Using Whole Genome Sequencing to investigate the inter-species transmission dynamics of *Mycobacterium bovis*

Joseph Crispell

Bachelor of Science in Zoology

Submitted in fulfilment of the requirements for the Degree of:
Doctor of Philosophy

School of Life Sciences
College of Medical, Veterinary, and Life Sciences
University of Glasgow

April 25, 2017

Abstract

Bovine tuberculosis (bTB), a disease of cattle caused by *Mycobacterium bovis*, presents considerable health and economic burdens in many countries. In the United Kingdom and New Zealand, wildlife reservoirs are implicated in the spread and persistence of bTB in cattle populations. Where multi-host systems exist, understanding the roles of different host species in the spread and persistence of *M. bovis* infection in livestock is paramount. This thesis describes how Whole Genome Sequenced (WGS) *M. bovis* data can be used to investigate inter-species transmission of bTB between livestock and wildlife populations.

WGS data must be processed before it can be used in downstream analyses. A filter sensitivity analysis was used to investigate whether the selection of quality criteria to be used in the processing of WGS data, could be informed by epidemiological data describing the sampled bTB system. WGS *M. bovis* data were available from three different bTB systems: in Northern Ireland, England and New Zealand. By using agreement between the genetic data, resulting from applying different quality filters, and the epidemiological data as an indication of the appropriateness of filtering, it was demonstrated that epidemiological data could inform the selection of quality criteria.

With appropriately processed WGS data from infected cattle and wildlife in New Zealand, the role of wildlife reservoirs was investigated. In addition, the WGS data were compared to the types defined by a different molecular typing method used heavily in New Zealand until 2012. The high resolution WGS data agreed well with the previous typing method, and was used to demonstrate that inter-species transmission had occurred between the sampled cattle and wildlife populations.

The large spatial range of the data available from New Zealand made it difficult to determine the direction of inter-species transmission. In the southwest of England, Woodchester Park is home to a badger population that is naturally infected with *M. bovis* and living in close proximity to many cattle herds. WGS *M. bovis* data, for infected cattle and badgers sampled from this system, combined with detailed epidemiological data, were used to provide evidence of inter-species transmission in both directions. The sampled badger population was also shown to be acting as a maintenance reservoir for bTB.

This thesis describes how WGS data can be used to investigate inter-species transmission, but also highlights how the underlying bTB systems must be sampled appropriately for these data to be most informative. Future work will involve addressing the limitations of the available data and the analyses conducted, as well as analysis of new data and implementation of new methods.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to take this opportunity to thank Rowland Kao and Ruth Zadoks, my supervisors, for their continued support throughout my PhD. The last four years have been extremely rewarding and have put me in a great position for my future. I would like to acknowledge all the help that I have received from all the members of Rowland's group. This group is packed with enthusiastic, intelligent researchers and I have benefited a great deal from being a part of it. Thank you to those working at Garscube for making a welcoming and enjoyable environment to work in.

I am very grateful to the BBSRC for funding my studentship. All the analyses described are based upon the data resulting from a lot of other people's hard work; Marian Price-Carter at AgResearch in New Zealand, Dez Delahay and crew at Woodchester Park, Robin Skuce and co. at AFBNI, Simon Harris at the Sanger Institute, everyone with Noel Smith at APHA and Graham Hamilton at the University of Glasgow Polyomics facility.

Thank you to my parents, brothers and sisters, who have provided me with a great deal of love and support over the course of my PhD.

Lastly I would to thank Joanna, my beautiful wife. Thank you for all your care, love and support. These last four years have been the best of my life so far, and I have you to thank!

Author's Declaration

I, Joseph Crispell, declare that, with exception to where explicit reference has been made to the contribution of others, this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow, or any other institution.

A version of Chapter 3 was submitted to the *BMC Genomics* journal in September 2016 and returned with minor revisions. The revisions were addressed and the manuscript was re-submitted on the 2nd of December 2016.

Table 1: List of abbreviations used in the thesis.

Abbreviation	Definition
bTB	Bovine tuberculosis
TB	Tuberculosis
OTF	Officially Tuberculosis-Free
REA	Restriction Endonuclease Analysis
VNTR	Variable Number Tandem Repeat
NGS	Next Generation Sequencing
WGS	Whole Genome Sequencing
PCR	Polymerase Chain Reaction
CFU	Colony Forming Unit
CMI	Cell Mediated Immune
SICCT	Single Intradermal Comparative Cervical Tuberculin test
CFT	Caudal Fold Single Intradermal Tuberculin test
LCS	Latent Class Statistical
ELISA	Enzyme-Linked Immunosorbent Assay
MI	Multiplex Immunoassay
VRA	Vector Risk Areas
VFA	Vector Free Areas
γ-IFN	γ-interferon
STAA	Special Testing Areas - Annual
STAB	Special Testing Areas - Biennial
SA	Surveillance Areas
FMD	Foot-and-Mouth Disease
RFLP	Restriction Fragment Length Polymorphism
Spoligotyping	Spacer-oligotyping
DNA	Deoxyribonucleic Acid
UK	United Kingdom
US	United States
SNP	Single Nucleotide Polymorphism
PhD	Doctor of Philosophy
RBCT	Randomised Badger Culling Trial
TVR	Test, Vaccinate or Remove
DPP	Dual-Path Platform VetTB
BEAST	Bayesian Evolutionary Analysis by Sampling Trees
AFBNI	Agri-Food and Biosciences Institute in Northern Ireland
LJ	Löwenstein-Jensen medium
CTAB	Salt hexadecyl Trimethyl Ammonium Bromide
FERA	Food and Environment Research Agency
PE	Proline-Glutamate
PPE	Proline-Proline-Glutamate
VCF	Variant Calling Format
DP	Read Depth
HQDP	High Quality Base Depth
MQ	Mapping Quality
SUP	Allele Support
COV	Site Coverage across isolates
PROX	Site Proximity
OOB	Out Of Bag
RF	Random Forest
MRCA	Most Recent Common Ancestor
HKY	Hasegawa-Kishino-Yano
GMRF	Gaussian Markov Random Field
DATM	Discrete Ancestral Trait Mapping
CTS	Cattle Tracing System
SAM	System for recording bTB testing results
QUAL	Quality score
SIR	Susceptible-Infected-Removed

Chapter 1

Introduction

1.1 Background

Bovine tuberculosis (bTB) is a globally important zoonotic disease, which is caused by *Mycobacterium bovis* bacteria. Domestic cattle are a principal host species, their infection has significant economic and public health impacts. Around the world, milk pasteurisation, test and slaughter regimes, movement controls, and abattoir surveillance limit the zoonotic impact of bTB, but at a considerable cost [1, 2, 3, 4, 5, 6, 7].

Combining test and slaughter, movement controls and abattoir surveillance to control *M. bovis* infection in cattle can be successful. In Europe, the "officially tuberculosis-free" (OTF) status is one that confers considerable advantages for trading [8]. A number of European countries are considered to be OTF, having <0.1% of cattle herds infected [4]. In 1997, after a 30 year campaign, Australia gained "TB Free" status, the requirements for which are similar to those in the EU [9]. More recently, Scotland gained the European OTF status despite bordering England, a source of endemic bTB [10].

The broad host range of *M. bovis* [11] means many wildlife species can act as infection reservoirs creating complex epidemiological systems and complicating control efforts [12, 13, 14, 15]. For example, in New Zealand unnaturally high brush-tailed possum (*Trichosurus vulpecula*) densities represent a large reservoir for *M. bovis*, which causes periodic outbreaks in livestock [15, 16]. Across Europe there are a number of wildlife species implicated in the transmission and persistence of bTB in livestock (Table 1.1). In the UK, it is thought that the European badger (*Meles meles*) is promoting persistence of bTB in cattle [11]. In contrast to the invasive and ecologically damaging brush-tailed possum of New Zealand, the European badger is a protected species - under the 1992 Badger Protection Act [17].

Given that wildlife populations are known to play a role in the persistence of cattle bTB,

Table 1.1: Some examples of wildlife reservoirs for *Mycobacterium bovis* present in European Countries. Data summarising review article by Gortázar *et al*. [18]

Country	Main Wild Reservoir(s)
United Kingdom	European Badger
Republic of Ireland	European Badger
Portugal	Red Deer, Fallow Deer, Wild Boar
France	Red Deer and Wild Boar
Italy	Wild Boar
Slovakia	Wild Boar

it is important to quantify the extent of this role. Unfortunately investigating inter-species transmission events is greatly limited by the poor sensitivity of available tests, which may result in infected animals being missed [19, 20]. In addition, upon exposure the development of symptoms can take many months, depending on the species infected, making it difficult to estimate the time of exposure for an infected animal [18].

Molecular techniques have been employed to investigate the spread of *M. bovis* infection within and between populations [21, 22, 23]. More recently, the traditional typing methods such as Restriction Endonuclease Analysis (REA), and Variable Number Tandem Repeat (VNTR) analysis, are being replaced by Whole Genome Sequencing (WGS). The continuing technological advances, culminating in the current Next Generation Sequencing (NGS) platforms, make WGS a feasible epidemiological tool. WGS won't necessarily provide a sequence for the entire *M. bovis* genome as some areas - such as the repeat regions that VNTR typing utilise - are more difficult to sequence and may have low coverage. That being said, the application of WGS to bTB systems have revealed transmission at an unprecedented scale [24, 25, 26].

WGS data combined with detailed epidemiological data were available for bTB systems in the United Kingdom and New Zealand. This thesis provides a description of the methods, techniques and analyses applied to these data. Initially a summary of *M. bovis* infection in cattle and wildlife populations, and its investigation using molecular methods is presented.

1.2 Bovine Tuberculosis in Cattle

1.2.1 Pathology

In cattle the primary route of infection is thought to be through inhalation, which results in an infection, initially, in the respiratory tract [27]. A very small dose is required for infection to develop, as demonstrated in an experimental study completed by Dean *et al*. [28]. One Colony Forming Unit (CFU), containing 6-10 viable bacilli was sufficient to establish infection. During the early stages of *M. bovis* infection in cattle, the Cell Mediated Immune (CMI) response dominates, it is only as the disease progresses that an antibody response develops [27]. A granuloma forms around the introduced bacilli and, in the majority of cases, the targeted CMI response is enough to inhibit any further disease progression - this early stage is often described as a latent infection (Figure 1.1, [29, 27]).

