately important within the contact network. Volkova et al. (2010) emphasised this by demonstrating that removing the 20% most important farms within a contact network of all Scottish sheep movements it was possible to reduce the size of a simulated disease epidemic by more than 80%.

Different diseases spread via different mechanisms, and therefore different types of contact are important in their spread. This essentially means that the contact network underlying the spread of one disease may be very different from that involved in the spread of another disease. Table 2.1 lists several infectious diseases of sheep and the timescale over which they transmit. For a highly infectious disease that transmits rapidly, such as foot and mouth disease, all movements could be important for the spread of the infection as a contact of any duration with an infected animal could transmit the infection. This would therefore involve a network referred to here as the "short-stay network", as it consists of any movement of sheep including those resulting in only a short-stay. In contrast, for scrapie, where a significant amount of disease transmission is believed to be associated with infected placental material (and therefore lambing) (Boden et al., 2010), only movements for breeding purposes resulting in long term stays on the destination holding, would be expected to be important in spreading the disease. It would be expected for other slowly transmitting diseases also that only movements resulting in extended stays (the "long-stay" network) would be important, as a movement resulting in a stay of short duration is unlikely to result in a transmission event and therefore unlikely to propagate the infection.

The British sheep industry

As British sheep movements are recorded at a batch level only, the reasons behind them are difficult to evaluate, and knowledge of the industry is necessary to interpret them. The sheep industry has a markedly stratified structure that has evolved to make the best use of the varying quality of British farmland while producing good quality lamb for the table. The traditional stratification of the industry is based around hill, upland and lowland farms and follows the pattern described below (Arnold et al., 2002):

- Hill: hill breeds are kept for their hardiness and, while young, hill ewes are well adapted to rear one lamb a year under the harsh conditions and poor grazing of the hill pastures (marginal land which would otherwise remain largely unutilised).

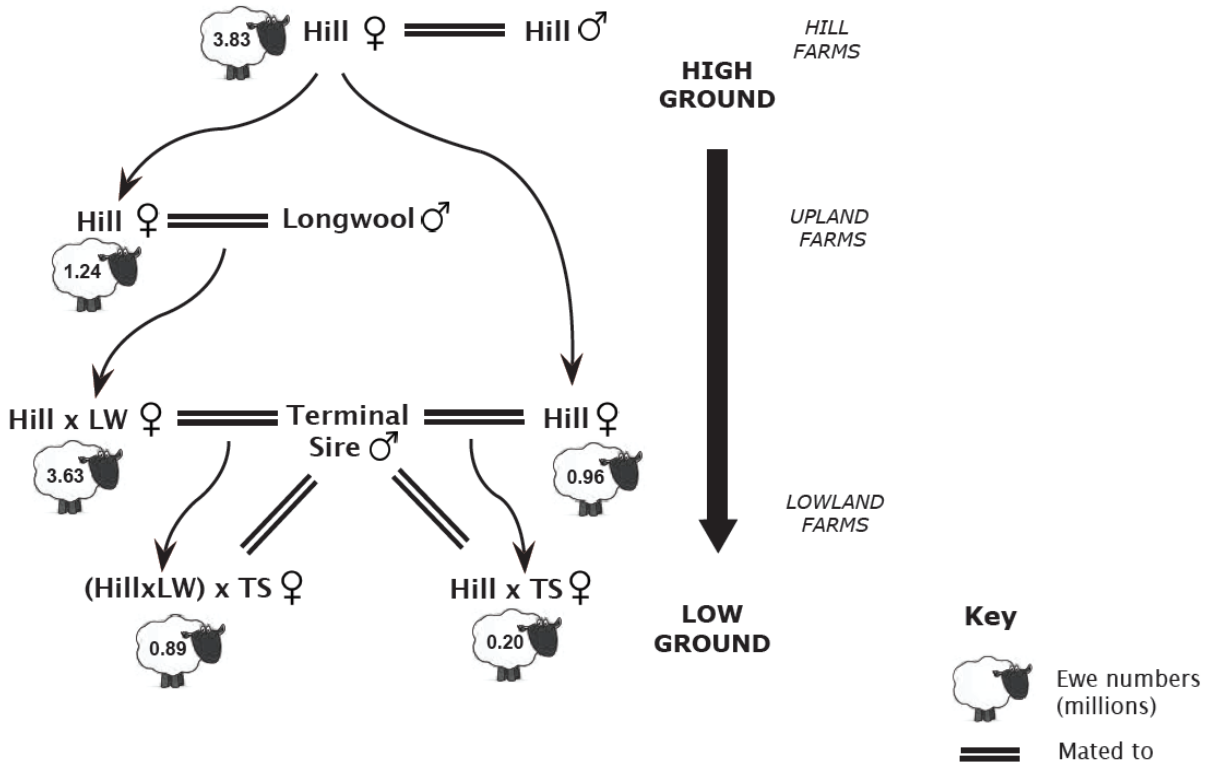


Figure 2.1 - Overview of the British sheep industry, giving numbers of ewes (millions) at each stage.

Adapted from Pollott and Stone (2003)

- Upland: after four to five years on the moors, older hill sheep (known as "draft" ewes) are moved down to the kinder conditions of the upland pastures. Here they are mated to longwool rams to produce first-cross ewes with good mothering ability.
- Lowland: these first-cross ewes are mated to the well-fleshed terminal sire breeds in the lowlands, producing the end product - a well mothered second-cross lamb, which grows fast and gives a good quality carcass on the lush grazing of the lowlands.

The actual structure of the industry is in reality more complex, and this is shown in Figure 2.1 (adapted from Pollott and Stone (2003)). It can be seen from the numbers of ewes at each stage in Figure 2.1, that although the above description does make up the backbone of the industry in terms of ewe numbers, there are also significant numbers of hill ewes mated directly to terminal sire breeds, and also second-cross ewes kept on to be mated to terminal sire rams. In addition to this complexity, several breeds of sheep operate in a self-contained manner, separate from the breeding structure described above.

The timing of sheep movements is driven by the strong seasonality of the sheep reproductive cycle. There are marked increases in sheep movements in autumn due to

movements of lambs (both "finished" lambs going direct to slaughter and "store" lambs sold to be fattened on a different farm) and ewes (including the draft ewes mentioned above, as well as cull ewes going to slaughter). A summary of the types of sheep involved in movements between the different categories of farm (hill, upland or lowland) is given in Table 2.2, although it must be remembered there is often a degree of overlap between the three different categories of farm.

Table 2.2 - Table giving a summary of the different types of sheep involved in movements between different sectors of the British sheep industry

Constructed with the help of F. Houston, D. Leggat and D. Logue

Departure	Arrival	Sheep
Hill	Hill	Breeding sheep
	Upland	Draft ewes
		Finished lambs
		Store lambs
		Pasture moves
	Lowland	Draft ewes
		Finished lambs
		Store lambs
		Pasture moves
Upland	Hill	Pasture moves
	Upland	Breeding sheep
	Lowland	Draft ewes
		Finished lambs
		Store lambs
		Pasture moves
Lowland	Hill	Pasture moves
	Upland	Pasture moves
	Lowland	Breeding sheep

It can be seen from the above overview of the industry and from Table 2.2 that there exists in general a flow of sheep movements down the hill. Movements uphill are much rarer and are generally associated with "pasture moves", consisting of young hill ewes which are sent to spend their first winter in the uplands or lowlands, and moved back up to the hill pastures in spring. Movements between farms within the same sector (i.e. hill-hill, upland-upland, lowland-lowland) are expected to represent sheep moved for breeding purposes. Hill flocks are largely self-sustaining and tend to breed their own replacement ewes, while also providing a source of sheep for farms further down the hill, although even mainly closed flocks do tend to buy in some breeding stock (McLean et al. 1999). Both the stratification of the sheep industry and the flow of sheep from hill to upland and lowland farms are important factors when considering the spread and control of disease within the British sheep flock.

This project

This project was undertaken to identify the sheep farms crucial to the spread of both fast-transmitting and slow-transmitting diseases, in order to assess whether targeting a subset of sheep farms might be a valuable way to control multiple diseases. As discussed above, fast- and slowly-transmitting diseases can be thought of as spreading along different

underlying contact networks, referred to here as the "short-stay" and "long-stay" networks respectively. The aim of this project was to identify the sheep farms and the associated sheep movements making up the short-stay and long-stay networks for the Scottish sheep industry, and to identify each farm's importance within each of the two networks. The project was confined to Scottish movements as Scottish movement data explicitly identify farm-to-farm moves, in contrast to movement data from the rest of Britain where the farm of final destination is not indicated if the animals were moved via a market holding. Thus Scottish sheep movement data are much better suited to this project, where farm-to-farm movements specifically are under consideration.

The project consisted of three major sections:

1. Initial review of the British sheep industry. This was conducted to consolidate current knowledge of the industry with regards sheep movements. Sheep movements are recorded in terms of numbers of animals moved rather than identifying individual sheep, therefore an understanding of the industry was essential to clarify the potential reasons for different movements. This section also allowed evaluation of the appropriateness of the project aims and datasets. Problems identified here with the original project aims led to the restriction of this project to hill flocks only.

2. Descriptive investigation of Scottish Animal Movement Licensing Scheme (SAMS) data. This enabled summary data on the movements between Scottish sheep farms to be collated and assessed against knowledge of the industry gathered in section 1. The analyses in this section provided the basis for the work done in section 3.

3. Values for farm network importance. The numbers of sheep movements involved in the short-stay and long-stay networks for Scottish hill flocks were extracted. These were used to calculate the importance of each farm in each of the networks. A comparison of the short-stay and long-stay network importance values for each farm enabled the identification of a subset of farms important in both networks. Temporal comparisons were also undertaken to assess whether the importance of a farm was predictable year to year.

MATERIALS AND METHODS

All the analyses described below were carried out in MS Access 2003 unless otherwise stated.

Scottish Animal Movements Scheme (SAMS): The SAMS animal movement database (www.scotland.gov.uk) gives details of movements of sheep, pigs and goats in Scotland. The movements are recorded on a daily basis at a batch level, giving information on the numbers of animals moved from premises to premises, without identifying the animals individually.

Each animal holding is represented by a unique identifier giving details of the County, Parish and individual Holding (the CPH number). The SAMS data give the CPH numbers of the departure and arrival holdings involved in the movement, the date of the movement, the number and species of animals moved and the CPH number and date of any market that animals passed through during the movement.

SAMS movement data were filtered using the "County" identifier to include only movements of sheep with Scottish arrival and departure holdings and for dates between 01/03/2008 and 28/02/2010. This gave 371,180 movement records spanning two years, and avoided the period 03/08/2007 – 31/12/2007 when movement restrictions were in place due to a Foot and Mouth Disease (FMD) outbreak.

Animal Movements Licensing Scheme (AMLS) list of premises location types: The British government also hold information on the type of premises for all British CPH numbers including those in Scotland (www.defra.gov.uk), along with details on their location.

Ram Genotyping Scheme (RGeS): The RGeS (www.defra.gov.uk) is a voluntary government scheme where rams are genotyped to evaluate their genetic susceptibility to scrapie. It includes the CPH number giving the rams' location and the breed(s) of the rams genotyped. Ram breed is a good indicator of whether a farm runs a hill, upland or lowland flock. Because the aim of the RGeS is to encourage the development of a national scrapie resistant flock, it is presumed that any genotyped rams are intended for breeding purposes, and thus should be representative of the main breed in their resident flock.

Hill Codings: Prior to this study RGeS participants had been classified as hill flocks or not-hill flocks (i.e. upland and lowland) according to ram breed present. All other, non-RGeS CPHs were then also classified as hill or not-hill based on the classification assigned to RGeS participants in their area. These data were provided for use here courtesy of P.

Bessell and L. Boden. Geographical proximity to an RGeS-classified flock was shown to be a better predictor of flock type than other meteorological or geographic parameters.

These data were linked to the AMLS premises data by CPH number, and modified so that hill and non-hill categories included only CPHs classed as Animal Residences (i.e. farms) in the AMLS premises data. Another category, "Other", was added for CPH numbers that were not included in the original coding data, or that were not listed as Animal Residence holdings in the AMLS premises data. This modified list of hill codings consisted of 39,254 hill CPH numbers, 537,158 non-hill CPH numbers and 21,939 "other" CPHs.

1. Initial review of the British sheep industry

An initial review was conducted to clarify the knowledge of the sheep industry, with particular relevance to sheep movements, and with the aim of validating the objectives and assumptions made in this project. Three areas were investigated in more depth as follows:

Hill vs. not-hill categories: The division of farms into hill and not-hill categories described above overlooks a variety of potentially important movements between upland and lowland farms. To assess the appropriateness of the hill/not-hill categories compared with hill/upland/lowland categories, sheep breeds with over 50 rams sampled in the RGeS, or those which were listed in the 50 most common British ram breeds (Pollott and Stone, 2003) were identified. Each breed's hill/not-hill classification was then re-evaluated and the breeds previously classified as "not-hill" were assigned either upland or lowland status with help from D. Logue and J. Vipond.