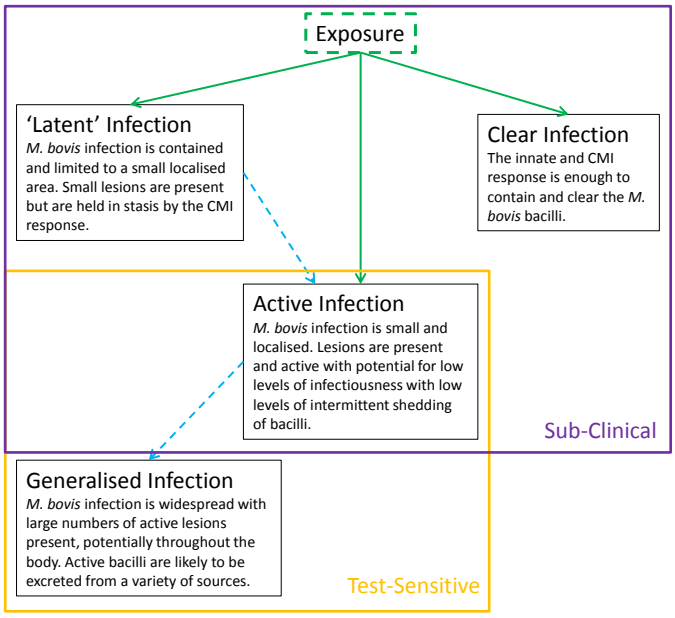

Figure 1.1: Summary of the possible infection and disease stages experienced upon exposure to *Mycobacterium bovis* and potential routes of disease progression. Whilst an animal is in a test-sensitive state it may be possible to detect infection, depending of the extent of shedding and sensitivity of the test(s) used.

Progression of the infection can lead to an infective stage, during which time the cow is likely to be intermittently shedding *M. bovis* bacilli [29]. Although rare, where a test and slaughter regime is in place, a generalised state of infection can develop. This generalised state is highly infectious, with high levels of shedding. The generalised state can further develop into an anergic state as a result of the immune system being compromised; animals in this state are likely to be insensitive to the currently available tests (Figure 1.2, [29]).

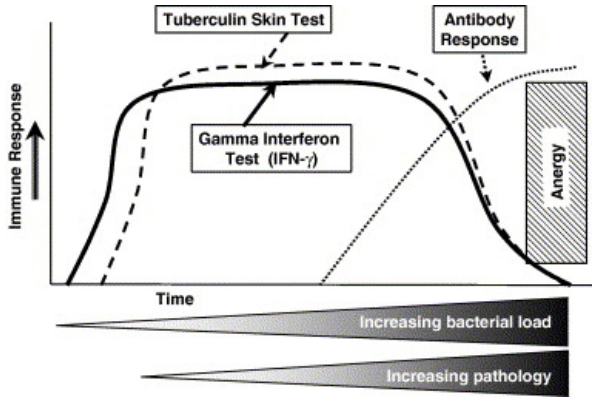

Figure 1.2: Schematic representation of the spectrum of responses exhibited by the bovine immune system to various tests for TB. This diagram illustrates the changing sensitivity of different tests, pathology, and bacterial load across different stages of infection. Figure taken directly from De la Rua-Domenech *et al*. [30].

In the majority of infected cattle, *M. bovis* infection is a slow, progressive, chronic infection, with little or no external signs of disease. Depending on the cow, infection can persist undetected for many years, as is seen in many human TB cases [31]. The low detectability of individuals suffering a 'latent' infection, means it is difficult to quantify the importance of the early stages of infection [32]. For those cattle with a more active and progressive infection, the infectious lifespan of the individual is likely to be dictated by the frequency of test and slaughter regimes.

1.2.2 Testing and Control

1.2.2.1 Tuberculin Skin Test

The test and slaughter regimes that have successfully controlled bTB in domestic cattle populations used a tuberculin skin test. Tuberculin consists of purified protein derivatives of the *M. bovis* pathogen [30]. These derivatives are injected into the cow's skin and a delayed-type hypersensitivity reaction, measured as swelling at the injection site after 72 hours, is indicative of *M. bovis* infection [30]. As part of the CMI response to *M. bovis* infection, T (Thymus) cells that come into contact with breakdown/secreted *M. bovis* are sensitised to these antigens [27]. Clonal expansion of sensitized T cells is a form of immunological memory. If an infected cow is further exposed to *M. bovis* antigens (via a tuberculin injection) it is expected to exhibit a visible hypersensitivity reaction. Since the data analysed in subsequent chapters are sourced from the UK and New Zealand, the differences between the test-and-slaughter regimes in the UK and New Zealand are discussed.

Although the use of the tuberculin skin test is standard, the method of application often differs [30]. For example, in the UK the Single Intradermal Comparative Cervical Tuberculin test (SICCT) is employed [33, 10, 34]. In addition to the *M. bovis* tuberculin, a *M. avium* tuberculin is injected in the SICCT [35]. The hypersensitivity reaction to the *M. bo-*

vis tuberculin is measured against the swelling resulting from the injection of the *M. avium* tuberculin. This comparison is used to reduce the number of false positive reactions to the skin test that are the result of exposure to environmental or *M. avium* mycobacteria [30]. In contrast to the UK, in New Zealand the Caudal Fold Single Intradermal Tuberculin test is used (CFT) [36]. The CFT is administered at the base of the tail, in contrast to the neck site used for the SICCT. In addition, there is no comparative aspect to the CFT.

To evaluate the efficacy of the tuberculin skin test, the test sensitivity and specificity are estimated (Figure 1.3). Sensitivity defines the probability of detecting an infected individual. The probability of a negative test result whilst conducting the test on a non-infected individual is defined by the specificity.

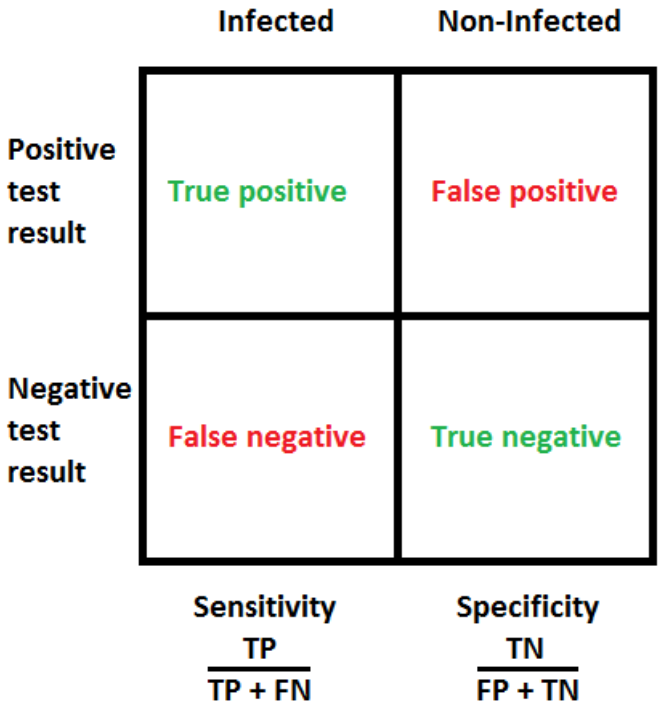

Figure 1.3: How the sensitivity and specificity of a test are calculated. TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

The specificity of the SICCT is estimated to range from 78.8% to 100%, the median estimate lies at 99.5% [30]. The median specificity of the CFT is slightly lower, 96.8% across a range from 75.5% to 99.0%. The difference in specificity between the SICCT and CFT is likely to be most pronounced in areas where *M. avium* has a high prevalence, which provides the motivation for the SICCT to be used across the UK and Republic of Ireland.

The sensitivity of the tuberculin skin test is estimated to range from 52 to 100% [30]. More recent published estimates are lower (50-70%) [37, 38, 39]. Costello *et al.* [35] used the presence of visible tuberculous lesions, found in 453 of 2528 SICCT tested cattle taken from depopulated herds, to estimate a comparatively high estimate (90.5%) for the test sensitivity. This high estimate may have resulted from selecting cattle that were likely to be in a

late stage of infection. Recently, Latent Class Statistical (LCS) models have been employed to estimate the skin test sensitivity. Clegg *et al*. [37] analysed routine bTB skin test results from low and high risk areas of the Republic of Ireland. The observed results of the SICCT were used to estimate the latent/unknown test sensitivity to be approximately 58.1%. These two studies demonstrate that there is much variation in the estimates of the tuberculin skin test sensitivity. The difference between these two estimates is likely resulting from testing animals in different stages of infection. There are many additional factors that could affect the test sensitivity such as: co-infection, immunosuppression, vaccination, and variation introduced by different people carrying out the test (see [30]).

Co-infection with liver fluke (*Fasciola hepatica*) was recently shown, by Claridge *et al*. [40], to be associated with a reduced hypersensitivity reaction to the SICCT. Claridge *et al*. aren't able to provide a definitive reason for this relationship but they explored a number of different hypotheses: their association is coincidentally resulting from unknown correlated factors influencing the bTB and *F. hepatica* distributions; co-infection with liver fluke compromises the cow's immune system resulting in a dampened immune response to the tuberculin injection; or *F. hepatica* infection provides a protective effect against bTB. Garza-Cuartero *et al*. [41] recently found a protective effect of co-infection with *F. hepatica*. The mycobacterial burden of co-infected animals was found to be lower and an *in-vitro* experiment demonstrated a lower uptake rate and recovery of mycobacteria as a result of co-infection.

1.2.2.2 γ-Interferon Assay

In an attempt to increase the sensitivity of bTB testing an *in-vitro* γ-interferon (γ-IFN) assay is often used to complement the tuberculin skin test [30]. During the initial stages of the CMI response when T-cells are interacting with the invading *M. bovis* bacilli, γ-IFN (a cytokine) is released to stimulate macrophages [27, 30]. The γ-IFN assay is an *in-vitro* blood test that exposes samples to either avian or bovine tuberculin and the amount of γ-IFN produced is quantified using an Enzyme-Linked Immunosorbent Assay (ELISA) [42]. Importantly, the γ-IFN assay is able to identify animals in an earlier stage of infection than the tuberculin skin test (Figure 1.2) and its interpretation and implementation are subject to less individual-based variation. In the UK the application of the γ-IFN assay is limited to large or persistent breakdowns [1] (due to its lower specificity) in order to improve the chances of identifying any infected cattle [27, 2, 42].

A range from 66-100% has been estimated for the sensitivity of the γ-IFN assay, though generally values are nearer 85% [30, 42, 37, 39]. The specificity of this assay is estimated to

[1] A herd breakdown is when a cow on the herd has reacted positively to the SICCT, and its infection has been confirmed during a post-mortem exminmation or through *M. bovis* culture.

fall between 87-97% [30, 37, 38]. The lower specificity of the γ-IFN assay, in comparison to the tuberculin skin test (specificity is estimated to be >99%), is the reason why the use of this assay is limited to an ancillary role. With a lower specificity there is a higher chance that a non-infected animal will be incorrectly identified as infected. When a large number of infected cattle have already been detected in a herd or the infection is persisting on a herd the γ-IFN assay will be used alongside the skin test [30]. In these situations the priority is to remove infected individuals and the relative cost of accidentally removing non-infected individuals is low.

In New Zealand the γ-IFN assay is used to increase both specificity and sensitivity [36]. Blood samples from CFT positive cattle are tested to increase specificity and the γ-IFN assay is used in parallel with the CFT in high risk areas. In areas of relatively low risk, where the cost of a false positive is high, a more specific but less sensitive *M. bovis* antigen is used in the γ-IFN assay to increase the specificity [36].

1.2.2.3 Bacteriological Culture

Bacteriological culture of *M. bovis* bacteria is used as the definitive means of confirming infection in cattle. Samples (taken during post-mortem examinations from lesions, pooled lung and lymph nodes, or sputum, for example) are decontaminated and cultured. Any growth is examined for the presence of *M. bovis*. Bacteriological culture can take several weeks. How successful culture is at identifying *M. bovis* is thought to be much higher when bTB lesions are found and can be used [43], although this is difficult to quantify given that there is no test with 100% sensitivity. The presence and distribution of bTB lesions in cattle is dependent on the stage of infection [44, 43]. During the early stages of infection, where at most only a few small lesions are present, the likelihood of finding any bTB lesions during post-mortem examinations is low. PCR (Polymerase Chain Reaction) based methods can also be used to test for the presence of *M. bovis* bacteria in samples [45, 46]. These PCR-based methods are considerably faster, taking hours rather than weeks but neither culture or PCR based methods are used to identify bTB routinely, however, culture is used routinely as part of the disease confirmation process.

1.2.2.4 Vaccination

A vaccine against bTB in cattle is available; the Bacillus Calmette-Guerin (BCG) vaccine originates from an attenuated and non-virulent strain of *M. bovis*. Neither New Zealand nor the United Kingdom have implemented the use of the cattle vaccine. The vaccine has been shown to provide at least partial protection, reducing the infection extent and bacterial load [47, 48, 49, 50, 51]. The current tuberculin skin tests and γ-IFN assay are unable to distuinguish between infected and vaccinated cattle. As long as these tests remain the foundation of

the test and slaughter programs in New Zealand and the UK, the use of the cattle bTB vaccine is restricted [50]. Research into the use of additional non-BCG *M. bovis* specific antigens in bTB tests is ongoing and promising [52, 47, 42, 53, 54].

1.2.2.5 Alternative bTB Tests

Currently the tuberculin skin test and the γ-IFN assay are deemed the most cost-effective and accurate techniques for surveilling bTB in cattle herds across the UK and New Zealand, but their problems mean there is a considerable drive to find alternatives. As research into improving the existing and new tests continues; one of the most promising technologies is perhaps the Multiplex Immunoassay (MI). Unlike the SICCT and the γ-IFN assay, the MI investigates the antibody response in a blood sample following the introduction of a range of antigens - up to 25 [55]. The MI has an estimated sensitivity of 68.6% and specificity that ranges between 92 and 99.8% [37, 56].

1.2.2.6 Control in the UK

In the United Kingdom, the compulsory national eradication program has been in place since the 1950s [57]. The frequency of routine herd testing varies across the UK. In Northern Ireland and Wales, herds are tested annually. In contrast, in England the testing is scaled in an attempt to limit the spatial spread of endemic bTB areas in the southwest. Three different regiments are used, 6 month, annual and quadrennial testing. The quadrennial testing area extends into Scotland, where despite being recognised as bTB free, herd breakdowns do occur, but these can always be linked to cattle imports [10]. In the UK, if an animal tests positive to the SICCT or γ-IFN assay and infection is confirmed via post-mortem examination and/or culture, the whole herd is subject to repeated testing at 60-day[2] intervals until two successive herd tests are passed (i.e. all the animals in the herd test negative for bTB) [57]. In addition, neighbouring herds are tested more frequently. Until the herd has passed its second successive test it is placed under strict movement controls and a severe interpretation of the SICCT is employed [57]. For the severe interpretation, the difference between the swelling resulting from the bovine and avian tuberculin injections necessary for the animal to be deemed infected is reduced [30]. This threshold reduction increases sensitivity at the cost of specificity. The severe interpretation may also be used retrospectively after an infected animal is found in a herd [44]. If a reactor's[3] infection is not confirmed during a post-mortem examination or through bacteriological culture, the herd is re-tested (under the standard interpretation) 60 days and 6 months later [44]. If both the follow up tests are clear the herd is returned to its normal testing regimen.

[2] Skin reactivity to the SICCT is known to be depressed for a period following a previous SICCT - desensitisation - a minimum of 42 days is stipulated [30]

[3] A reactor is a cow that has reacted positively to the bTB skin test.

1.2.2.7 Control in New Zealand

A compulsory test and slaughter regime was introduced in New Zealand in the early 1960s for dairy herds and in 1970 for beef herds [58]. The frequency of herd testing, and the strictness of their interpretation, is dependent on the risk of infection to the herd from other farms and wildlife [36]. New Zealand's land is classified according to the infection status of wildlife [59]. Areas are defined as either TB Vector[4] Risk Areas (VRAs) or TB Vector Free Areas (VFAs). In New Zealand it is estimated that 70% of new herd breakdowns in VRAs can be attributed to wildlife [7]. Within VRAs and VFAs the land is further classified [59]. The VRAs are made up of Movement Control Areas, where routine testing is conducted annually and pre-movement tests must be conducted, and Special Testing Areas - Annual (STAA) with only annual testing. VFAs are made up of Special Testing Areas - Biennial (STAB), where herds are tested every two years, and Surveillance Areas (SAs) with triennial testing.

In New Zealand, herds are defined by how many years they have retained their non-infected status [60]. A herd's infected status[5] is removed once the herd has passed two CFT tests at least six months apart [36]. Any CFT positive animals must be negative in a follow up γ-IFN assay to increase specificity and avoid economic losses due to false positives. For the second test, the γ-IFN assay is used in parallel to the CFT to increase testing sensitivity and increase the chance of finding all the infected animals. In addition to the routine test and slaughter regimes implemented in the UK and New Zealand, surveillance of all cattle carcasses is conducted in abattoirs [36, 57]. Every cow slaughtered is examined for the presence of bTB-like lesions and trace-back to the source herds is instigated if bTB lesions are found [30].

1.2.2.8 Success of bTB Control

The progress of the control regimes in New Zealand and the UK has differed quite dramatically. The initial success of New Zealand's eradication campaign, which targets bTB in wildlife and cattle, meant funding was reduced in the late 1970s [61]. The number of infected cattle herds spiked and control funding was reinstated in the early 1990s [7]. Since the mid 1990s bTB prevalence in New Zealand has dropped dramatically, with only 39 infected herds (0.06% of the total herds) remaining by 2015 [60]. Control of bTB in New Zealand is funded by both the government and the farmers themselves [58]. Being farmer led has meant that the eradication programme has a high level of acceptance and could go some way to explaining its success. Similarly to New Zealand, the test and slaughter schemes across the UK were effective (especially in the 1960s and 70s) in drastically reducing the prevalence of bTB in cattle herds [57]. By the late 1970s there were less than 1000 infected cattle,

[4] In the context of bovine tuberculosis, a vector is defined as a species that are susceptible to *M. bovis* infection and capable of transmitting it onwards to other species

[5] Herds where TB infection in a cow has been confirmed in the laboratory by either culture or PCR are defined as infected [36]

dropping from almost 17,000 in the early 1960s [30]. Similarly to New Zealand, this early success did not continue. However, in contrast, increasing bTB prevalence has continued in recent decades in the UK [29, 62, 57]. The Foot-and-Mouth Disease (FMD) outbreak in 2001 was very detrimental to bTB control in England and Wales, it led to a spike in prevalence - back up to pre-control levels [57]. This increase in prevalence was most likely the result of the lack of bTB control during the FMD outbreak, and increased movement of cattle during restocking once the outbreak had been controlled [63]. An additional contrast to New Zealand, is that bTB control in the UK is entirely funded by the government [34].

The attitudes towards the wildlife bTB reservoirs in UK and New Zealand are very different. In New Zealand, the brustail possum (*Trichosurus vulpecula*) is considered to act as a bTB reservoir [61, 36, 16]. The brushtail possum was introduced into New Zealand in 1858 to initialise fur production and trade, but populations quickly spread across almost all of New Zealand and became a significant conservation issue [64]. Control of possum populations is a major part of New Zealand's eradication campaign and it is generally well supported by the public [60]. In the UK, up until the 1970s the European badger was controlled to maintain low population densities [65]. These operations were halted in response to the badger becoming a protected species in the UK [66] and any wildlife control targeted at this species is met with considerable resistance.

1.2.2.9 Cattle to Cattle Transmission

When a herd breakdown occurs, it is important to understand where the infection came from. The low sensitivity of the currently available tests mean that cattle-to-cattle transmission often has time to occur [29]. Under the stringent testing regiments of the UK and New Zealand, it is likely that *M. bovis* infection in cattle populations manifests mainly as a sub-clinical infection, with a low-level of infectiousness [35, 44, 43]. As discussed earlier, cattle are susceptible to extremely low doses of infectious bacilli and, therefore, the low and intermittent shedding by bacteria of these sub-clinically infected cattle may be enough to promote cattle-to-cattle transmission and maintain bTB persistence.

Transmission of bTB from experimentally infected cattle has been demonstrated, although high doses were necessary to reliably achieve transmission [67, 68]. The fact that the majority of breakdowns in the UK involve less than three reactors would suggest that cattle-to-cattle transmission is rare, relative to amount of contact between cattle in a herd [44, 69]. However, the potential for bTB transmission within and between cattle herds is large. Schoenbaum *et al.* [70] traced a single infected beef cow, discovered during routine abattoir surveillance, to a single farm in Oklahoma from which 13 cattle had been sold, which further exposed >4000 cattle. In Scotland the four year testing intervals mean that should infection be introduced, it could be a number of years before cattle are tested. Chalmers *et al.* [71] reported a bTB outbreak in the south west of Scotland. Tracing of a single test positive

calf identified a heavily infected herd where a third of its 261 cattle were skin test positive [71]. A modelling study done by Barlow *et al*. [72], based on cattle herds in New Zealand, suggested that within-herd bTB transmission occurred and that it made an important contribution to maintenance of bTB in New Zealand.

Long distance movements of cattle from high risk areas of infection are recognised as an important driver for bTB spread [73, 69, 63, 74]. Gilbert *et al*. [73] demonstrated that cattle movements, particularly in endemic areas of the UK, were a strong predictor of disease occurence. In 2008, Green *et al*. [75] further examined the role of cattle movements in the spread of bTB. Green *et al*. [75] used a sotchastic model, which explicitely used recorded cattle movements, to demonstrate that cattle movements were an important mechanism of spread and that their surveillance could greatly limit the growth of, and spread from, high bTB prevalence areas. This recognition endorses the pre-movement testing that is implemented for movements from high risk areas in both the UK and New Zealand [57, 59].

1.2.2.10 Environmental Persistence

Numerous studies have demonstrated that *M. bovis* can survive outside of the host for extended periods [76, 77, 78, 79]. Environmental persistence of bacteria in the soil could be an additional means by which bTB can persist in cattle herds. Duffield and Young [77] examined the persistence of *M. bovis* bacteria in a range of laboratory conditions, demonstrating that viable bacteria could be isolated from soil kept in the shade after 4 weeks. Jackson *et al*. [76] examined environmental persistence outside of the laboratory finding that survival time was greatly influenced by season. Jackson *et al*. [76] concluded that in summer it was likely to have a limited role in bTB persistence.