The revised list was linked with the RGeS dataset on breed, and arrival and departure CPHs were classified as hill, upland or lowland. This was then used to assign hill, upland or lowland status to the arrival and departure CPH numbers present in the SAMS movement dataset and to define the types of movement taking place (potentially either Lo-Lo, Lo-U, Lo-H, U-U-U, U-U, U-H, H-Lo, H-U or H-H where Lo=lowland, U=upland and H=hill). The nature of the relationship used in MS Access to link RGeS to SAMS ensured that only movements involving farms present in both datasets were included. Numbers of movements and numbers of sheep moved were then calculated in total and by season (as described in section 2 below) in order to assess whether the movement signatures of upland flocks were similar to hill flocks or lowland flocks, and therefore whether combining upland and lowland flocks into the one "not-hill" category was appropriate.

Pasture moves: This term refers to movements of sheep most likely to be associated with over-wintering of hill ewe-lambs in the lowlands. These were isolated by identifying linked pairs of CPHs moving sheep from hill to upland/lowland farms in autumn, and from upland/lowland to hill farms in the spring.

Draft ewes: Once pasture moves were accounted for, the numbers of draft ewes involved in hill to upland/lowland movements was estimated using values identified in section 2 for numbers of sheep participating in pasture moves, hill to non-hill moves and hill-to-slaughter moves (14,8623, 456,175 and 111,339 sheep per annum respectively), and based on the following assumptions:

- Ewes are removed from the hills at an average of five years old and therefore the number of ewes moved should equal approximately 20% x flock size. These ewes consist of 75% draft ewes sold to upland/lowland flocks and 25% cull ewes going to slaughter.

- One lamb is raised per breeding ewe per year in hill flocks. Given that a ewe does not breed in her first year of life, and that 20% x flock size of lambs raised will be kept as replacements on their farm of birth, the number of lambs moved to hill flocks should number 60% x flock size. These will include store lambs sold to upland/lowland flocks for fattening and finished lambs moving directly to slaughter.

- Sheep moved from hill to upland/lowland farms potentially consist of draft ewes, store lambs and pasture moves

- Sheep moved from hill farms to slaughter should be made up primarily of finished lambs and cull ewes

2. Descriptive investigation of SAMS movement data

Summary values: The arrival and departure CPH numbers for animal movements present in the filtered SAMS movement data were assigned a numerical coding according to whether they were listed as hill, not-hill or other in the modified Hill Codings dataset. The information on the type of arrival and departure premises was then combined and each movement was classified as being one of nine types of movement: UL-UL, UL-H, H-UL, H-H, UL-other, H-other, other-UL, other-H, other-other (where H=hill and UL=not-hill).

Values for number of movements and number of sheep moved were calculated for each type of movement over the two years present in the data. The AMLS premises data were also used to identify the different types of premises involved in each movement category. Potential reasons for each of the different movement types were defined with the help of D. Logue, F. Houston and D. Leggat.

The filtered SAMS data, in combination with the associated movement category for each movement, were then broken down by season based on movement date. Spring was taken to include the months of March, April and May; summer as June, July and August; autumn as September, October and November; and winter as December, January and February. Season was used instead of quarter as it has previously been shown as a more meaningful division of the sheep calendar (E.Waugh, unpublished results).

3. Values for farm network importance

The possibility of a farm within either the short-stay or long-stay networks becoming infected with a disease can be said to be proportional to the number of movements (or of individual sheep) moving on to that farm. The possibility of a farm transmitting an infection once infected is proportional to the number of movements or of sheep moving out of that farm. Therefore the importance of a farm in spreading disease through the network in question can be calculated by multiplying the number of movements or sheep coming in to that farm (proportional to the probability of becoming infected) with the number of movements or of sheep moving off the farm (proportional to the probability of spreading that infection). This concept is derived from work on highly active individuals involved in the spread of human sexually transmitted infections (Anderson and May, 1991) and is referred to here as the "importance" of a farm within the network. It was calculated here in terms of numbers of movements and numbers of sheep moved, for both the long-stay and the short-stay networks as detailed below.

Long-stay hill network: The long-stay network of hill flocks was defined as consisting of hill-to-hill movements only. These movements were isolated and the farms taking part in them were identified for the whole two-year period (n=1,925 farms over the two year period). The product of the numbers of movements or of sheep moving into and out of a farm was calculated, thus giving the farm's importance to the network both in terms of movements and in terms of sheep numbers.

Short-stay hill network: The short-stay network of hill flocks was defined as above but with the addition of movements to or from upland or lowland farms. The short-stay network consisted of 2,849 hill farms over the two-year period. Movement and sheep importance values were calculated for these as for the long-stay network, but with the addition of a weighting for market importance. Market weighting was added because, for acute diseases, movements through markets are more important in spreading the disease than direct farm-to-farm movements would be. Market importance was defined as the number of sheep present on all market days connected by movements to the farm of interest multiplied by 0.004 (the latter value estimated by Green et al. (2006) to approximate the risk of transmission of foot and mouth disease should an infected animal be present at a market). The value for short stay importance is therefore defined as follows:

short-stay importance = in degree x out degree + market weighting

Importance rankings: These importance values for farms in terms of sheep and of movements in the long-stay and the short-stay networks were combined and exported to MS Excel 2003. Here the importance values were ranked and the farms present in the 20% most important long-stay and short-stay premises were identified.

Seasonal importance values: The steps described above were also carried out individually for the eight seasons present in the filtered SAMS movement data.

RESULTS

1. Initial review of the British sheep industry

A review of current knowledge of the British sheep industry was conducted, with particular respect to movements between farms. Table 2.2 shows the type of sheep expected to be involved in movements between different sectors of the sheep industry. Particular attention was paid to the two following areas in order to validate the assumptions and aims of the project:

Hill vs. not-hill categories: Following the recoding the SAMS farms present in the RGeS data as either hill, upland or lowland according to breed of ram sampled in the RGeS, the types of movement between farms were then re-categorised using three categories of farm. Movements could be Lo-Lo, Lo-U, Lo-H, U-U-U, U-U, U-H, H-Lo, H-U or H-H (where Lo=lowland, U=upland and H=hill). We calculated χ^2 values for comparisons between

upland movements and hill or lowland movements in order to assess whether movements involving upland flocks were significantly different from others not involving upland farms. The χ^2 values for all comparisons, both in terms of total numbers over the two-year period and when the movements were stratified by season, were significant at $p < 0.001$. Movements involving an upland farm appeared to be as different from hill movements as they were from lowland movements in terms of the magnitude of the χ^2 value. These high levels of significance may be in part due to the large numbers involved in the comparisons, but none the less these results may raise questions as to the appropriateness of the hill vs. not-hill division.

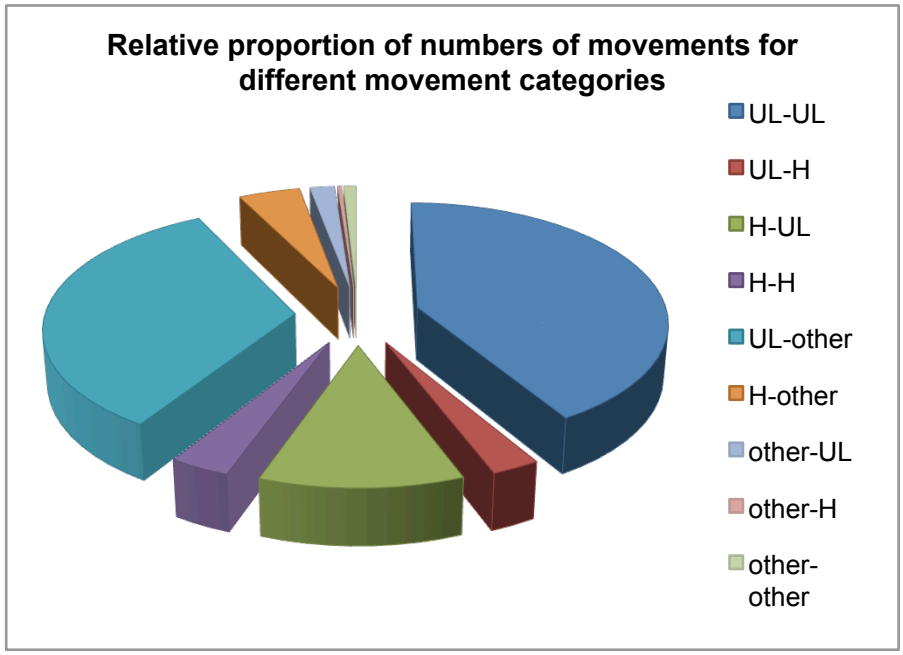
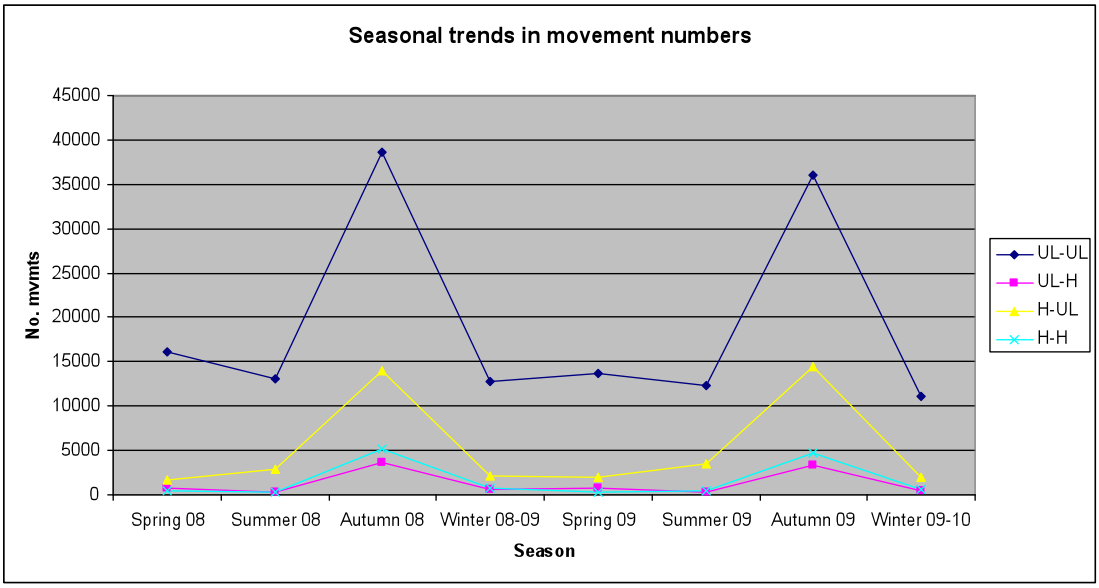


Figure 2.2 - The relative contributions of different types of sheep movements to the total movements over the two-year period
H = hill farm
UL = upland/lowland farm
Other = non-coded or non-farm animal holding

Estimated number of draft ewes: "Draft ewes" are defined as older hill ewes that are moved to upland/lowland flocks to produce crossbred progeny. The numbers of draft ewes taking part in the hill to upland/lowland movements were calculated using the numbers of sheep moved averaged over the two-year period, and the assumptions detailed in the Materials and Methods. Given these assumptions, the number of draft ewes moved from hill to upland/lowland farms was estimated at 42,046 sheep per annum. These movements would be expected to take place in autumn in variably sized batches, similar to the movements of store lambs.

a)



b)