Recurrence of cattle herd breakdowns is a problem for control in both the UK and New Zealand [80, 81, 82, 83]. This recurrence could be the result of not removing all the infected cattle from the herd. The low sensitivity of the current tests mean that infected cattle can be missed. Dawson *et al*. [83] evaluated the potential factors associated with herd breakdown recurrence in New Zealand. This research found that there are likely to be a number of reasons for recurrence including missed infection, environmental persistence and the presence of a wildlife reservoir. Young *et al*. [78] investigated a herd in the Republic of Ireland with history of breakdowns, which was subjected to a whole herd cull and restocking. Four months after the herd depopulation, molecular typing methods revealed the presence of viable *M. bovis* bacteria. Herd depopulation is one means of combating recurrence of bTB on a cattle herd. Good *et al*. [84] found that the Irish herd bTB depopulation policy was effective.

1.2.2.11 The use of Whole Herd Depopulation

An additional contrast in the bTB control regimes in place in the United Kingdom and New Zealand is in the use of whole herd depopulation. In New Zealand if a high number of re-

actors with an extensive infection (as observed during post-mortem examination) are found on a herd, the herd is depopulated [85]. Depopulation is rarely used in the UK; mandatory conditions for whole herd depopulation are that $\geq$25% of the herd reacted positively to the test and the herd infection must be confirmed through culture. Although these mandatory conditions are similar to those for New Zealand, there are many additional factors that are considered before a whole herd cull will be implemented in the UK. These factors include the extent of disease in the reactors, herd size, the history of bTB on the herd and local incidence of bTB - lower favours depopulation. Herd depopulation will be most effective when the outside risk of re-infection is low [85, 86, 87]. Therefore in areas of New Zealand where the infection risk from wildlife is minimal, herd depopulation is likely to be effective. In contrast, given the higher and increasing prevalence of bTB in endemic areas of the UK the use of herd depopulation is likely to be ineffective. Although whole herd depopulation has been shown to be an effective method for eradicating disease in a herd, it is not generally favoured, both for animal welfare and economic reasons [84, 87, 88]. Without depopulation, more emphasis is placed on the accurate identification of individual infected animals.

1.3 Molecular Epidemiology of Bovine Tuberculosis

1.3.1 Restriction Fragment Length Polymorphism Typing

Molecular tools have been instrumental in the investigations into the potential sources of infection into cattle herds. Over the last three decades there has been a continuous progression through different molecular typing methods for *M. bovis*. This progression is motivated by improvements in speed, price, repeatability and ease. Early methods that used Restriction Fragment Length Polymorphisms (RFLPs) provided evidence of the role of the brushtail possum in New Zealand [89, 90]. For RFLP, *M. bovis* DNA is digested using a range of restriction enzymes and the resulting fragments are separated using gel electrophoresis [91]. The banding patterns on the gel can be then be used to discriminate strains of bTB and define molecular types. Restriction Endonuclease Analysis (REA), a method that uses RFLP, was used routinely in New Zealand up until 2012 [92]. Early research by Collins *et al.* [93] found evidence of inter-species transmission by showing that livestock and wildlife from the same region shared REA types. In addition, regionalisation found by Collins *et al.* [94], means REA typing could be used to investigate whether a new herd breakdown was locally sourced or not. Skuce *et al.* [95] conducted a study to determine the utility of RFLP typing methods for discriminating *M. bovis* isolates sourced from cattle. By combining RFLP typing data with knowledge of cattle movements it was possible to determine that the sampled herd outbreak could be linked to a herd nearby via the movement of an infected cow.

1.3.2 Spacer-oligotyping

In the UK, spacer-oligotyping (spoligotyping) took the place of RFLP typing methods in the late 1990s. Spoligotyping offers a more rapid, repeatable and easier to perform technique [96, 97, 98, 99]. Spoligotyping characterises strains by using a single direct repeat region of the *M. bovis* genome that is interspersed with non-repetitive short sequences of nucleotides (spacers) [96]. A strain type is defined by the presence, or absence, of known spacer sequences within this repeat region. The ease and speed of spoligotyping enabled Smith *et al.* [100] to conduct an analysis of 11,500 *M. bovis* isolates from Great Britain. Smith *et al.* [100] revealed a high degree of geographic localisation, similar to that shown for REA types in New Zealand, and hypothesised that it was the result of clonal expansion of *M. bovis* across Great Britain. Geographic localisation of different strains is an important characteristic of *M. bovis* populations, which increases the utility of molecular typing methods since particular types can often be tied to certain areas [62]. Elsewhere in Europe, spoligotyping has provided evidence of inter-species transmission of *M. bovis* between wildlife and livestock. In Spain, Aranaz *et al.* [101] found that deer, wild boar, Iberian lynx and cattle from the same geographical area were infected with the *M. bovis* strains of identical spoligotypes.

1.3.3 Variable Number Tandem Repeat Typing

Variable Number Tandem Repeat (VNTR) typing represents the successor of spoligotyping and REA typing. The genome annotations available for the full genome of *M. bovis*, published in 2003 by Garnier *et al.* [102], help to identify polymorphic locations that are made up of tandemly repeated DNA [103]. Variation can occur in the number of repeats present at these polymorphic locations. VNTR typing utilises a number of the locations on the *M. bovis* genome and uses the number of repeats present at the locations, as a means of discriminating and typing *M. bovis* strains. Spoligotyping uses a single direct repeat region, whereas VNTR typing utilises multiple tandem repeat regions on the genome. Using multiple polymorphic regions enables VNTR typing to have an increased discriminatory ability in comparison to spoligotyping [103, 104]. The selection of which polymorphic regions for VNTR typing to use can affect the discriminatory ability and also limits the generalisability of results [105, 106]. Furphy *et al.* [22] recently used VNTR typing to provide further evidence of inter-species transmission between badger and cattle populations, by showing that these populations shared a large number of VNTR strain types. In 2012, VNTR typing replaced REA typing in routine bTB surveillance in New Zealand [92]. Although REA typing is a highly discriminatory tool, more so than VNTR typing [92], the technique is complex and time consuming and the gels produced (during the gel electrophoresis) must be interpreted manually [36].

1.3.4 Whole Genome Sequencing

The history of the molecular typing methods used for *M. bovis* represents a progression. As technology improves, and research continues, new methods appear that are faster, easier, or less complex. The recent advent of Next Generation Sequencing (NGS) platforms has opened up the possibility to use Whole Genome Sequences (WGS) in epidemiological investigations [107]. Automated Sanger sequencing platforms, the first generation, were based on the chain termination method [108]. A single stranded fragment of DNA is repeatedly replicated using DNA polymerase and during the replication a terminating nucleotide is present and can be bound at any time. Different sized fragments of DNA will result, following the binding of the terminating nucleotide at different stages of the replication. Gel electrophoresis can be used to separate the fragments by their size and the pattern resulting will correspond to the DNA sequence. Currently there are a number of NGS platforms available, which are competing with the Sanger method. Sequencing with these methods is considerably faster and cheaper than with Sanger techniques [108]. The sequencing, for NGS platforms, is completed during a DNA replication phase. The replication is continuous, unlike in the chain termination method, and interpretation doesn't rely upon gel electrophoresis [109]. However, NGS methods are more error prone than Sanger platforms [110]. The huge volume of data that NGS platforms produce must be interpreted carefully to remove and limit the effect of any sequencing errors that may be present.

Illumina sequencing platforms are an example of NGS technology [109]. A fragmented genome is bound into a flow cell, and each fragment is clonally amplified to form clusters on the cell. The clusters are repeatedly sequenced using fluorescently labelled nucleotides. Each nucleotide bound during sequencing releases a signal that is recorded. This sequencing-by-synthesis method rapidly produces vast quantities of sequence data.

NGS technology could allow WGS to replace VNTR typing in the routine surveillance of bTB [111, 112, 113]. In recent applications to human tuberculosis, WGS has added a great deal of resolution to epidemiological investigations [114, 115, 116, 117, 118, 119, 120]. Gardy *et al*. [115] combined WGS of *M. tuberculosis* isolates with detailed social network information to investigate a recent human tuberculosis outbreak in Canada. The WGS data revealed that the previously thought clonal outbreak was actually two co-circulating outbreaks. Given this finding, the social network data was refined to reveal several transmission events and identify three individuals that the two outbreaks originated from. In addition, it was possible to link these two outbreaks to an increase in drug usage a few years prior. Bryant *et al*. [117] also examined the potential to use WGS data to elucidate patient-to-patient transmission. WGS were the highest resolution genetic data available but it is important to understand that what can be learnt about a system is dependent on how the rate of change of the pathogen relates to the speed of the process being investigated [121]. Bryant *et al*.'s [117] research found that the substitution rate of *M. tuberculosis* was low and variable, which made it difficult to track direct patient-to-patient transmission. This variability could result from there being different stages of infection for human tuberculosis, similar to those seen for bTB [120].

Recently WGS techniques have been applied to bTB systems [24, 25, 26]. The earliest work, by Biek *et al*. [24], found evidence of badger populations in Northern Ireland being involved in the persistence of bTB in local cattle populations. This work, built upon by Trewby *et al*. [25], emphasises the utility of WGS in surveillance. Following a recent outbreak of bTB in Minnesota, a previously TB free state in the United States (US), Glaser *et al*. [26] conducted a retrospective analysis of WGS data from isolates. This analysis revealed that the outbreak was the result of a recent introduction into the area from southwestern US or Mexico, rather than neighbouring states Michigan and Manitoba where bTB remains endemic. This retrospective analysis also found that the outbreak was circulating in both cattle and deer populations.

1.4 Bovine Tuberculosis in Wildlife

In the UK, New Zealand and elsewhere, wildlife populations have been implicated in the spread and persistence of bovine tuberculosis ([18, 122, 16]). *M. bovis* is capable of infecting a wide range of mammals [11, 13, 18]. Understanding the role of different species in the maintenance of a disease is not a problem specific to bTB, it is an important problem in the control of a wide range of diseases [123]. The different wildlife species implicated in the spread and persistence of bTB in livestock in the UK and New Zealand are discussed.

Corner [13] established five points to investigate when examining the role a particular species has in the maintenance of bTB:

1. How does the animal get infected?
2. Where does the infection manifest in the body?
3. Where is *M. bovis* excreted and to what extent?
4. By what means does the infection pass into livestock populations?
5. What is the minimum infective dose necessary by each route?

1.4.1 Brushtail Possum (*Trichosurus vulpecula*)

It is estimated that there are approximately 30 million brushtail possums (*Trichosurus vulpecula*) in New Zealand, at densities varying from 1-20 individuals per hectare [124, 16]. These possums are solitary animals with overlapping home ranges, which vary in size depending on the landscape [125, 16]. The brushtail possum is recognised as a maintenance reservoir[6] for *M. bovis* infection in New Zealand [16]. As the prevalence of bTB in New Zealand's cattle populations continues to drop, now with only 39 infected herds [60], possum populations are increasingly implicated as a source [7, 16]. Prevalence in a naturally infected possum population has been estimated to be less than 5%, although this has been shown to vary with the possum population density [16]. Nugent *et al.* [16] argue that possum populations were only infected with bTB in the 1960s, in line with when an infected possum was first observed in 1967 as hunters would have encountered infected possums earlier had they been present. The current thinking is that *M. bovis* was introduced into possum populations from deer as a result of commercial deer hunting [126, 7, 16].

Generally, brushtail possums are highly susceptible to bTB, suffering a heavy generalised infection and dying within a few months [127, 128, 13, 16]. Buddle *et al.* [127] demonstrated this high susceptibility through experimental infections. Upon challenge with low infectious doses, the artificially infected possums died within 8-10 weeks [127]. Ramsey and

[6] A maintenance reservoir is a population that is capable of intra- and inter-species transmission and in which *M. bovis* infection can independently persist.

Cowan [128] tracked naturally infected possums and found that they survived an average of 4.7 months. The infection in possums manifests mainly in the lungs and associated lymph nodes, and excretion of infectious bacteria occurs in an aerosol form [129, 16].

Possum-to-possum transmission has been observed in captivity. Corner *et al.* [129], by housing experimentally infected and susceptible possums together, recreated transmission of bTB. The transmission rates were low and affected by how social the possums involved were. A field trial conducted by Whitford *et al.* [130] confirmed this low transmission rate. Transmission between possums is likely to occur via aerosol or open wounds during direct contact [16]. Transmission of bTB from possums to cattle is thought to occur through heavily infected possums. Terminally ill possums have been shown to no longer avoid cattle and cattle, being inquisitive, will often investigate dying possums by licking and biting them [131]. The increased ranging behaviour of infected possums was investigated by Ramsey and Cowan's [128] research and home ranges were, on average, larger but not significantly so.

In addition to the brushtail possum, there are a number of other species thought to be important in the maintenance of a wildlife bTB reservoir in New Zealand: ferrets (*Mustela furo*), pigs (*Sus scrofa*), and red deer (*Cervus elaphus*) [126, 122]. These species have considerably larger home ranges, in comparison to possums, and are thought to act as spatial vectors of disease linking otherwise isolated possum populations [126, 132, 122].

1.4.2 Other Wildlife Hosts in New Zealand

There are three species of wild deer known to carry bTB in New Zealand, of which the red deer are predominant [122]. Nugent, in his PhD thesis [133], estimated prevalence in deer populations to vary from 8-37%. Nugent estimated these prevalences by examining deer populations living in the same area as uncontrolled infected possum populations. Deer are thought to become infected through contact with terminally ill or dead infected possums [134]. Their infection is mainly restricted to the head and neck and there is little or no potential for deer-to-deer transmission [134, 133]. Deer populations are unable to maintain infection in the absence of possums unless they are at high densities [134, 122]. Importantly though, deer-to-possum transmission can occur through possums feeding on deer carcasses [133]. The long lifespan of deer means they can act as temporal vectors of disease, enabling bTB infection to persist for many years following the removal of possums [126]. Since prevalence of bTB in possums is generally low it is often difficult to determine if a population is infected and, therefore, infection in sympatric[7] deer populations is used as a indicator of an infected possum population [134, 135].

Feral pigs and ferrets can additionally be used as sentinel species to detect an infected

[7] Sympatric populations are those that occupy the same geographical space at the same time.

possum population [134, 126, 122]. Although very high prevalences of *M. bovis* infection have been observed in these populations, drops in prevalence following the removal of possums would suggest that they generally aren't acting as maintenance hosts [126, 122]. That being said ferret populations, in pockets of high density, have been observed to maintain bTB infection in the absence of possums [136]. Ferrets and pigs are scavenger species and will readily pick up infection from infected possum carrion [122].

1.4.3 European Badger (*Meles meles*)

In the United Kingdom, European badger (*Meles meles*) populations are implicated in the maintenance of bTB infection in cattle populations [137, 138, 6]. In contrast to the solitary lifestyle of the possum, badgers live in relatively stable social groups [65]. Each badger group inhabits a territory containing multiple setts, which consist of a main sett and a few subsidiary setts that they'll bed in. Group size can vary considerably, in high density areas, such as in southwest England, group sizes of 8-20 individuals are observed [65]. Whereas, in lower density areas, for example in the Republic of Ireland, a group usually only includes around 3 badgers [65, 139]. The dispersal rates of badgers between groups is thought to vary with group size, being lower in high density populations [139]. Despite living in stable social groups, extra-group mating events have been shown to be relatively common, accounting for up to 50% of cubs in a badger population in Luxembourg [65, 140].

M. bovis infection in badgers results in a chronic infection primarily of the respiratory system, although biting presents a secondary source of infection [141, 142, 143]. As with possums the infection is generally restricted to the lungs and associated lymph nodes [144]. In contrast to possums, in the majority of badgers, *M. bovis* infection is thought to manifest itself as a small localised infection with little or no infectiousness [145, 65]. A small proportion of infected badgers may succumb to a heavy widespread generalised infection [143]. In these badgers the bacteria appears to spread from the lungs into a variety of other organs throughout the body [141, 143, 65, 144]. These individuals excrete large amounts of infectious *M. bovis* in faeces, urine and as an aerosol and have a significantly higher mortality rate [146].

In the southwest of England, badger and cattle populations live in close proximity [147]. Their close proximity mean that there is great potential for inter-species transmission of *M. bovis* infection. Cattle and badgers have frequently been observed interacting both directly and indirectly [148, 149, 150, 151, 152]. The evidence would suggest that the European badger is capable of acting as a maintenance reservoir for *M. bovis* in some areas of the United Kingdom. The badger is able to survive for long periods (several years) with *M. bovis* infection, even whilst suffering from a highly infectious generalised infection [153, 141, 146]. Although a number of other british wildlife species are susceptible to *M. bovis* infection, the badger populations alone are thought to be acting as the wildlife reservoir of disease [154, 11].

1.4.4 Control of bTB in Wildlife: New Zealand

The contrasting views regarding the different wildlife hosts in the UK and New Zealand, have led to markedly different control strategies. In New Zealand, possum population control has been implemented since the possums were implicated in the persistence of bTB in cattle populations in the early 1970s [155]. Research conducted in the Buller district, found on the northwest coast of New Zealand's south island, provided strong evidence for possum populations infecting cattle [7]. Test-negative cows were grazed on pasture where infected possums had been captured 6 months previously. After 6 months, 26 of the 29 cows were test-positive. These findings prompted extensive possum control throughtout the district that resulted in a 30% drop in the number of infected herds [7].

In 2011, following a 90% reduction in the number of cattle reactors since 1994, the objective of New Zealand's bTB control campaign was changed from control to eradication [7]. The current campaign centres around Vector Risk Areas (VRAs) of New Zealand, defined as areas where bTB infection of wildlife has been confirmed or is strongly suspected [7]. VRAs are split into four categories as described in Figure 1.4 and below.

1. TB Eradication - removal of bTB from cattle and wildlife by 2026.
2. TB-Free Area Protection - maintain low possum population densities to prevent spread of VRAs.
3. Infected Herd Suppression - control of possum populations surrounding farmland to minimise bTB transmission into cattle herds.
4. Proof of Concept Areas - designated areas where trials are on-going to determine whether bTB eradication is feasible.

M. bovis has been successfully eradicated in the Proof of Concept Area found on New Zealand's south island, shown in Figure 1.4 and described above, and eradication is close in the other Proof of Concept Area [60].

Possum populations are reduced through the use of traps and distribution of 1080 (metabolic toxin sodium fluoroacetate) poison [64]. The 1080 poison, via interference with the metabolic system, is lethal to possums and causes cardiac or central nervous system failures [156]. Distribution of 1080 via helicopter is conducted routinely in order to reduce possum density across large areas of inaccessible forest [64]. Aerial operations are restricted on farmland and near residential areas; here control is limited to ground based use of traps and poison.

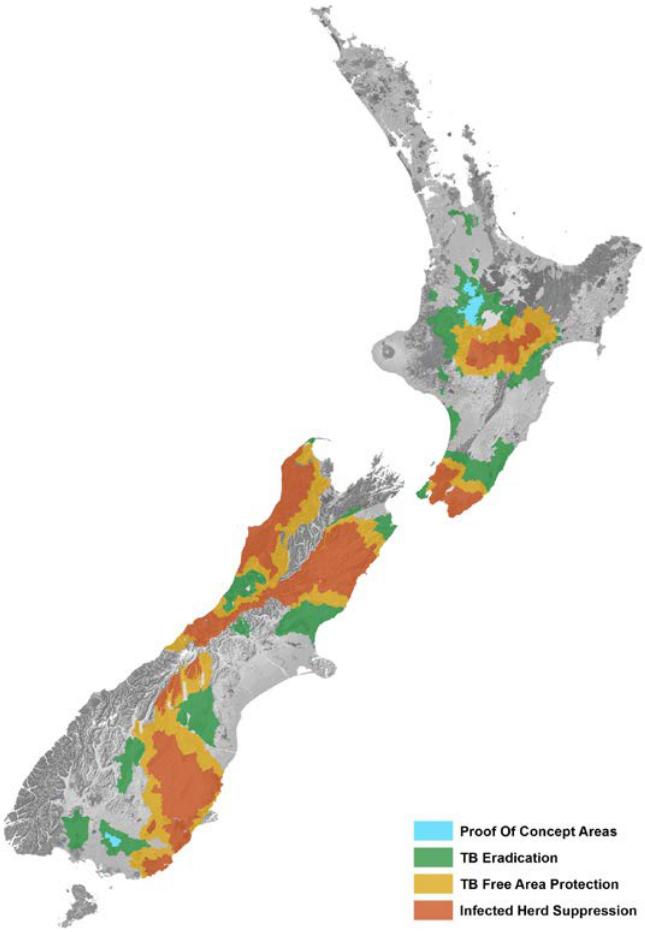

Figure 1.4: The categorisation of Vector Free Areas across New Zealand. Note that areas coloured in cyan are Proof of Concept Areas, where trials are on-going. Taken directly from OSPRI Annual Report [60].

The lethality of 1080 isn't limited to possums, death of non-target species has occurred. Initially when the poison was baited using carrots, from 1960-80, important native bird species, such as the New Zealand robin (*Petroica longipes*), would commonly succumb [64]. Cereal baits replaced carrots, and the risk to native bird species is now considered low. Following the findings that wild deer were dying as a result of consuming the 1080 baits, special deer repellent baits, which are effective for reducing deer casualties, are now routinely used [64]. 1080 is highly poisonous to people and contamination of water sources was a concern but there is no evidence to support it being a problem. In addition, strict legislation is in place to limit the risk for those involved in possum control operations [156]. Given the controversy surrounding the use of 1080, vaccination and fertility controls have been investigated but these are currently not considered feasible alternatives [50, 64].