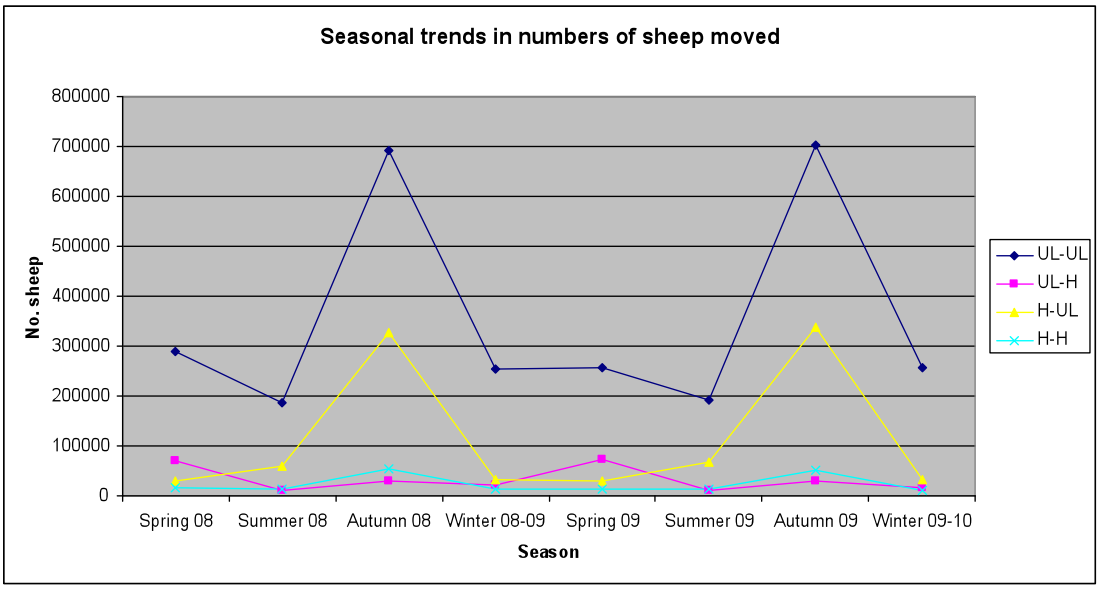


Figure 2.3 - Graphs showing trends in a) numbers of movements and b) numbers of sheep moved by season for each of the between-farm movement classes.
H = hill farm
UL = upland/lowland farm

2. Descriptive investigation of SAMS movement data

For movements recorded within the SAMS data, the departure and arrival holdings for each movement were classified as hill farm, upland/lowland farm, or "other" premises (these latter consisting of either premises that were not farms or that were not coded in the original Hill Codings data). Using this, movements were then classified into nine different types according to type of departure and arrival premises. Figure 2.2 shows the relative proportions of sheep movements of different types recorded in SAMS for the two years

between 01/03/2008 – 28/02/2010. It can be seen that the majority of these consist of movements from an upland/lowland flock to either another upland/lowland flock (41%) or to an "other" premises (33%). Following these, hill to upland/lowland movements are the most common (11%).

Seasonal movement trends: Figure 2.3 a) and b) shows seasonal trends in the numbers of movements and the number of sheep moved respectively, by season and in each of the farm-to-farm movement categories identified in the SAMS data. They show consistent autumnal increases in almost all movement classes, demonstrating the seasonal nature of

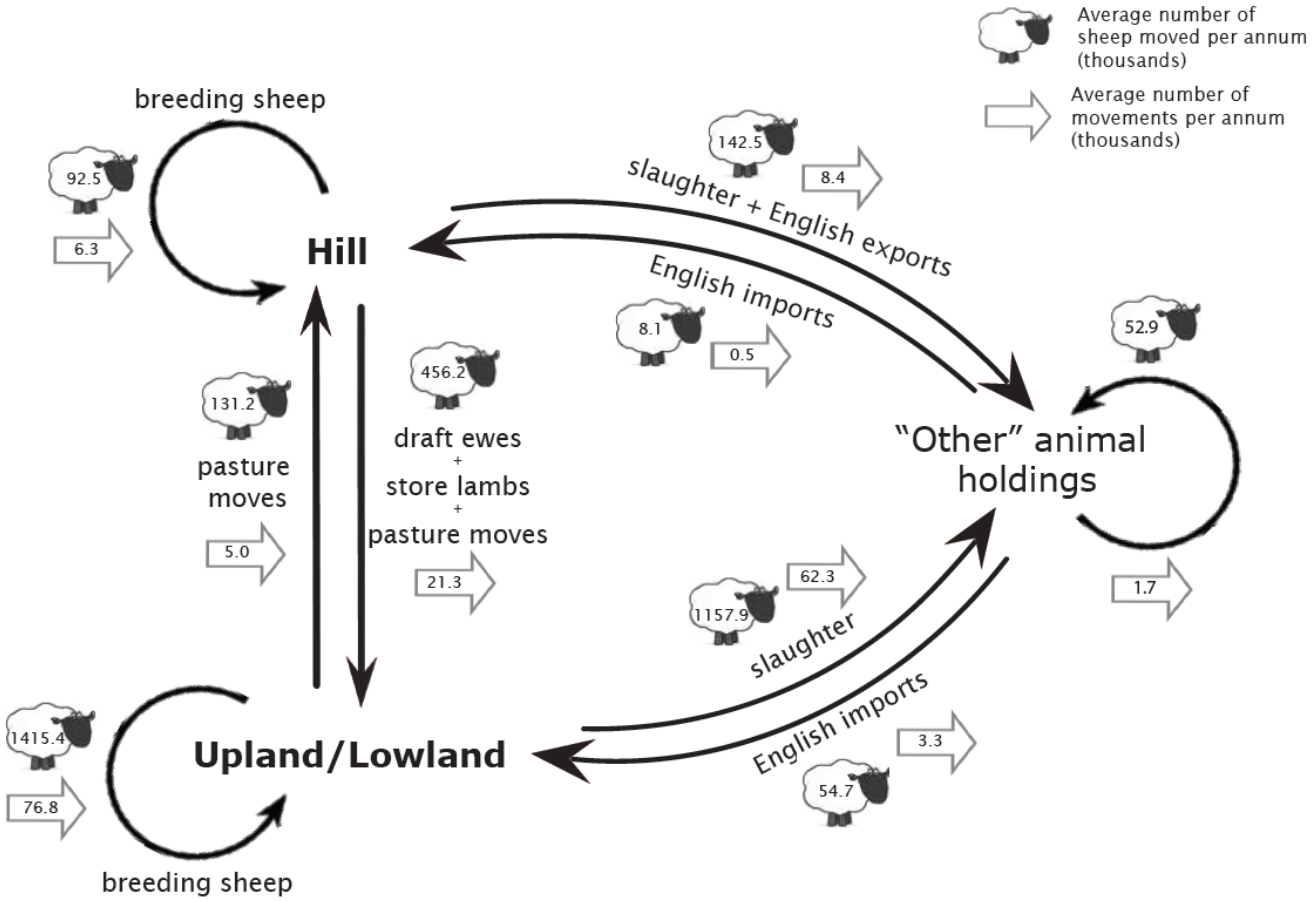


Figure 2.4 - Diagram showing the average numbers of movements per annum and the numbers of sheep moved per annum (in thousands) between hill flocks, upland/lowland flocks and other animal holdings and the potential reasons for the different types of movements.

Table 2.3 - Table giving details of the type of animal holding sheep depart from and arrive at for each movement classification, giving the numbers of movements and numbers of sheep moved to and from each type of premise as totals for the whole two-year period

Mvmt Class	Departure Premise	No. Mvmts	No. Sheep	Arrival Premise	No. Mvmts	No. Sheep
UL-UL	Animal Residence	153611	2830872	Animal Residence	153611	2830872
UL-H	Animal Residence	10075	262321	Animal Residence	10075	262321
H-UL	Animal Residence	42512	912349	Animal Residence	42512	912349
H-H	Animal Residence	12687	184889	Animal Residence	12687	184889
UL-other	Animal Residence	124609	2315897	Animal Residence	431	6377
				Gathering	9453	106603
				Non-AMLS	480	17623
				Research Centre	6	33
				Slaughter Premises	114238	2185256
				Veterinary	1	5
H-other	Animal Residence	16724	284918	Animal Residence	158	2466
				Gathering	3274	52428
				Non-AMLS	401	7302
				Port	1	39
				Research Centre	2	4
				Slaughter Premises	12887	222677
				Veterinary	1	2
other-UL	Animal Residence	477	8072	Animal Residence	6609	109446
	Gathering	5722	86054			
	Non-AMLS	401	15277			
	Slaughter Premises	9	43			
other-H	Animal Residence	48	720	Animal Residence	1015	16102
	Gathering	866	10819			
	Non-AMLS	97	4503			
	Slaughter Premises	3	59			
	Veterinary	1	1			
other-other	Animal Residence	254	2494	Animal Residence	65	408
	Gathering	2176	71673	Gathering	306	4737
	Non-AMLS	291	4743	Non-AMLS	51	829
	Slaughter Premises	617	26912	Slaughter Premises	2916	99848

an industry still closely linked to the seasonality of the ovine reproductive cycle. The exception to these autumnal peaks is seen in the numbers of sheep moved from upland/lowland farms to the hills, which peaks in the spring rather than autumn. This increase is likely to be caused by "pasture moves" where hill ewes in their first year spend the winter in the lowlands and are then moved back up to the hills in the spring. These pasture moves were identified in the SAMS data and consisted of 528 linked pairs of hill and upland/lowland farms moving an average of 148,623 sheep between them per annum.

Movements between different premises: Figure 2.4 summarises the average numbers of movements and the average numbers of sheep moved per annum between the three different categories of animal holdings (hill, upland/lowland and "other"), along with potential reasons for those movements.

In addition to this, Table 2.3 gives more detail about the "other" premises involved in sheep movements, and the numbers of movements and of sheep moved by premises type over the two-year period. It also shows that premises coded as hill or upland/lowland consist of animal residences (i.e. farms) only, thereby confirming that the modification to the original Hill Codings data was successful.

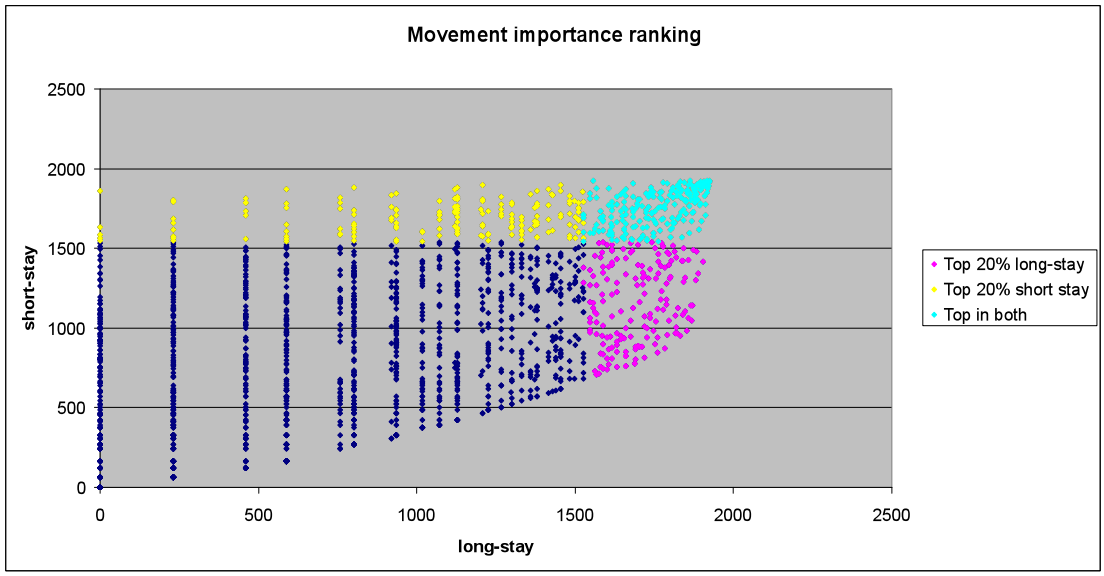
It can be seen in Table 2.3 that for movements departing from an "other" premises, the majority (over 80%) of these "other" premises were listed as Gatherings (meaning primarily markets). Within Scotland, the CPH from which the sheep originated is listed as the departure holding, irrespective of any market that the sheep were sold through. This is not the case in the rest of Britain, and so these movements originating from Gathering (market) premises are likely to represent imports via markets from outside of Scotland.

Looking at movements arriving into an "other" premises, over 90% of the "other" arrival premises are slaughterhouses. This is consistent with the overview of the sheep industry given in the introduction, where the major output from the industry is meat. Interestingly, a lower proportion (77%) of hill-to-"other" moves represent slaughter moves compared with the overall proportion for farm-to-other moves. This again is consistent with knowledge of the industry, as the hills produce not only lambs for slaughter but also ewes for crossbreeding, and some of the latter will be sold out of Scotland through markets and therefore be registered as hill-to-Gathering moves.