1.4.5 Control of bTB in Wildlife: United Kingdom and the Republic of Ireland

The first tuberculous badgers were discovered in the early 1970s in England and the Republic of Ireland [144]. This discovery prompted badger population culling operations, which included gassing setts, baiting, trapping and/or shooting [65]. Unsanctioned culling was halted with the introduction of legislation to protect badgers in 1973 [66]. Since the implementation of the badger legislation there have been a number of government sanctioned badger culling trials in the UK, and in the Republic of Ireland. These trials were conducted to determine the extent to which *M. bovis* persistence in badgers influenced the risk and occurrence of bTB in cattle.

The Thornbury trial, which ran from 1975 to 1981 in southwest England, was the first badger culling trial [157]. Badger setts were gassed within a geographically isolated area covering 104km^2. Herd breakdown levels were drastically reduced and maintained for almost a decade. Similar success was noted for the East Offaly Project culling trial conducted in the Republic of Ireland between 1989 and 1994 [158]. The East Offaly Project was conducted over a 738km^2 area, which included a buffer area implemented to limit badgers re-colonising the area.

Following the success of the East Offaly Project, in 1996 a large-scale badger removal program was licenced in four counties of the Republic of Ireland - the Four Areas Project [159]. Within four geographically distinct areas, removal and reference areas were matched and buffer areas were combined with natural barriers to limit re-colonisation. In removal areas repeated trapping was used to reduce the badger population density. The herd breakdown level was reduced by 60-95% across the four areas [159]. The success of the trial acts to endorse the current use of badger culling in the national bTB control scheme [33, 160]. In the Republic of Ireland, for any herds where the source of their breakdown cannot be attributed to cattle movements, badgers caught within a 2km radius are removed.

One of the most recent, and potentially most controversial, badger culling trials was the Randomised Badger Culling Trial (RBCT) based in southwest England, which took place from 1998 to 2006 [161]. A proactive cull was compared to a reactive cull (as is implemented in the Republic of Ireland) and no culling. In the proactive culling areas, culling aimed to reduce and maintain badger population densities as low as possible. In the reactive culling areas, culling was only undertaken in response to a herd breakdown and aimed to remove any badger groups that had access to the herd. The proactive cull reduced the bTB herd breakdown levels although had negatively associated effects in the surrounding areas [162]. The reactive culling was prematurely halted in 2003 because it was causing an increased risk of herd breakdown in the associated areas [161]. Woodroffe *et al*. [163] demonstrated that badger culling, proactive or reactive, increased the ranging behaviour of badgers and disruption of social organisation and, thereby, affected the prevalence of *M. bovis*. The increased ranging behaviour of badger populations was linked to a rise in cattle bTB incidence [162]. As a result of the limited success of the RBCT, badger culling lacks support and isn't currently incorporated into bTB control in England or Wales.

Amongst other things these trials emphasise that the UK and Republic of Ireland represent two different bTB systems, despite the same protected wildlife host being implicated. These differences extend through policy into farming practices and resident badger populations [164, 165]. Generally, high density badger populations and large intensive farming in the UK, contrasts with the sparser badger populations and smaller scale farming practices of the Republic of Ireland [165].

Research into the use of a vaccine for badgers has been extensive. The Bacille Calmette Guerin (BCG) vaccine is the only one being used and it doesn't provide complete protection, but it does reduce the extent of disease progression in experimental and field conditions [166, 167, 168, 169]. Corner *et al*. [168] established that the vaccine could be administered orally if encapsulated in a lipid matrix formulation, which protects the vaccine whilst it is in the stomach. As discussed previously, most bTB infected badgers will have a slowly progressing disease of the lungs that leads to little excretion of infectious material. A small proportion may develop a heavy generalised infection. If the vaccine is successful in reducing the number of badgers suffering a heavier infection, it could greatly reduce the transmission potential of badger populations. The Badger Vaccine Deployment Project was initiated in 2010 by the UK government, giving people the opportunity to train and apply for licensing to conduct badger trapping and vaccination on their land [164, 34]. This project, which finished in 2014, was not designed as a scientific trial but rather aimed at developing the practical knowledge necessary for the large scale deployment of badger vaccinations [170]. The data from the project were collated and the impact of the vaccination on bTB in cattle has yet to be examined.

In Northern Ireland, a Test, Vaccinate or Remove (TVR) program is currently ongoing. Trapped badgers are tested using an animal-side bTB test [171] and any positive animals are culled [172]. This vaccination strategy should act to limit the transmission potential by reducing both the number of badgers with a heavy generalised infection (through the removal of test-positive animals that are more likely react to the test) and the number of badgers that develop one (by distributing the vaccine). In addition, culling of only infected badgers is more likely to be supported by those that are against badger culling operations. The animal-side bTB test used is the Dual-Path Platform VetTB (DPP) assay. The DPP assay has two bands of antigens (proteins derived from *M. bovis* on a membrane strip) [171]. Blood serum is added to the assay well and a visible reaction at the two bands is interpreted as a positive reaction.

1.5 The cost of bTB control

The bTB systems in the UK and New Zealand are contrasting in many aspects, not least in the extent of wildlife control. A major limitation to both campaigns is cost. England and Wales are reported to have expended 0.5 billion pounds in the last decade [6], whilst the cost of control in New Zealand is comparable at more than 50 million New Zealand dollars (£43.5million) per year [61, 7]. As is highlighted by Hone [173], as the prevalence of bTB drops the control of *M. bovis* infection in cattle becomes increasingly costly, relative to the amount by which prevalence is reduced. This trend is well illustrated in Figure 1.5, which shows how the cost of wildlife control in New Zealand has increased to its highest level, despite the number of infected cattle and deer herds being at its lowest. This relationship prompts the questioning of the motivations for bTB control, given its low zoonotic risk now that pasteurisation is routine [174]. But as history has shown, both in New Zealand and the UK, if control of bTB lapses, either through lack of funds or a more pressing disease, prevalence quickly increases threatening this low zoonotic potential. In the UK the Foot-and-Mouth Disease (FMD) outbreak in 2001, prompted a rapid increase in the number of herd breakdowns [57]. In the late 1970s, the reduction of funds due to the initial success of bTB control in New Zealand resulting in an increase from 540 infected herds in 1980 to 1,694 by 1994 [7].

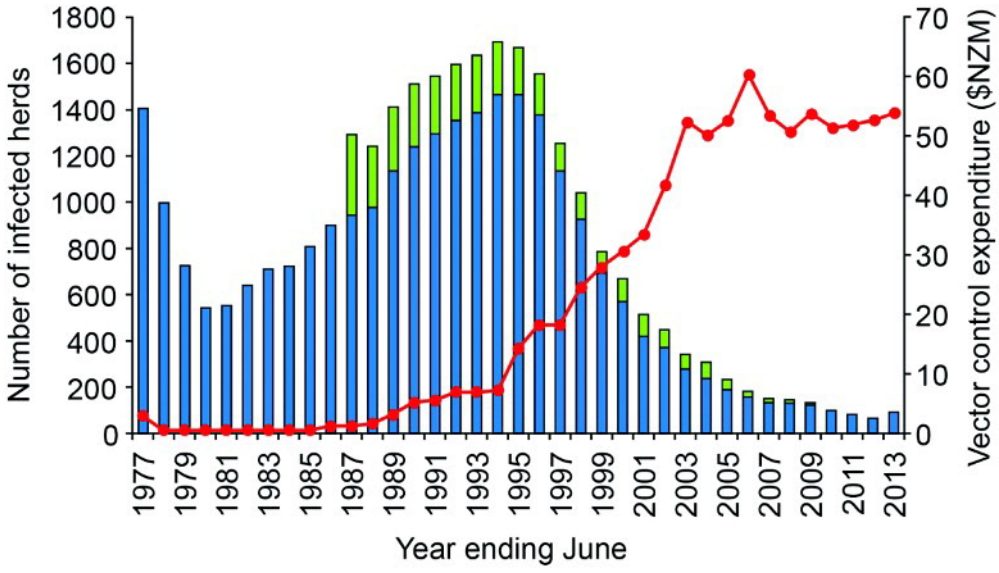

Figure 1.5: The number of cattle (blue bars) and deer (green bars) herds that were found to be infected with *M. bovis* in New Zealand from 1977 to 2013. The cost (in New Zealand dollars (millions)) of wildlife control operations is plotted as a red line over the same period. This figure was taken directly from Livingstone *et al.* [7].

1.6 Chapter Plan

The research for this thesis has centred around the use of Whole Genome Sequencing (WGS) to investigate the transmission of *Mycobacterium bovis* within and between cattle and wildlife populations. The chapters that follow will describe the use of statistical, mathematical and simulation based approaches to investigate three main themes:

1. The utility of WGS in epidemiological investigations of *M. bovis* transmission
2. The application of phylogenetic inference methods to examine the dynamics of bovine tuberculosis
3. The role of wildlife reservoirs in the transmission and persistence of *M. bovis* infection in domestic cattle populations

The layout of the thesis will be as follows:
Chapter 2 - Informing the selection of Single Nucleotide Polymorphisms with epidemiological data
This chapter, on the assumption that the genetic relatedness between the *M. bovis* isolates will, to some degree, be explained by epidemiological data that describes the sampled animals, investigates Single Nucleotide Polymorphism selection criteria. A selection of quality filter criteria were assessed by comparing the genetic data they produced to available epidemiological data. The analyses conducted here were used to inform the selection of filters for the genetic data used in Chapter 3 and Chapter 4.

Chapter 3 - Using Whole Genome Sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand
A selection of WGS *M. bovis* isolates were available, sampled from infected cattle and wildlife across New Zealand. Using the program Bayesian Evolutionary Analysis by Sampling Trees (BEAST), the substitution rate of *M. bovis* was estimated as well as inter-species transmission rates. The dataset used for this chapter had some evident biases in the sampling, these were investigated further in Appendix A.

Chapter 4 - Evidence of inter-species transmission between cattle and badger populations residing within and surrounding Woodchester Park
M. bovis was isolated from samples taken from infected cattle and badgers living within and surrounding Woodchester Park (found in the southwest of England) and the genomes were sequenced. These isolates with WGS data were examined in the context of detailed population data to investigate the presence and extent of inter-species transmission. In addition, the maintenance role of the sampled badger population was examined. Evidence of mislabelling was uncovered during the analyses described in Chapter 4, these are discussed in Appendix B.

Chapter 5 - General Discussion
The final chapter of this thesis will provide a critical examination of the research described

above, commenting on any improvements and how the work could be continued. The results will be put into the context of bTB in the United Kingdom and New Zealand.

Chapter 2

Informing the selection of Single Nucleotide Polymorphisms with epidemiological data

2.1 Introduction

Molecular typing methods are frequently used in epidemiological investigations to trace transmission, determine the extent of outbreaks, identify sources and understand the role of reservoirs. Genetic similarity is taken as an indication of a recent common infection source and used to infer transmission events. The utility of molecular typing methods, which quantify genetic similarity, is measured by their discriminatory ability. Whole Genome Sequencing (WGS) characterises almost all of the genetic data available for a sampled pathogen and maximises the potential to discriminate isolates [107, 111, 121, 112]. *Mycobacterium bovis*, the causative agent of bovine tuberculosis (bTB), is a pathogen that, despite continued control efforts, is endemic in large parts of the UK and Ireland [57]. In New Zealand, although bTB is no longer considered endemic in livestock, it is in wildlife populations, which present a considerable spill-over risk [7, 60]. Current typing methods routinely used in bTB surveillance, such as Variable Number Tandem Repeat (VNTR) typing and spoligotyping, target specific repeat regions of the *M. bovis* genome. These methods are, and have been, very informative in epidemiological investigations but their resolution is limited [62, 175, 92]. For a slowly evolving pathogen like *M. bovis* [24, 25], WGS ensures that as much of the variation between isolates is captured as possible, providing a considerably higher discriminatory power.

Multiple studies of *M. tuberculosis* have demonstrated the utility of WGS in epidemiological investigations [115, 116, 117, 118]. Walker *et al.* [116] used WGS to characterise outbreaks and identify important superspreader individuals in a study of human TB in the UK. Roetzer *et al.* [118] conducted a large analyses of a human TB outbreak in Hamburg and demonstrated the much improved resolution of WGS, in comparison to traditional typing methods. The higher discriminatory ability of WGS has also been applied to a number of different bTB systems [24, 26, 25]. In the UK, WGS data has enabled genetic comparisons at an unprecedented scale, distinguishing herds, though not usually down to the individual ani-

mal level [24, 25]. Glaser *et al*. [26] recently applied WGS to investigate a recent outbreak in Minnesota, and revealed an unexpectedly distant source and demonstrated transmission between the resident cattle and wildlife populations. The demonstrated utility of WGS for epidemiological investigations of tuberculosis now mean that there is a drive to incorporate these methods into routine bTB surveillance.

It is the recent advent of Next Generation Sequencing (NGS) technologies, offering a cheap and fast alternative to the previously popular Sanger sequencing approaches [108, 107], that enables WGS to be a feasible epidemiological tool for the surveillance of bTB systems. In Sanger sequencing, fragments of DNA are repeatedly replicated in the presence of terminating nucleotides [108]. The random binding of terminating nucleotides results in many fragments of varying sizes, which can be separated by gel electrophoresis and used to deduce the nucleotide sequence of the original fragment. In contrast, NGS methods don't use terminating nucleotides but instead use fluorescently tagged nucleotides [109]. The binding of these tagged nucleotides results in a detectable nucleotide-specific reaction and doesn't disrupt the DNA replication process. This type of sequencing is known as sequencing-by-synthesis and NGS technologies massively parallerise it to be able to sequence thousands of short DNA fragments simultaneously. Although NGS methods are considerably faster than the previous Sanger techniques they are error prone.

NGS methods produce thousands of short reads corresponding to the fragmented DNA that is repeatedly sequenced. During the sequencing, nucleotides are bound in consecutive cycles. If no or multiple nucleotides are bound in a single cycle, de-phasing occurs [108]. De-phasing makes it difficult to detect the flourescent signal from nucleotides being bound and can produce errors where the nucleotide is incorrectly called. These errors, especially for a slowly mutating pathogen such as *M. bovis*, can have important consequences and must be considered during the handling of the vast quantities of data produced [109]. Quality metrics, describing the number and quality of reads at each position on the genome, are used in downstream analyses in an attempt to remove as many sequencing errors as possible, whilst ensuring that correct information is retained [176].

The quality metrics that have been used to process WGS *M. bovis* data in the published literature are varied (Table 2.1). The selection of metrics and their thresholds is data-specific and done to ensure the most accurate data is used. This is informed by the literature, a good overview of which is provided by Bishop [177]. The research described here investigated whether the selection of metrics and their criteria could additionally be informed by epidemiological data.

Epidemiological concordance is often used as a means of evaluating the utility of molecular typing methods [178, 179, 180, 181, 182]. Epidemiological concordance defines the probability that epidemiologically related pathogen strains would be correctly grouped to-

Table 2.1: The different quality filtering thresholds used by the published literature using Whole Genome Sequencing to investigate *Mycobacterium tuberculosis*, and *M. bovis* systems.

	Read Depth	High Quality Base Depth	Mapping Quality	Proportion Reads Supporting Allele Called	Proximity between Variant Positions	Quality Score	Species
Ford *et al.* [114]	5		20				*M. tuberculosis*
Biek *et al.* [24]			60	0.95			*M. bovis*
Walker *et al.* [116]	5	1		0.75	12		*M. tuberculosis*
Bryant *et al.* [117]			30	0.75		50	*M. tuberculosis*
Perez Lago *et al.* [119]	10		20				*M. tuberculosis*
Roetzer *et al.* [118]	10			0.80			*M. tuberculosis*
Glaser *et al.* [26]						150	*M. bovis*
Trewby *et al.* [25]	25-50	2-6	35-40	0.95	200		*M. bovis*

gether using a molecular typing tool. This concept relies upon the assumption that the genetic relatedness of isolates will reflect their epidemiological relationships. *M. bovis* is a clonal bacteria [100], and as such the genetic and epidemiological data should have a good agreement. Research on *Mycobacterium* systems has shown that inter-isolate relationships are highly related to epidemiological associations at a range of scales [116, 24, 26]. For WGS data sampled from three different bTB systems, a range of metrics and thresholds were evaluated within an automated framework, and the genetic data that resulted was examined in the context of the available epidemiological data. This automated procedure was designed to find the quality filtering which resulted in genetic data with a high level of epidemiological concordance, and thereby inform the selection of appropriate quality metrics and their thresholds.

2.2 Materials and Methods

2.2.1 Sampling and Isolate Preparation

Three different bTB systems were sampled, in Northern Ireland, New Zealand, and the south-west of England (Figure 2.1). Sets of genetically similar WGS data from isolates were available from these systems as well as sampling information and population data (Table 2.2). Genetically similar isolates were selected using molecular typing data (VNTR, REA and Spoligotyping data for the isolates from Northern Ireland, New Zealand, and England, respectively) and by the clade structure of preliminary phylogenetic trees. Genetically similar isolates were selected for the current analyses, as the genetic relationships between these isolates were more likely to reflect the epidemiological relationships described by the data available for each system.

Table 2.2: The final number, and temporal range, of isolates from the different bTB systems selected for this analysis. In addition, the type of sampling information and population data available is stated. Note that for the Woodchester Park system some isolates came from the same badger.

	N. Isolates	N. Cattle Isolates	N. Wildlife Isolates	N. Individuals Sampled	Time Range	Spatial Data	Temporal Data	Network Data
Northern Ireland	72	71	1	72	1998-2011	YES	YES	YES
New Zealand	72	41	31	72	1991-2009	YES	YES	NO
Woodchester Park	106		106	60	2000-2010	YES	YES	YES

2.2.1.1 Northern Ireland Isolates

A subset of the *M. bovis* isolates from cattle samples used in Trewby *et al.* [25] were available for further analyses. For these isolates, *M. bovis* isolation and confirmation of granuloma tissue samples, taken from suspect cattle, were completed at the Agri-Food and Biosciences Institute in Northern Ireland (AFBNI) using standard protocols. The archived isolates were grown on Löwenstein-Jensen medium (LJ) slopes to form single colonies. DNA was extracted from these single colonies using standard high salt hexadecyl trimethyl ammonium bromide (CTAB) and solvent extraction protocols [183, 91]. The cattle isolates selected were all Variable Number Tandem Repeat (VNTR) type 10 and were selected to create a representative of a recent single outbreak in Northern Ireland. 72 DNA extracts, originating from 71 individual cattle and 1 badger, were sequenced using an Illumina Genome Analyser IIx that produced paired end (2 x 70bp) reads. The sequencing was completed at the University of Glasgow's Sir Henry Wellcome Functional Genomics Facility. For each cow sampled, additional anonymised sampling information were available. These data described the year of sampling and which herds were sampled. In addition, any cattle movements that involved the sampled herds between 1998 and 2011 in the northeast region of Northern Ireland (centred on Newtownards) were available.

Figure 2.1: The sampling locations for the isolates from the bTB systems in Northern Ireland, New Zealand, and the southwest of England. Locations have been jittered to ease interpretation.

2.2.1.2 New Zealand Isolates

M. bovis isolates taken from infected cattle and wildlife across New Zealand's South Island were available as a result of the on-going routine surveillance of bTB in cattle and wildlife populations. During surveillance any lesions discovered during post-mortem examinations are investigated using conventional tests (described in [184]). All cultures are frozen and stored in the strain archive at AgResearch Ltd. The selected isolates were defrosted and re-cultured to generate DNA extracts for WGS. Isolates were selected based upon their Restriction Endonuclease Analysis (REA) type, species, and geographical location. The selection aimed to produce a set of genetically similar isolates from different species, which lived within the same geographical region. 72 DNA extracts (sourced from 41 cattle, 17 ferrets, 8 possums, 5 pigs, and 1 stoat) were sequenced at the University of Glasgow Polyomics Facility using an Illumina MiSeq platform. Paired end (2 x 300bp) reads were produced for each isolate.

For each of the WGS *M. bovis* isolates sourced from infected cattle and wildlife in New Zealand the year of sampling, species sampled, approximate latitude and longitude, parish and region where the sampling took place, and the REA type for the isolate were available. The REA method was used to type the isolates prior to their sequencing according to a previously described method [90, 94]. This method uses three different restriction enzymes to fragment DNA, and the resulting fragment patterns were used to define types [92].

2.2.1.3 Woodchester Park, Southwest England

Since 1977 the high density badger population present in Woodchester Park has been routinely monitored [185]. Woodchester Park, a densely wooded valley, is found in the southwest of England. The badger population is naturally infected with *M. bovis* [186]. Each season, badgers are trapped within annually delimited badger group territories. Each badger captured is given a unique tattoo. Upon capture, samples of urine, feaces, and tracheal and pharyngeal aspirates are taken. In addition, if any wounds or abscesses are present, swabs are taken. If a badger is found dead, a post-mortem examination is completed and samples are taken from the lymph nodes and major organs [185]. Any *M. bovis* that is successfully cultured from samples taken from captured badgers is archived at the Food and Environment Research Agency (FERA) in York. In addition, the blood samples are used in the following tests: γ-IFN test [42], Brock Test ELISA (Enzyme-Linked Immunosorbent Assay) [187], and the Stat-Pak test [188]. In the current research, a positive reaction to any of these tests, or a successful *M. bovis* culture was taken as an indication of infection.