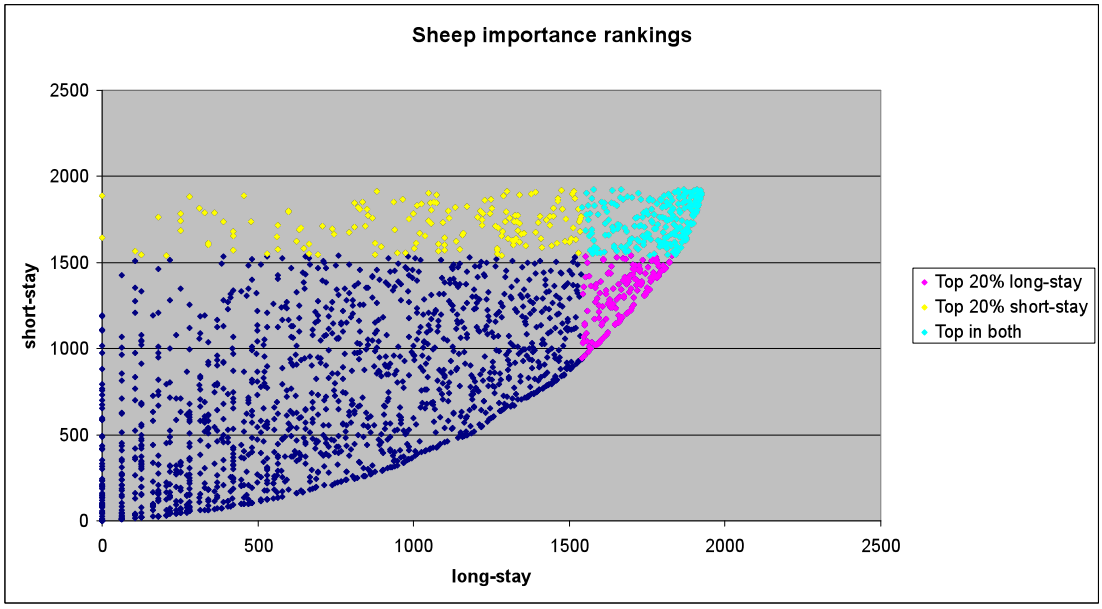
3. Values for farm network importance

The importance of a farm in the long stay network was calculated as number of long-stay movements (or sheep) coming in to the farm multiplied by the numbers of long-stay sheep or movements going out. This same calculation was used to estimate the importance of a farm within the short-stay network with the addition a value related to whether the movement had taken place through a market.

These importance values within the long-stay and short-stay networks were then ranked and compared for each farm. This comparison is shown in Figure 2.5 for importance in



a)



b)

Figure 2.5 - Graphs comparing the rank importance of farms within the long-stay network (x-axis) and short-stay network (y-axis) for numbers of movements and numbers of sheep moved over the whole two-year period (graphs a) and b) respectively). Farms present in the top 20% of the long-stay importance rankings, the short-stay importance rankings, and both, are highlighted.

terms of movements and of sheep for the whole two year period (2.5 a) and b) respectively). The graphs of the same comparisons for the autumn and spring of 2008 (Figure 2.6 a)-d)) are also given to illustrate that a similar pattern, although with less data points, is seen when the data are broken down by season. The top 20% ranked farms in either and in both networks are highlighted. The interesting shape of the graphs (whereby the long-stay importance is never larger than the short-stay importance) is accounted for by the fact that the long-stay network is an integral part of the short-stay network. It can be seen that there is a degree of positive correlation between the short-stay and long-stay importance rankings.

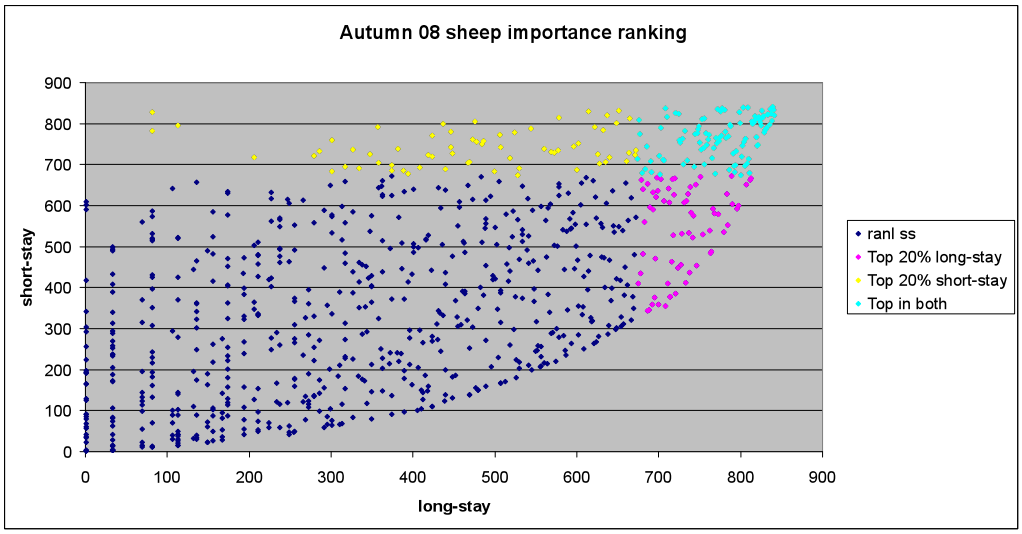
Table 2.4 gives the proportion of farms that are present in the top 20% importance rankings for *both* the long-stay and the short-stay networks, for movements over the whole two-year time period and by individual season, along with the number of farms present in both networks for each time period. It can be seen that the autumn networks are much more densely populated than the other seasons, again demonstrating that the majority of sheep movements take place in autumn. The number of farms present in the 20% most important farms for both networks is consistently above the percentage that would be expected if no correlation existed between short-stay and long-stay importance values (4%).

DISCUSSION

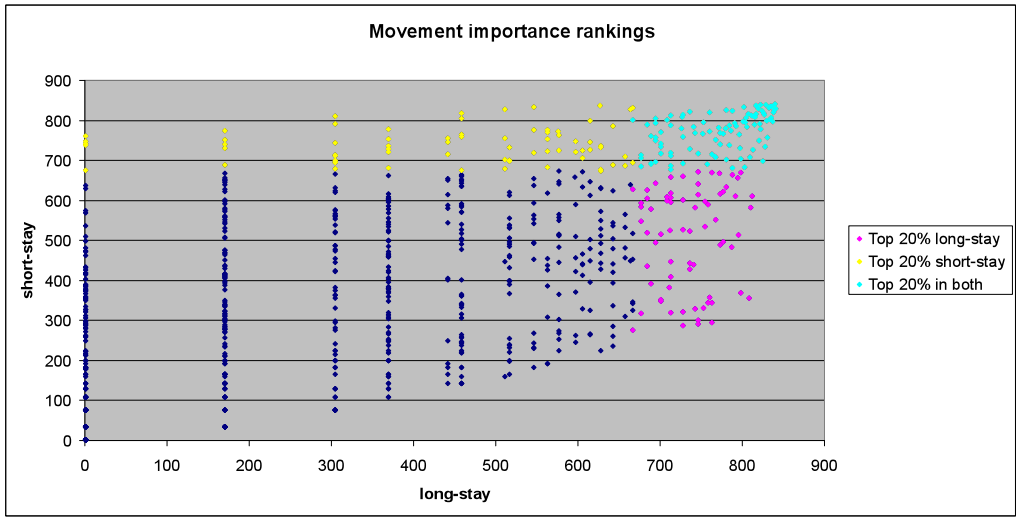
Section 1 of this project involved an investigation into the structure of the British sheep industry, with particular reference to the types and reasons for movements between different sectors. This review was conducted in order to verify that the aims of the project and the assumptions made were appropriate to the system under study. A summary of the types of sheep moving between different sectors is given in Table 2.2 and also shown in a slightly different manner in Figure 2.4. Two particular areas were studied in depth to assess the project aims and assumptions:

Hill vs. not-hill categories: Previous work classified the type of flock associated with each British sheep farm, dividing them into hill and not-hill categories, and provided the Hill Codings data used in this project. This hill/not-hill division may disregard potentially important differences between upland and lowland flocks.

a)



b)



c)

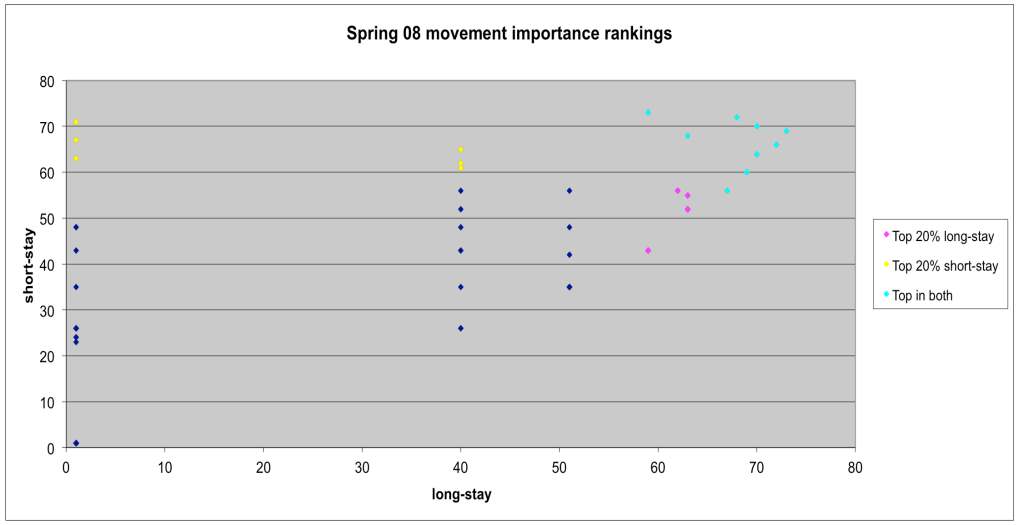


Figure 2.6 (continued overleaf) - Graphs comparing the rank importance of farms within the long-stay network (x-axis) and short-stay network (y-axis) for numbers of movements and numbers of sheep moved for autumn 2008 (graphs a) and b) respectively) and for spring 2008 (graphs c) and d)). Farms present in the top 20% of the long-stay importance rankings, the short-stay importance rankings, and both, are highlighted.

d)

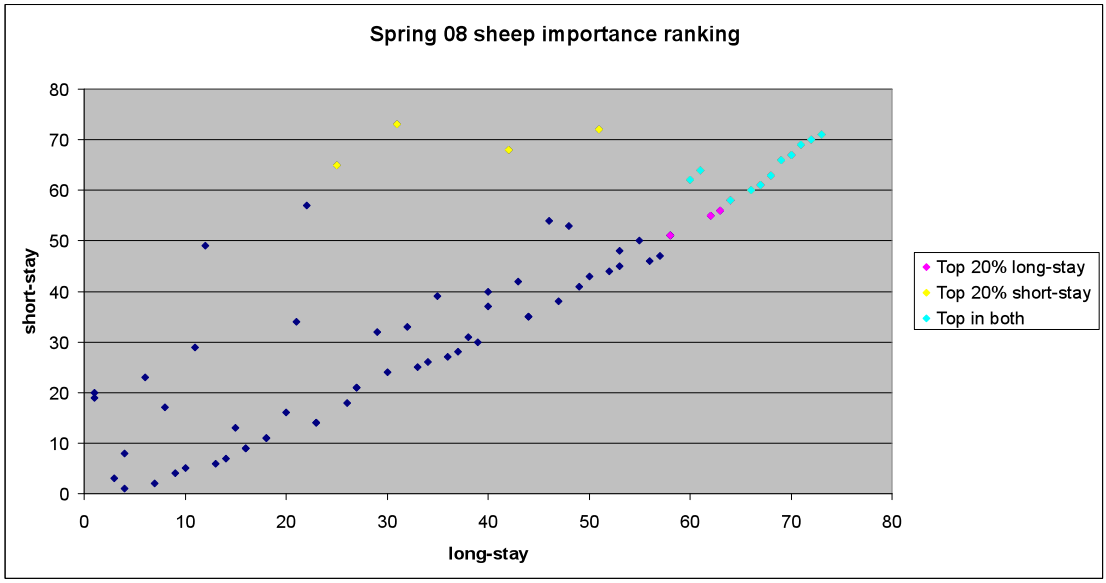


Figure 2.6 (*cont.*) - Graphs comparing the rank importance of farms within the long-stay network (x-axis) and short-stay network (y-axis) for numbers of movements and numbers of sheep moved spring 2008 (graphs c) and d)). Farms present in the top 20% of the long-stay importance rankings, the short-stay importance rankings, and both, are highlighted.

Therefore it appears that hill flocks do function in a separate manner than upland or lowland flocks. However, using the hill/not-hill classification to distinguish between long- and short-stay movements would potentially result in the misclassification of some upland-to-lowland moves as long-stay movements (as it would be assumed that all upland/lowland-to-upland/lowland movements were within-sector breeding moves resulting in long-stays, which is not the case).

Draft ewes: Draft ewes are older hill ewes that are moved to upland/lowland flocks. They go on to produce one or more crops of lambs at their farm of destination and up to half of these may then be sold on as breeding stock themselves. Therefore movements of draft ewes can be seen to contribute to the long-stay network, and are a potential mechanism for spreading disease between sectors from the hills into upland/lowland flocks.

The numbers of sheep moved as draft ewes are not insignificant, and we estimated them to consist of around 42,000 sheep per annum. However, they are very difficult to isolate from other sheep moving from hill to upland/lowland farms using the SAMS movement data.

Table 2.4 - Table showing the number of farms participating in both the short-stay and the long-stay networks for the entire two-year time period and for each of the seasons involved. Also indicated are the number and percentage of farms for each time period that were present in the 20% most important farms for both short-stay and long-stay networks, in terms of movements and of number of sheep moved

Time period	No. farms present in both networks	No. farms in top 20% most important in both networks	
		Movement importance (% of total in brackets)	Sheep importance (% of total in brackets)
Entire two years	1925	214 (11%)	228 (12%)
Spring 2008	73	9 (12%)	11 (15%)
Summer 08	84	7 (8%)	14 (16%)
Autumn 08	841	98 (12%)	102 (12%)
Winter 08-09	147	20 (14%)	16 (11%)
Spring 09	58	3 (5%)	10 (11%)
Summer 09	75	5 (7%)	11 (15%)
Autumn 09	843	93 (11%)	98 (12%)
Winter 09-10	138	19 (14%)	19 (14%)

Although it is possible to distinguish them from autumn pasture moves, the draft ewe movements have a very similar signature to movements of store lambs. These would not be expected to contribute to long-stay moves, despite moving at a similar time of year to the draft ewes, in variable batch sizes. Broadly speaking, we would expect all hill flocks to contribute to draft ewe movements, and all upland/lowland flocks to potentially act as recipients (though depending on the degree of separation between pedigree and commercial upland/lowland flocks, the latter assumption may not hold true). Therefore we would expect the long stay network to include a random subset of all hill to upland/lowland movements.

Both of the above issues were solved in this project by restricting the analysis to hill flocks only. Contact networks in the hill sector act as drivers for the flow of sheep within the industry, as shown in Figure 2.1. Therefore hill flocks are also potentially drivers for the spread of disease through the industry, as demonstrated by the role of hill sheep in the 2001 British foot and mouth outbreak (Gibbens et al., 2001). An infection may pass into the hill sector via pasture moves of sheep or through routes of transmission other than sheep movements. By restricting the analysis to the hill flocks, this project still gives a valuable insight into the industry, while avoiding the problems described above in applying the available data to a real-life system. .

A descriptive investigation was carried out in section 2 of the project to further characterise the movements occurring within the industry, the results of which are summarised in Figures 2.4 and Table 2.3. These results correspond with the understanding

of the sheep industry gained in the initial review, and provided a basis for the construction of network importance values in section 3.

In Section 3 of this project, two contact networks for the hill sector of the Scottish sheep industry were identified: the short-stay network, consisting of all sheep movements, along which a fast transmitting disease such as foot and mouth disease could spread; and the long-stay network, consisting of breeding moves only, along which a slowly transmitting disease such as scrapie might transmit. The importance of a node within a directional network is calculated as the number of in connections multiplied by the number of out connections, and with here the addition of a weighting in the short-stay network to correct for the additional importance of movements passing through a market.

In calculating importance values for farms within the long- and short-stay networks and comparing the ranked values for each of them a positive correlation between the two values was identified. Although this correlation might seem self-evident (i.e. it would seem obvious that a farm which participates in many short-stay moves would also be involved in a lot of long-stay movements) this is not the case. For example in the British poultry and pig industries it would be expected that the highly specialised breeding population would participate in almost entirely long-stay movements, while the end-stage commercial units would be involved in short-stay movements only, thus potentially giving a scenario where minimal correlation exists between short- and long-stay importance. The 20% of farms important to either the long-stay or the short-stay network were also identified, as was the subset of farms present in the top 20% of *both* networks (shown in Figure 2.5), and this latter value is consistently above that which would be expected if there were no correlation between importance values in the different networks.

Given the known heterogeneity between farms in terms of sheep movements (Kiss et al., 2006) and the fact that 20% of farms make up over 80% of the potential for disease transmission (Volkova et al., 2010) it is likely that the most important farms identified in the networks here are also disproportionately important in the spread of disease through the networks. The next stage of this research will be to identify, using disease simulations on both networks discussed here, whether removal of the subset of farms that are present in the top 20% most important farms in both networks acts to reduce the size of a disease outbreak to a significant extent. If this were the case, it would validate the possibility that targeting this single group of farms enables the holistic control of multiple infectious diseases within the national flock.

Chapter 3

The implications of post-infection immunity for the epidemiology and control of *Escherichia coli* O157 infection of cattle

Supervised by Dr Louise Matthews
(louise.matthews@glasgow.ac.uk)

Chapter 3 – THE IMPLICATIONS OF POST-INFECTION IMMUNITY FOR THE EPIDEMIOLOGY AND CONTROL *Escherichia coli* O157 INFECTION OF CATTLE

INTRODUCTION

Escherichia coli O157 causes severe bloody diarrhoea in humans and disease can in some cases progress to haemolytic uraemic syndrome, a potentially fatal complication (Tarr et al., 2005). The severity of human disease seen with *E. coli* O157 is largely due to the production of shiga toxin (Ethelberg et al., 2004). Cattle are the maintenance host of *E. coli* O157 and infection in this species is asymptomatic; human outbreaks are generally caused either by faecal contamination of food or water, or by direct contact with cattle (Pennington, 2010). Methods to reduce carriage and shedding of *E. coli* in cattle should therefore act to reduce the risk of human infection, and for this purpose a vaccine to reduce shedding of *E. coli* O157 in cattle is currently licensed in Canada (Bioniche Food Safety, 2011).

Scotland has one of the world's highest incidences of human O157 infection (Locking et al., 2011), and as a result the epidemiology of the pathogen has been very well studied in this country. An extensive series of studies combining epidemiological modelling with field prevalence surveys has done much to clarify the dynamics of *E. coli* O157 in Scottish cattle (Matthews et al., 2009, Matthews et al., 2006a, Liu et al., 2007a, Liu et al., 2007b, Chase-Topping et al., 2007, Matthews et al., 2006b, Pearce et al., 2009). These studies have highlighted the importance of between-individual heterogeneity and "supershedder" animals in maintaining *E. coli* O157 in cattle (Chase-Topping et al., 2008). This finding has also been validated biologically, with *E. coli* O157 colonisation at the recto-anal junction linked to very high levels of bacterial shedding (Omisakin et al., 2003, Low et al., 2005, Cobbold et al., 2007, Naylor et al., 2003). The presence of supershedders and individual heterogeneity has important consequences for attempts to control the pathogen.

All the previous Scottish modelling work has been carried out under the assumption of an immediate return to susceptibility once an animal has recovered from infection. However, recent experimental work has questioned this assumption, suggesting that there may be a period of immunity to re-infection following recovery, rather than immediate return to the susceptible state (Naylor et al., 2007, Hoffman et al., 2006). This period of post-infection

immunity has been suggested to last a similar amount of time as the infectious period, approximately 2-4 weeks (D. Gally, personal communication).

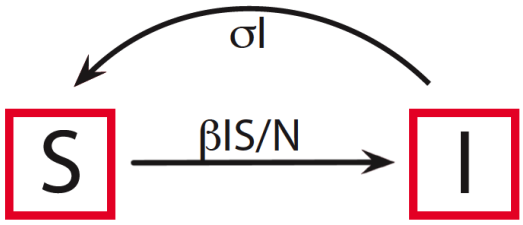


Figure 3.1 – Diagram representing the SIS model structure. S = susceptible compartment, I = infectious compartment

In addition to the O157 work described above, O'Reilly et al. (2010) and Liu et al. (2007b) have used a combination of modelling and field data to investigate the transmission dynamics of non-O175 shigatoxigenic strains of *E. coli* in cattle. These studies again use a model assuming immediate return to susceptibility, and though less is known about the recovery period in non-O157 *E. coli* strains, post-infection immunity is a possibility here as well. Although this project primarily concerns the dynamics of *E. coli* O157, the results described here can also help to understand the impact of post-infection immunity on the dynamics of non-O157 shigatoxigenic *E. coli*. The importance of understanding non-O157 strains has been highlighted recently by the German outbreak of shigatoxigenic *E. coli* O104:H4, which affected thousands of people and killed 47 (European Centre for Disease Prevention and Control, 2011).

The aim of this project was to clarify the impact that a period of post-infection immunity might have on the conclusions drawn by previous work assuming an immediate return to susceptibility. This was achieved by comparing the outputs of epidemiological models based on the two different assumptions of transmission dynamics (i.e. models with and without post-infection immunity). A stochastic Susceptible-Infectious-Susceptible (SIS) model (Figure 3.1) was used to represent the model structure in previous work assuming immediate return to susceptibility, and this was here compared with a stochastic Susceptible-Infectious-Recovered/Immune-Susceptible (SIRS) model (Figure 3.2) to allow investigation of the effects of a period of post-infection immunity.

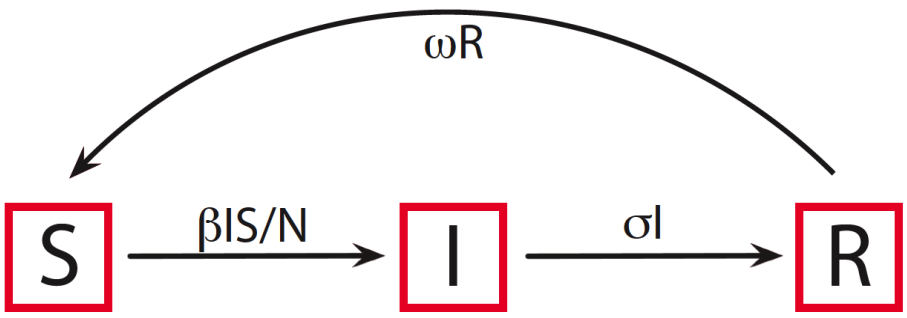


Figure 3.2 – Diagram representing the SIRS model structure. S = susceptible compartment, I = infectious compartment, R = recovered/immune compartment

MATERIALS AND METHODS

Models

This project was carried out in the R programming environment (R Development Core Team, 2011). Infection dynamics involving a period of post-infection immunity were simulated using the SIRS model outlined in Figure 3.2. In this model, infection, recovery and return-to-susceptibility events arise in the population according to the rates given in Table 3.1. The SIS model, outlined in Figure 3.1 and representing the previous work assuming immediate return to susceptibility was used as a comparison for the SIRS model. The rates of infection and recovery events in the SIS model are as shown in Table 3.1, however in this model there is no Recovered/Immune compartment and therefore a recovery event is the same as return-to-susceptibility, occurring at rate σI.

As cattle populations tend to be managed in relatively small group sizes, chance effects are expected to play a substantial role in their infection dynamics. The Gillespie algorithm was used (Gillespie, 1977) to incorporate this stochasticity.

Table 3.1 – Rates of occurrence of different kinds of event in the SIRS model. S = number of susceptible individuals at a particular time point, I = number of infectious individuals at a particular time point, R = number of recovered/immune individuals at a particular time point. β, σ and ω represent rate constants

Event type	Rate of occurrence
Infection	$\beta IS/N$
Recovery	σI
Return-to-susceptibility	ωR

Input parameters

Throughout these simulations, the recovery parameter σ was kept constant at 0.1.

Simulations iterating through a range of values of the transmission parameter β (0.05 - 0.7) were run for the above models. β is related to the basic reproduction number R_0 by the formula $R_o=\beta/\sigma$. As σ was kept constant throughout these simulations; changing β is equivalent to iterating through a range of R_0 values.

These iterations through β were carried out for a series of values of ω, the rate constant for return-to-susceptibility, as well as for the SIS model. As ω is inversely proportional to the time an individual spends in the recovered/immune compartment, increasing ω means reducing the period of post-infection immunity (the SIS model, with no recovered/immune

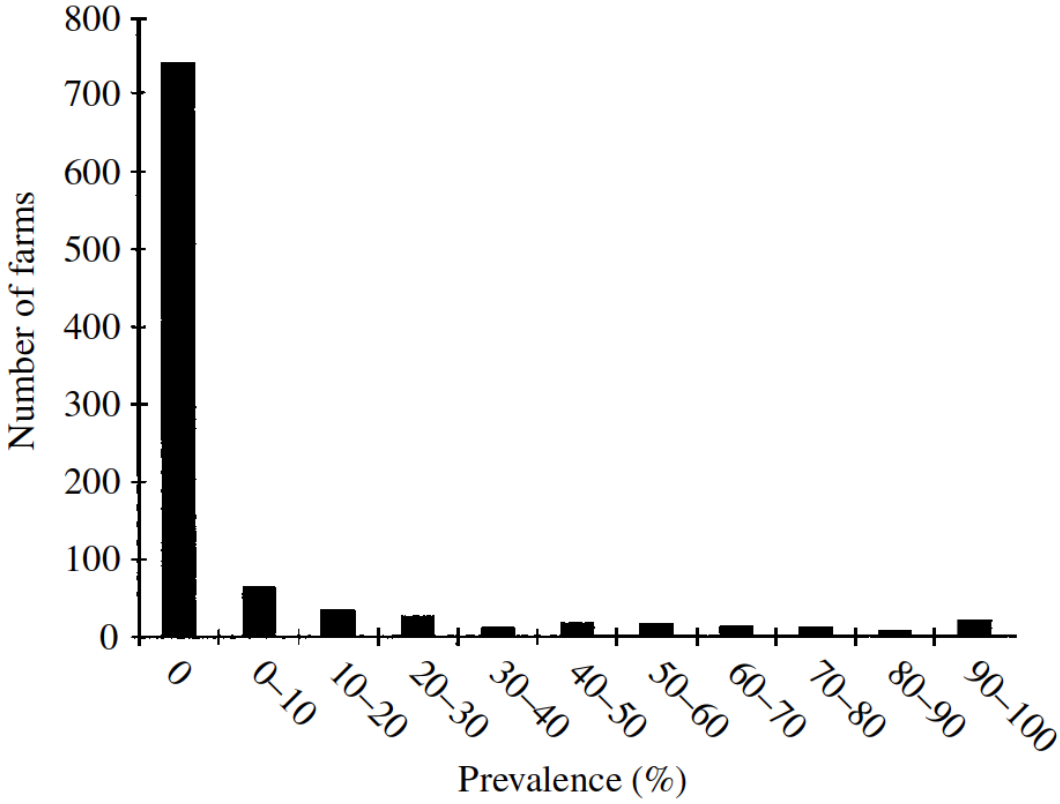


Figure 3.3 - The distribution of prevalences of *E. coli* O157 in faecal pats sampled from finishing groups of beef cattle on 952 Scottish cattle farms (Matthews et al., 2006b)

compartment, essentially has an infinite ω rate). Experimental data suggest the potential period of immunity in cattle following infection with *E. coli* O157 is approximately the same length as the infectious period. This indicates that biologically plausible values of ω are in the region of 0.1, equal to our default value for the recovery rate constant σ.

Outputs

For each combination of β and ω, 500 stochastic simulations were run. The mean duration of infection, mean infection prevalence and the variance in infection prevalence were calculated for each combination. These factors were selected as they capture key features of the epidemiology as follows:

- Mean prevalence gives information on the expected mean on-farm prevalence, which provides information on R_0
- Variance in prevalence provides information on the expected range of infection incidence on different farms (and therefore information on the degree of individual heterogeneity in transmission)

- Average duration is related to the number of farms that would be expected to be free of infection at any one time and therefore provides information on the rate of incursion of infection onto farms.

These types of output have been used extensively in previous work fitting models to field data, such as the prevalence distribution shown in Figure 3.3 (taken from Matthews et al. (2006b). In particular, the skewed distribution of prevalences (which is related to the variance in prevalence) was used in the Scottish modelling work to demonstrate the importance of individual heterogeneity and supershedders in the dynamics of *E. coli* O157 transmission in cattle.

RESULTS

Figure 3.4 shows the duration of infection and mean and variance in prevalence over a biologically plausible range of the basic reproduction number R_0, for a series of different values of the rate of return-to-susceptibility ω and for the SIS model. R_0 is here directly proportional to the transmission parameter β.

Higher values of R_o (reflecting increased β) give an increase in mean duration of infection and mean and variance in prevalence. The increases in duration and mean prevalence are to be expected from an increase in the transmission rate, but the increase in variance would have been less straightforward to predict without the use of stochastic simulation.

Figure 3.4 also shows that higher values of ω, the return-to-susceptibility rate, likewise give an increase in the three calculated output statistics. This occurs because increasing ω reduces the amount of time individuals spend in the recovered/immune compartment. Therefore at higher ω values a larger proportion of the population is susceptible to infection at any one time. The highest values for duration and prevalence are seen for the SIS model, which, without a recovered/immune compartment, effectively has infinite ω rate.

Figure 3.4c), showing variance in prevalence against R_0, demonstrates that the variance is lower for the SIRS model than the SIS model. Field data from extensive *E. coli* O157 studies exhibit considerable variance in on-farm prevalence, and it was the inability

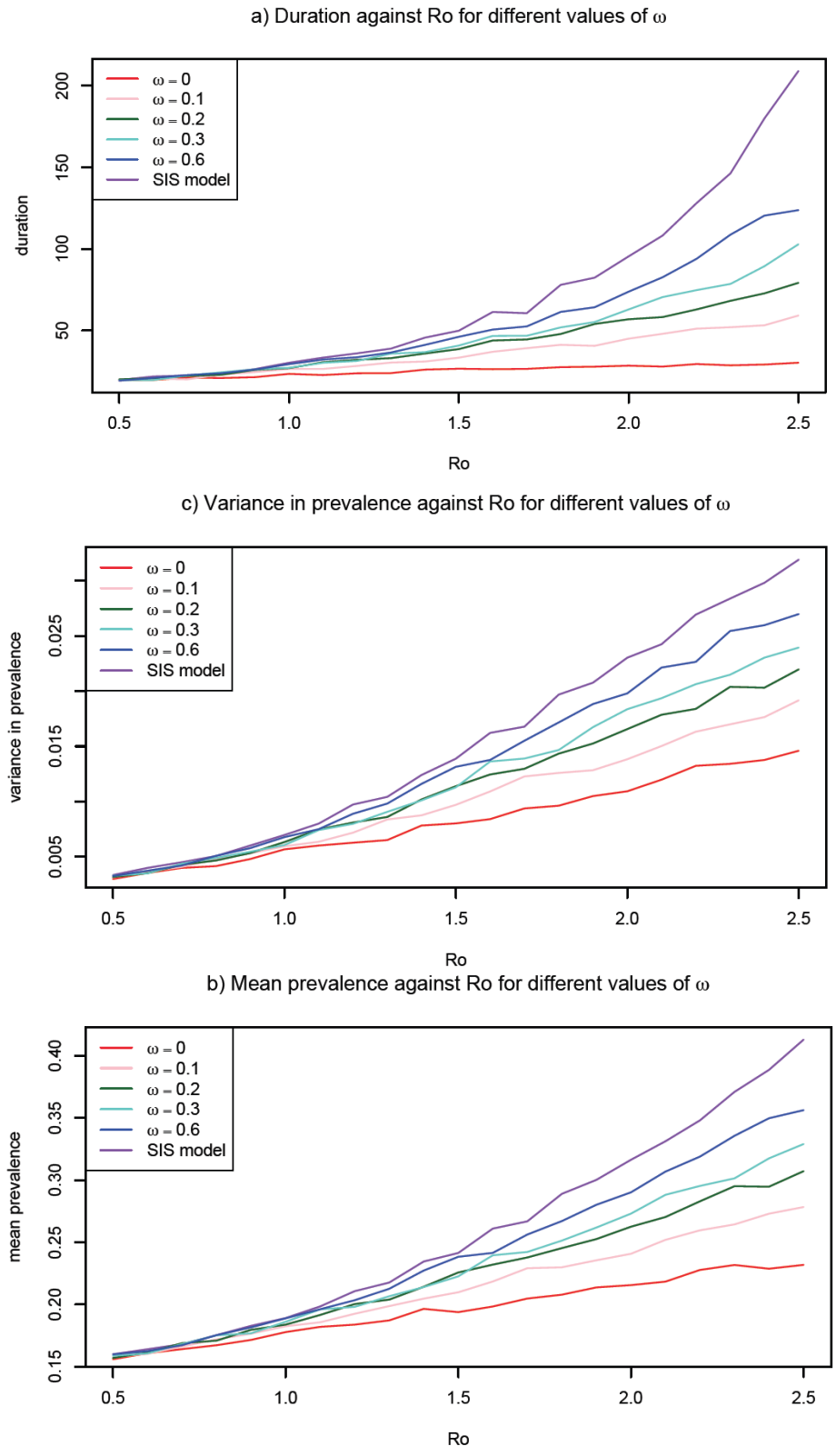


Figure 3.4 – Graphs to show the effect that varying R_0 has on a) duration of infection, b) mean infection prevalence c) variance in infection prevalence, given different values of ω in the SIRS model, and for the SIS model (population size = 10, σ = 0.1, averaged over 500 iterations of the stochastic model)

of models assuming a homogeneous population to explain this variance led to the characterisation of individual heterogeneity in the transmission of *E. coli* O157.

It can be seen in Figures 3.4 and 3.5 that for higher values of R_0, the outputs calculated for different values of ω are increasingly divergent. Therefore at low transmission rates and low R_0, the impact of post-infection immunity is likely to be less than at higher values of R_0 and higher transmissibility. Figure 3.5 also illustrates that in higher prevalence situations one would expect more divergent values of R_0 from the SIS model compared with the SIRS model.

The herd immunity is the proportion of the population that must be successfully vaccinated to prevent a from pathogen spreading. The threshold value at which herd immunity occurs is directly related the R_0 value by the following formula (Fine, 1993):

$$\text{Herd immunity threshold} = 1 - 1/R_0$$

Using these values of R_0 from the model simulations it was possible to calculate the values for herd immunity threshold at different prevalences for the SIS and SIRS (ω=0.1) model. These are shown in Table 3.2.

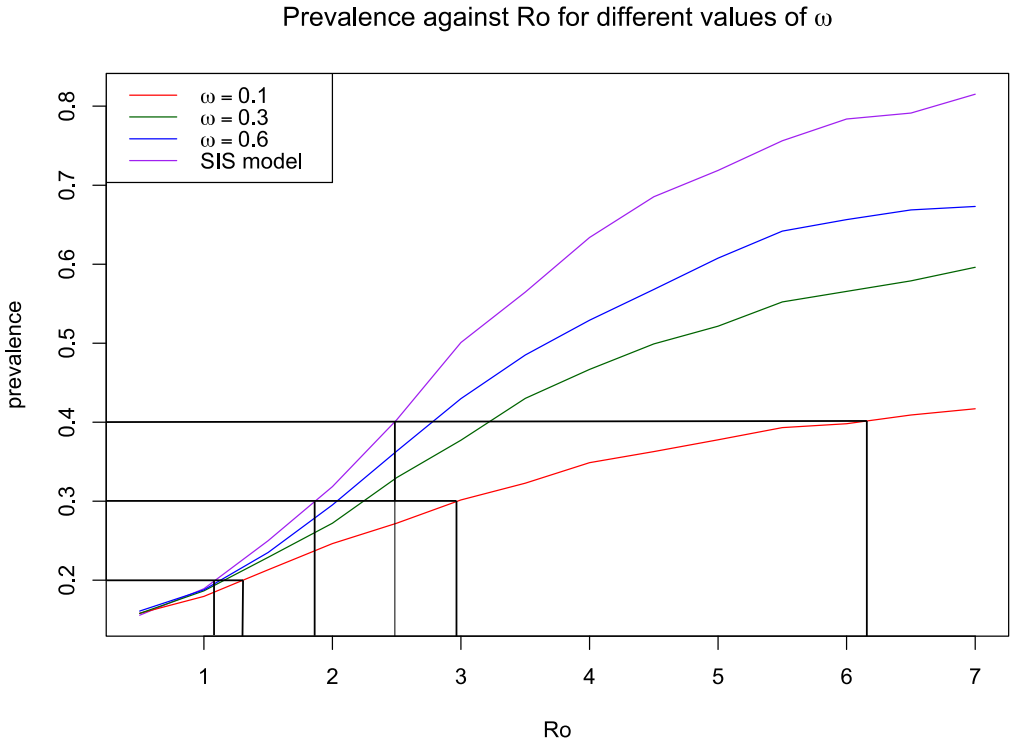


Figure 3.5 - Graph to illustrate the divergence in R_0 for ω = 0.1 and the SIS model at different mean prevalences

Table 3.2 - Herd immunity threshold at different average infection prevalences under the SIRS (ω = 0.1) model and the SIS model

Prevalence (%)	Herd immunity threshold (%)	
	SIRS model (ω=0.1)	SIS model
20	21	13
25	53	35
30	66	47
40	83	60

DISCUSSION

This work exploring the dynamics of *E. coli* infection in cattle has shown that for larger values of R_0, the basic reproduction number, and larger values of ω, the rate of return-to susceptibility, there is an increase in our three measured outputs, namely the mean duration of an outbreak and the mean and variance in prevalence of infection. This is due to an increase in the transmissibility of the disease with increased R_0, and a decrease in the length of post-infection immunity (and therefore increase in the proportion of the population susceptible to infection) with increased ω.

Relevance of post-infection immunity to control measures

The basic reproduction number R_o is defined as the average number of new infections resulting when one infectious individual is introduced into a totally susceptible population (Anderson and May, 1991). If R_0 is less than 1 then the infection will on average die out, if it is above 1 it is expected to spread through the population. The R_0 value of a pathogen is important in understanding its transmission dynamics and can be related to many facets of infection and control. An important example is herd immunity, the proportion of the population that must be successfully vaccinated to prevent the pathogen spreading.

The results of this study demonstrate that in higher prevalence situations, post-infection immunity is increasingly important in the infection dynamics of *E. coli*, resulting in a prediction of higher R_0 values for a given observed prevalence as shown in Figure 3.5. Table 3.2 shows that at higher infection prevalences, the differences in herd immunity threshold values for models with (SIRS) and without (SIS) a period of post-infection immunity are considerable. The current vaccine, to prevent shedding of *E. coli* O157 in cattle is at most 60-70% effective (Snedeker et al., 2011). The herd immunity thresholds calculated for the two different models indicate that on farms with *E. coli* O157 prevalences of 30-40%, the presence or absence of a period of post-infection immunity could mean the difference between success and failure of vaccination in elimination the infection. In the scenario involving post-infection immunity (the SIRS model) the herd immunity threshold is increased above the level of vaccine efficacy at these levels of infection. In contrast, the results from the SIS model suggest if cattle become susceptible immediately after recovering from infection, vaccination could be useful even at these higher prevalence levels.

Post-infection immunity and the role of supershedders

The calculation of mean duration, mean prevalence and variance in prevalence reflect the type of information that has been used to fit models to field data in previous studies

modelling the transmission dynamics of *E. coli* O157 in cattle (Matthews et al., 2009, Matthews et al., 2006a, Liu et al., 2007a, Liu et al., 2007b, Chase-Topping et al., 2007, Matthews et al., 2006b, Pearce et al., 2009). Although fitting to data is beyond the scope of this project, the results of our analyses provide important insights into the robustness of the conclusions drawn by previous modelling studies. In particular, our results concerning variance in prevalence are key. Variance in prevalence has previously been used to show that the very wide variability in the level of *E. coli* O157 present on different farms is best explained by models which allow some individuals within the population to have much higher levels of transmission than others (so-called supershedders): it was found that models assuming a homogeneous population with no individual variability were unable to reproduce the level of variation seen in the field (Matthews et al., 2009, Matthews et al., 2006a, Liu et al., 2007a, Liu et al., 2007b, Chase-Topping et al., 2007, Matthews et al., 2006b, Pearce et al., 2009)

The results shown in Figure 3.4c) indicate that models which do not allow for individual variation (implying a homogeneous population) such as those used here, give a lower variance in prevalence when they incorporate a period of post-infection immunity compared to immediate return to susceptibility. Therefore the SIRS model, with delayed return to susceptibility, would explain even less of the observed variation in the field data than does the SIS model used in the above work on supershedders. This suggests that a period of post-infection immunity may actually increase the degree of individual heterogeneity required to explain the on-farm dynamics of *E. coli* O157 in cattle, strengthening the case for supershedders.

Further work

This project has made the first steps towards understanding the importance of post-infection immunity in the transmission dynamics of *E. coli* in cattle. However it is based on the assumption of a homogeneous population, where all individuals contribute equally to transmission. Previous work has shown this is not the case in the field, and that heterogeneity between individuals is important in the maintenance and spread of the infection (Chase-Topping et al., 2008). The next step in investigating the impact that post-infection immunity has on *E. coli* infection would be to create and compare individual-based SIS and SIRS models, allowing the incorporation of individual variation into the analysis.

CONCLUSIONS

SPECIFIC CONCLUSIONS:

Chapter 1 - Examining the evolutionary history of Bovine Papillomavirus in equine sarcoids

Chapter 1 describes phylogenetic analyses of three genetic regions of BPV-1. Although two of these analyses were uninformative, the phylogeny of the LCR sequences suggests three interesting conclusions with respect to the evolutionary history of the virus, in its natural host the cow and regarding the association with equine sarcoids. The LCR phylogeny shows that the genetic diversity seen in BPV-1 isolates associated with equine sarcoids is ancient and predates bovine and equine domestication. The phylogeny also shows a clear separation between the sequences found in Africa/Brazil and those found in Europe, with the European variants showing greater evolutionary divergence from the root of the tree. Finally, the distribution of cattle and horse samples within the LCR phylogeny, in combination with experimental understanding of the viral biology, suggest that BPV-1 originally diversified within its bovine host followed by multiple, more recent crossover events into equids. The analysis also highlights the high prevalence within the equine samples of a potentially equine-adapted sequence variant.

Chapter 2 - Scottish sheep movements and their potential for disease transmission

Chapter 2 gives an overview of the structure of the British sheep industry and a descriptive summary of the data associated with sheep movements in Scotland. Using concepts drawn from network theory, values for the combined in and out movements of sheep farms in Scotland were calculated. These indicate that many farms which are important in the transmission of highly infectious epidemic diseases, such as FMD, are also likely to be important for the transmission of less infectious chronic diseases such as scrapie. These results therefore suggest it may be possible to improve the health of the national flock by targeting interventions towards a limited subset of flocks, thereby efficiently reducing the transmission of multiple infectious agents

Chapter 3 - The impact of post-infection immunity on *Escherichia coli* O157 infection in cattle

Chapter 3 uses mathematical modelling to show that a period of post-infection immunity may have a significant impact on the dynamics of *E. coli* O157 in cattle, when compared with the assumption of an immediate return to susceptibility. This has important implications for control measures and suggests that elimination of infection will be

especially difficult on higher prevalence farms. The model incorporating a period of post-infection immunity also predicts reduction in the variance in prevalence of infection. This suggests that accounting for a period of post-infection immunity will increase the predicted role of individual heterogeneity and supershedder animals in explaining the observed variability in the distribution of *E. coli* O157 between farms. To confirm these findings, future work is required to examine the behaviour of individual-based models incorporating both the effects of individual heterogeneity and a period of post-infection immunity.

OVERALL SUMMARY

These three projects have provided brief insights into how numerical methods and quantitative analyses can be used to inform and to draw conclusions from all three aspects of the study of infectious disease: laboratory science (Chapter 1); observational and field data (Chapter 2); and mathematical modelling (Chapter 3). The techniques used here have varied greatly in scale: from the molecular genetic (Chapter 1) right through to the national level (Chapter 2), and in perspective: from the pathogen (Chapter 1) to the population (Chapter 2). While Chapter 3 is intermediate in both scale and perspective, it differs in its use of mathematical modelling which allows us to capture processes underlying the observed data, thereby enabling extrapolation and prediction outwith the data.

As the capability of computers increases and new analytical tools are developed, quantitative methodologies are increasingly able to give a deeper insight and add power to all three aspects of the study of infectious disease, as has been briefly shown here. Most excitingly perhaps, the crossover between the three areas and the integration of the analytical tools and data from all three disciplines holds much promise. This could give a fusion of our understanding across the different scales and across the different perspectives touched on here, potentially providing even greater rewards in the future.

REFERENCES

AJMONE-MARSAN, P., GARCIA, J. F., LENSTRA, J. A. & GLOBALDIV, C. 2010. On the Origin of Cattle: How Aurochs Became Cattle and Colonized the World. *Evolutionary Anthropology,* 19**,** 148-157.

ANDERSON, R. M. & MAY, R. M. 1991. *Infectious diseases of humans : dynamics and control*, Oxford University Press.

ANON 2007. HPV and Cervical Cancer in the World 2007 Report - Preface. *Vaccine,* 25.

ARNOLD, M., MEEK, C., WEBB, C. R. & HOINVILLE, L. J. 2002. Assessing the efficacy of a ram-genotyping programme to reduce susceptibility to scrapie in Great Britain. *Preventive veterinary medicine,* 56**,** 227-249.

BIONICHE FOOD SAFETY 2011. *Escherichia coli* O157 bacterial extract vaccine - Technical Information. http://www.econichevaccine.com/docs/techinfo.pdf.

BODEN, L. A., HOUSTON, F., FRYER, H. R. & KAO, R. R. 2010. Use of a preclinical test in the control of classical scrapie. *The Journal of general virology,* 91**,** 2642-50.

BORZACCHIELLO, G. & ROPERTO, F. 2008. Bovine papillomaviruses, papillomas and cancer in cattle. *Veterinary Research,* 39**:**45.

BROWN, J. M., HEDTKE, S. M., LEMMON, A. R. & LEMMON, E. M. 2010. When trees grow too long: investigating the causes of highly inaccurate bayesian branch-length estimates. *Systematic biology,* 59**,** 145-61.

CHAMBERS, G., ELLSMORE, V. A., O'BRIEN, P. M., REID, S. W. J., LOVE, S., CAMPO, M. S. & NASIR, L. 2003. Association of bovine papillomavirus with the equine sarcoid. *Journal of General Virology,* 84**,** 1055-1062.

CHASE-TOPPING, M., GALLY, D., LOW, C., MATTHEWS, L. & WOOLHOUSE, M. 2008. Super-shedding and the link between human infection and livestock carriage of Escherichia coli O157. *Nature reviews. Microbiology,* 6**,** 904-12.

CHASE-TOPPING, M. E., MCKENDRICK, I. J., PEARCE, M. C., MACDONALD, P., MATTHEWS, L., HALLIDAY, J., ALLISON, L., FENLON, D., LOW, J. C., GUNN, G. & WOOLHOUSE, M. E. 2007. Risk factors for the presence of high-level shedders of Escherichia coli O157 on Scottish farms. *Journal of clinical microbiology,* 45**,** 1594-603.

COBBOLD, R. N., HANCOCK, D. D., RICE, D. H., BERG, J., STILBORN, R., HOVDE, C. J. & BESSER, T. E. 2007. Rectoanal junction colonization of feedlot cattle by Escherichia coli O157:H7 and its association with supershedders and excretion dynamics. *Applied and environmental microbiology,* 73**,** 1563-8.

DRUMMOND, A. J., ASHTON, B., BUXTON, S., CHEUNG, M., COOPER, A., HELED, J., KEARSE, M., MOIR, R., STONES-HAVAS, S., STURROCK, S., THIERER, T. & WILSON, A. 2010. Geneious v5.1. *Available from* http://www.geneious.com.

DRUMMOND, A. J., HO, S. Y. W., PHILLIPS, M. J. & RAMBAUT, A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biology,* 4, 699-710.

DRUMMOND, A. J. & RAMBAUT, A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *Bmc Evolutionary Biology,* 7:214.

DRUMMOND, A. J., RAMBAUT, A., SHAPIRO, B. & PYBUS, O. G. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution,* 22**,** 1185-92.

ETHELBERG, S., OLSEN, K. E. P., SCHEUTZ, F., JENSEN, C., SCHIELLERUP, P., ENGBERG, J., PETERSEN, A. M., OLESEN, B., GERNER-SCHMIDT, P. & MOLBAK, K. 2004. Virulence factors for hemolytic uremic syndrome, Denmark. *Emerging Infectios Diseases,* 10**,** 842-7.

EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL 2011. Shiga toxin-producing E. coli (STEC): Update on outbreak in the EU (27 July 2011, 11:00). http://ecdc.europa.eu/en/activities/sciadvice.

FEVRE, E. M., BRONSVOORT, B. M., HAMILTON, K. A. & CLEAVELAND, S. 2006. Animal movements and the spread of infectious diseases. *Trends in microbiology,* 14**,** 125-31.

FINE, P. E. M. 1993. Herd Immunity: History, Theory, Practice. *Epidemiologic Reviews,* 15**,** 265-302.

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. *Veterinary Record,* 149**,** 729-43.

GILLESPIE, D. T. 1977. Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry,* 81**,** 2340-61.

GOBEIL, P., GAULT, E. A., CAMPO, M. S., GOW, J., MORGAN, I. M. & NASIR, L. 2007. Equine sarcoids are not induced by an infectious cell line. *Equine Veterinary Journal,* 39**,** 189-91.

GOTTSCHLING, M., STAMATAKIS, A., NINDL, I., STOCKFLETH, E., ALONSO, A. & BRAVO, I. G. 2007. Multiple evolutionary mechanisms drive papillomavirus diversification. *Molecular Biology and Evolution,* 24**,** 1242-58.

GREEN, D. M., KISS, I. Z. & KAO, R. R. 2006. Modelling the initial spread of foot-and-mouth disease through animal movements. *Proceedings. Biological sciences / The Royal Society,* 273**,** 2729-35.

GUINDON, S. & GASCUEL, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology,* 52**,** 696-704.

HASEGAWA, M., KISHINO, H. & YANO, T. A. 1985. DATING OF THE HUMAN APE SPLITTING BY A MOLECULAR CLOCK OF MITOCHONDRIAL-DNA. *Journal of Molecular Evolution,* 22**,** 160-74.

HO, S. Y. W. & LARSON, G. 2006. Molecular clocks: when times are a-changin'. *Trends in Genetics,* 22**,** 79-83.

HOFFMAN, M. A., MENGE, C., CASEY, T. A., LAEGREID, W., BOSWORTH, B. T. & DEAN-NYSTROM, E. A. 2006. Bovine immune response to shiga-toxigenic Escherichia coli O157:H7. *Clinical and vaccine immunology : CVI,* 13**,** 1322-7.

HTTP://WWW.LIV.AC.UK/SARCOIDS/. 2010. *University of Liverpool sarcoid website* [Online]. University of Liverpool. Available: http://www.liv.ac.uk/sarcoids/ [Accessed 10/01/2011 2011].

HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R. & BOLLBACK, J. P. 2001. Evolution - Bayesian inference of phylogeny and its impact on evolutionary biology. *Science,* 294**,** 2310-4.

JACKSON, C. 1936. The incidence and pathology of tumours in domesticated animals in South Africa. *Onderspoort Journal of Veterinary Science,* 6**,** 375-85.

KAO, R. R., DANON, L., GREEN, D. M. & KISS, I. Z. 2006. Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proceedings. Biological sciences / The Royal Society,* 273**,** 1999-2007.

KAO, R. R., GREEN, D. M., JOHNSON, J. & KISS, I. Z. 2007. Disease dynamics over very different time-scales: foot-and-mouth disease and scrapie on the network of livestock movements in the UK. *Journal of the Royal Society, Interface / the Royal Society,* 4**,** 907-16.

KIMURA, B., MARSHALL, F. B., CHEN, S. Y., ROSENBOM, S., MOEHLMAN, P. D., TUROSS, N., SABIN, R. C., PETERS, J., BARICH, B., YOHANNES, H., KEBEDE, F., TECLAI, R., BEJA-PEREIRA, A. & MULLIGAN, C. J. 2011. Ancient DNA from Nubian and Somali wild ass provides insights into donkey ancestry and domestication. *Proceedings of the Royal Society B-Biological Sciences,* 278**,** 50-57.

KIMURA, M. 1980. A SIMPLE METHOD FOR ESTIMATING EVOLUTIONARY RATES OF BASE SUBSTITUTIONS THROUGH COMPARATIVE STUDIES OF NUCLEOTIDE-SEQUENCES. *Journal of Molecular Evolution,* 16**,** 111-20.

KISS, I. Z., GREEN, D. M. & KAO, R. R. 2006. The network of sheep movements within Great Britain: Network properties and their implications for infectious disease spread. *Journal of the Royal Society, Interface / the Royal Society,* 3**,** 669-77.

KNOTTENBELT, D. C. 2005. A suggested clinical classification for the equine sarcoid. *Clinical Techniques in Equine Practice,* 4**,** 278-95.

LANCASTER, W. D. & OLSON, C. 1982. ANIMAL PAPILLOMAVIRUSES. *Microbiological Reviews,* 46**,** 191-207.

LIU, W. C., MATTHEWS, L., CHASE-TOPPING, M., SAVILL, N. J., SHAW, D. J. & WOOLHOUSE, M. E. 2007a. Metapopulation dynamics of Escherichia coli O157 in cattle: an exploratory model. *Journal of the Royal Society, Interface / the Royal Society,* 4**,** 917-24.

LIU, W. C., SHAW, D. J., MATTHEWS, L., HOYLE, D. V., PEARCE, M. C., YATES, C. M., LOW, J. C., AMYES, S. G., GUNN, G. J. & WOOLHOUSE, M. E. 2007b. Modelling the epidemiology and transmission of Verocytotoxin-producing Escherichia coli serogroups O26 and O103 in two different calf cohorts. *Epidemiology and infection,* 135**,** 1316-23.

LOCKING, M. E., POLLOCK, G. J., ALLISON, L. J., RAE, L., HANSON, M. F. & COWDEN, J. M. 2011. *Eschericia coil* O157 infection and secondary spread, Scotland, 1999-2008. *Emerging Infectios Diseases,* 17**,** 524-7.

LOW, J. C., MCKENDRICK, I. J., MCKECHNIE, C., FENLON, D., NAYLOR, S. W., CURRIE, C., SMITH, D. G., ALLISON, L. & GALLY, D. L. 2005. Rectal carriage of enterohemorrhagic Escherichia coli O157 in slaughtered cattle. *Applied and environmental microbiology,* 71**,** 93-7.

MATTHEWS, L., LOW, J. C., GALLY, D. L., PEARCE, M. C., MELLOR, D. J., HEESTERBEEK, J. A., CHASE-TOPPING, M., NAYLOR, S. W., SHAW, D. J., REID, S. W., GUNN, G. J. & WOOLHOUSE, M. E. 2006a. Heterogeneous shedding of Escherichia coli O157 in cattle and its implications for control. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 547-52.

MATTHEWS, L., MCKENDRICK, I. J., TERNENT, H., GUNN, G. J., SYNGE, B. & WOOLHOUSE, M. E. 2006b. Super-shedding cattle and the transmission dynamics of Escherichia coli O157. *Epidemiology and infection,* 134**,** 131-42.

MATTHEWS, L., REEVE, R., WOOLHOUSE, M. E., CHASE-TOPPING, M., MELLOR, D. J., PEARCE, M. C., ALLISON, L. J., GUNN, G. J., LOW, J. C. & REID, S. W. 2009. Exploiting strain diversity to expose transmission heterogeneities and predict the impact of targeting supershedding. *Epidemics,* 1**,** 221-9.

NASIR, L. & CAMPO, M. S. 2008. Bovine papillomaviruses: their role in the aetiology of cutaneous tumours of bovids and equids. *Veterinary Dermatology,* 19**,** 243-254.

NASIR, L., GAULT, E., MORGAN, I. M., CHAMBERS, G., ELLSMORE, V. & CAMPO, M. S. 2007. Identification and functional analysis of sequence variants in the long control region and the E2 open reading frame of bovine papillomavirus type 1 isolated from equine sarcoids. *Virology,* 364**,** 355-61.

NASIR, L. & REID, S. W. J. 1999. Bovine papillomaviral gene expression in equine sarcoid tumours. *Virus Research,* 61**,** 171-175.

NAYLOR, S. W., FLOCKHART, A., NART, P., SMITH, D. G., HUNTLEY, J., GALLY, D. L. & LOW, J. C. 2007. Shedding of Escherichia coli O157:H7 in calves is reduced by prior colonization with the homologous strain. *Applied and environmental microbiology,* 73**,** 3765-7.

NAYLOR, S. W., LOW, J. C., BESSER, T. E., MAHAJAN, A., GUNN, G. J., PEARCE, M. C., MCKENDRICK, I. J., SMITH, D. G. E. & GALLY, D. L. 2003. Lymphoid Follicle-Dense Mucosa at the Terminal Rectum Is the Principal Site of Colonization of Enterohemorrhagic Escherichia coli O157:H7 in the Bovine Host. *Infection and immunity,* 71**,** 1505-1512.

O'REILLY, K. M., LOW, J. C., DENWOOD, M. J., GALLY, D. L., EVANS, J., GUNN, G. J., MELLOR, D. J., REID, S. W. & MATTHEWS, L. 2010. Associations between the presence of virulence determinants and the epidemiology and ecology of zoonotic Escherichia coli. *Applied and environmental microbiology,* 76**,** 8110-6.

OLSON, C. & COOK, R. H. 1951. CUTANEOUS SARCOMA-LIKE LESIONS OF THE HORSE CAUSED BY THE AGENT OF BOVINE PAPILLOMA. *Proceedings of the Society for Experimental Biology and Medicine,* 77**,** 281-284.

OMISAKIN, F., MACRAE, M., OGDEN, I. D. & STRACHAN, N. J. C. 2003. Concentration and Prevalence of Escherichia coli O157 in Cattle Feces at Slaughter. *Applied and environmental microbiology,* 69**,** 2444-2447.

PEARCE, M. C., CHASE-TOPPING, M. E., MCKENDRICK, I. J., MELLOR, D. J., LOCKING, M. E., ALLISON, L., TERNENT, H. E., MATTHEWS, L., KNIGHT, H. I., SMITH, A. W., SYNGE, B. A., REILLY, W., LOW, J. C., REID, S. W., GUNN, G. J. & WOOLHOUSE, M. E. 2009. Temporal and spatial patterns of bovine Escherichia coli O157 prevalence and comparison of temporal changes in the patterns of phage types associated with bovine shedding and human E. coli O157 cases in Scotland between 1998-2000 and 2002-2004. *BMC microbiology,* 9**,** 276.

PENNINGTON, H. 2010. Escherichia coli O157. *The Lancet,* 376**,** 1428-1435.

POLLOTT, G. E. & STONE, D. G. 2003. The breeding structure of the British sheep industry 2003. *In:* DEFRA (ed.).

POSADA, D. 2003. Using MODELTEST and PAUP* to select a model of nucleotide substitution. *Curr Protoc Bioinformatics,* Chapter 6**,** Unit 6.5.

POSADA, D. 2008. jModelTest: phylogenetic model averaging. *Molecular biology and evolution,* 25**,** 1253-6.

R DEVELOPMENT CORE TEAM 2011. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

RAGLAND, W. L., MCLAUGHLIN, C. A. & SPENCER, G. R. 1970. Attempts to relate bovine papilloma virus to the cause of equine sarcoid: horses, donkeys and calves inoculated with equine sarcoid extracts. *Equine Veterinary Journal,* 2**,** 168-172.

RECTOR, A., LEMEY, P., TACHEZY, R., MOSTMANS, S., GHIM, S.-J., VAN DOORSLAER, K., ROELKE, M., BUSH, M., MONTALI, R. J., JOSLIN, J., BURK, R. D., JENSON, A. B., SUNDBERG, J. P., SHAPIRO, B. & VAN RANST, M. 2007. Ancient papillomavirus-host co-speciation in Felidae. *Genome Biology,* 8**,** R57.1-R57.12.

RONQUIST, F. & HUELSENBECK, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics,* 19**,** 1572-1574.

SHAH, S. D., DOORBAR, J. & GOLDSTEIN, R. A. 2010. Analysis of Host-Parasite Incongruence in Papillomavirus Evolution Using Importance Sampling. *Molecular Biology and Evolution,* 27**,** 1301-1314.

SNEDEKER, K. G., CAMPBELL, M. & SARGEANT, J. M. 2011. A Systematic Review of Vaccinations to Reduce the Shedding of Escherichia coli O157 in the Faeces of Domestic Ruminants. *Zoonoses and public health*.

TACHEZY, R., DUSON, G., RECTOR, A., JENSON, A. B., SUNDBERG, J. P. & VAN RANST, M. 2002. Cloning and genomic characterization of Felis domesticus papillomavirus type 1. *Virology,* 301**,** 313-321.

TAJIMA, F. 1989. STATISTICAL-METHOD FOR TESTING THE NEUTRAL MUTATION HYPOTHESIS BY DNA POLYMORPHISM. *Genetics,* 123**,** 585-595.

TAMURA, K., DUDLEY, J., NEI, M. & KUMAR, S. 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution,* 24**,** 1596-1599.

TARR, P. I., GORDON, C. A. & CHANDLER, W. L. 2005. Shiga-toxin-producing Escherichia coli and haemolytic uraemic syndrome. *The Lancet,* 365**,** 1073-1086.

VILA, C., LEONARD, J. A., GOTHERSTROM, A., MARKLUND, S., SANDBERG, K., LIDEN, K., WAYNE, R. K. & ELLEGREN, H. 2001. Widespread origins of domestic horse lineages. *Science,* 291**,** 474-7.

VOLKOVA, V. V., HOWEY, R., SAVILL, N. J. & WOOLHOUSE, M. E. J. 2010. Sheep Movement Networks and the Transmission of Infectious Diseases. *PLoS One,* 5**,** e11185.

VOSS, J. L. 1969. TRANSMISSION OF EQUINE SARCOID. *American Journal of Veterinary Research,* 30**,** 183-&.

WEBB, C. R. 2005. Farm animal networks: unraveling the contact structure of the British sheep population. *Preventive veterinary medicine,* 68**,** 3-17.

WEBB, C. R. 2006. Investigating the potential spread of infectious diseases of sheep via agricultural shows in Great Britain. *Epidemiology and infection,* 134**,** 31-40.

WIRTH, T., MEYER, A. & ACHTMAN, M. 2005. Deciphering host migrations and origins by means of their microbes. *Molecular ecology,* 14**,** 3289-306.