The predominant spoligotype of *M. bovis* circulating in the Woodchester Park's badger population is 17. One hundred and six isolates of this spoligotype were selected for the current analysis. These 106 isolates were taken from the archive and were re-cultured at AFBNI. DNA extraction was also carried out at the AFBNI, and the same standard protocols were

used as for the Northern Ireland isolates. DNA isolates were sequenced using an Illumina MiSeq platform at the University of Glasgow Polyomics Facility producing 2 x 300bp paired end reads. As a result of the longitudinal monitoring of the badger population in Woodchester Park, detailed data regarding the life histories of the sampled badgers were available to accompany the WGS data from the isolates. Each isolate was linked to its sampled badger via a unique tattoo.

2.2.2 Processing Sequence Data

A standard pipeline was developed to handle the sequencing data resulting from the different Illumina platforms used to sequence the genomes of the *M. bovis* isolates:

1. Examination of each isolate's raw sequencing reads (FASTQ files) using FASTQC (v0.11.2 [189]) to determine the extent of trimming necessary.
2. Trim adapter sequences, if present, using TRIMGALORE (v0.4.1 [190]).
3. Trim poor quality ends of raw sequencing reads using PRINSEQ [191].
4. Alignment to the *M. bovis* reference genome, AF2122/97 [102] using the Burrow-Wheeler Alignment tool [192].
5. Use SAMTOOLS [176] to call Variants and store their information in a VCF (Variant Calling Format) file.

Once the raw read processing for each isolate was complete, the information was collated within the Northern Ireland, New Zealand and Woodchester Park datasets. Any sites that fell within regions that included the Proline-Glutamate (PE) and Proline-Proline-Glutamate (PPE) genes or annotated repeat regions were removed (see [193] for motivations). All the isolates used in the current analysis had $\geq$20 reads mapped to $\geq$90% of the *M. bovis* genome. Only sites on the *M. bovis* genome that varied amongst the isolates (variant positions) were used in the analyses to follow.

2.2.3 Filter Sensitivity Analysis

The following steps summarise the filter sensitivity analysis completed for each of the three sets of isolates:

1. Variant positions, on the *M. bovis* genome, were filtered based on a defined set of filters.
2. The inter-isolate genetic distances were calculated using the filtered data available for the selected variant positions.
3. The epidemiological metrics were fitted to the inter-isolate genetic distances using a Random Forest model in R (v3.2.1 [194]).

2.2.3.1 1. Filtering Variant Positions

To select variant positions, sites that show variation amongst the *M. bovis* isolates, quality thresholds were used. The thresholds were based upon the available quality information (Read Depth, High Quality Base Depth and Mapping Quality) that summarised the sequenced reads, which were aligned at each particular position on the *M. bovis* genome for an isolate. These quality metrics were used to define the following quality filters:

- Read Depth (DP) - the number of sequenced reads that aligned to the current position on the reference genome.
- High Quality Base Depth (HQDP) - the number of high quality reads that aligned to the current position on the reference genome. A high quality base, on a read, was defined as a base that had a Phred quality score of more than 20 [195]. The Phred quality score describes the probability that the base was sequenced wrong. A Phred score of 20 equates to a probability of 0.01.
- Mapping Quality (MQ) - describes the probability that the nucleotide at the current position on the genome was mapped correctly. This metric was calculated by examining the number of mismatches each mapped read had over its extent.
- Allele Support (SUP) - the proportion of the reads mapped to the current position on the reference genome that carry the allele called. The allele called at a position was the nucleotide (Adenine, Cytosine, Guanine, or Thymine) that was supported by the majority of the mapped reads.
- Site Coverage across isolates (COV) - the proportion of ALL isolates that had sufficient coverage at each variant position on the genome.
- Site Proximity (PROX) - the number of positions that separated variant positions to be selected.

A set of threshold combinations were used to filter the variant positions (Table 2.3). These ranges were chosen to include the range of values used in the published literature that utilised WGS to investigate bovine and human TB systems. The values of Mapping Quality and Allele Support were held constant as preliminary work found that varying these filters had little to no effect.

2.2.3.2 2. Calculating Genetic Distances

Once a specific set of thresholds (Table 2.3) had been applied, on a given set of isolates, the isolates were compared to one another to generate an inter-isolate genetic distance distribution. Each genetic distance was calculated based upon the p-distance - the number of variant positions that differ between the isolates being compared.

Table 2.3: The range of thresholds explored for each of the quality filters used to filter the variant positions (sites on the genome that show variation amongst the isolates examined).

Metric	Min	Max	Increment By
Read Depth	10	50	5
High Quality Base Depth	0	10	2
Mapping Quality	30	30	-
Allele Support	0.95	0.95	-
Site Coverage	0.5	1	0.1
Site Proximity	0	100	10

2.2.3.3 3. Defining Epidemiological Metrics

Each generated inter-isolate genetic distance distribution was compared to a distribution of epidemiological metrics, which was produced by making the same pairwise isolate comparisons. In the current research, epidemiological metrics, described in Table 2.4, were defined to describe comparisons made between the isolates. These metrics were based upon the available sampling information and sampled population data. The epidemiological metrics were designed to capture the patterns of direct and indirect contact between the sampled animals. It was expected that animals that were close in space, time and within a contact network, would be more likely to carry similar strains of bTB.

The available data for each sampled bTB system differed (Table 2.2). For the isolates from New Zealand, only sampling information (species, parish, year, location (latitudes and longitudes), and REA type) were available. These data were used to define epidemiological metrics that describe the spatial and temporal relationships between the isolates. In contrast, for the isolates from Northern Ireland as well as similar sampling information, the recorded movements of cattle for the sampled herds were available. Using these data, spatial, temporal and network based epidemiological metrics were defined to explain the inter-isolate relationships. For the isolates from Woodchester Park, the detailed badger capture database provided information about the population dynamics of the sampled high density badger population. Each isolate could be linked to its sampled badger, which were compared via spatial, temporal and network based epidemiological metrics.

2.2.3.4 Conducting Sensitivity Analysis

The filter sensitivity analysis aimed to use the changing agreement between the genetic and epidemiological data to inform the selection of quality filter thresholds. If certain combinations of thresholds resulted in a poor agreement between the genetic and epidemiological data, they should be avoided. This analysis assumes that genetic data that agrees well with the sampled system's epidemiology is true. Using the ranges of filter thresholds described in Table 2.3, 3564 unique combinations were defined. Each of the combinations was examined

Epidemiological metric	**Northern Ireland**	*New Zealand*	Woodchester Park
Isolates taken from the same ***species***/badger (yes/no)	SPECIES	SPECIES	BADGER
Isolates have the same REA type (yes/no)	NA	USED	NA
Isolates taken from cattle infected in the same outbreak (yes/no)	USED	NA	NA
Isolates taken from same *parish*/*region* or sampled/main/infected group (yes/no)	NA	PARISH/REGION	SAMPLED/MAIN/INFECTED GROUP
Spatial distance (km) between *sampling*/**sampled**/main/infected *locations*/**herds**/groups	SAMPLED HERDS	SAMPLING LOCATIONS	SAMPLED/MAIN/INFECTED GROUPS
The number of ***years***/days between the sampling dates of the isolates compared	YEARS	YEARS	DAYS
The number of days between the infection detection dates of the isolates compared	NA	NA	USED
The number of days that the sampled animals spent together in the same **herd**/group	HERD	NA	GROUP
The number of days overlap between the recorded lifespans of the sampled animals	USED	NA	USED
The number of days overlap between the infected lifespans of the sampled badgers	NA	NA	USED
The number of recorded animal movements between the **sampled**/main/infected **herds**/groups of sampled animals	SAMPLED HERDS	NA	SAMPLED/MAIN/INFECTED GROUPS
The number of badgers captured in both the sampled/main/infected groups of the sampled badgers	NA	NA	SAMPLED/MAIN/INFECTED
Shortest path length between the **sampled**/main/infected **herds**/groups of the sampled animals	SAMPLED HERDS	NA	SAMPLED/MAIN/INFECTED GROUPS
Mean number of animals dispersing along the edges of the shortest path between the **sampled**/main/infected **herds**/groups	SAMPLED HERDS	NA	SAMPLED/MAIN/INFECTED GROUPS

Table 2.4: The epidemiological metrics calculated based upon comparisons between the isolates available from Northern Ireland, New Zealand and Woodchester Park. Where it was necessary to refer to these different datasets, the text was in **bold**, *italics*, or underlined for the respective datasets. In each dataset's column, where necessary, the words referred to in the metric column, specific to the dataset were noted. The sampled group/herd was defined as the group/herd that the animal was sampled in. The main group was defined as the group that the badger spent the majority of it's recorded life in. The infected group was defined as the group that the badger was captured in when it's infection was first detected. The shortest path, calculated on an adjacency matrix that recorded the number of movements between herds/groups, was calculated using Dijkstra's algorithm [196]. Metrics that were based on spatial, temporal, or network based data were coloured red, gold, or blue, respectively.

in turn. Inter-isolate genetic distances were generated by using the filtered variant positions based on each combination of filters. Each resultant inter-isolate genetic distance distribution was compared to epidemiological metrics that were calculated for each respective pair of isolates. This comparison was done by fitting the epidemiological metrics to the corresponding genetic distances using a Random Forest regression model [197]. The fitting of the Random Forest model for each set of filters was completed in the statistical programming environment R (v3.2.1 [194]).

The Random Forest algorithm builds a series of decision trees based upon the predictor variables (epidemiological metrics) and a response variable (the genetic distances). For a regression model each tree was considered a regression tree - a means of using the predictor variables to predict a single response value [197]. Each regression tree was built based upon a random subset ($\sim$66%) of the input data. Nodes in the tree used a random selection of the predictor variables to split the input data into two groups. Splitting at nodes continued until a terminal node was reached. A terminal node provided an estimate of the response variable. This estimate represented the mean of the response values that remained, following the split. The predictive accuracy of each tree was tested on the remaining ($\sim$33% - the Out Of Bag (OOB)) data. A forest of independent regression trees was built. The Random Forest model aggregated the regression trees into a single predictive model.

To fit the Random Forest model, two parameters were necessary, one defining the number of trees to be built and another that specified the number of predictor variables (mTry), which were randomly drawn and used to define the nodes in each regression tree (Table 2.5). The parameters used, were estimated using the tuneRF function (available in the randomForest package in R [197]). Only 500 trees (instead of 1000) were built in the Random Forest models fitted on the genetic and epidemiological data for Woodchester Park to increase computational efficiency for analysis of this larger dataset.

Table 2.5: The input parameter settings for fitting the Random Forest models during the filter sensitivity analysis. The number of predictor variables (epidemiological metrics) randomly drawn and used to build the nodes of each regression tree was defined by mTry. The Number Trees parameter specified the number of regression trees to be built.

Isolate Set	mTry	Number Trees
Northern Ireland	6	1000
New Zealand	4	1000
Woodchester Park	13	500

2.3 Results

The isolates available for the current research were sourced from three bTB systems. The samples were collected, cultured and isolated separately. In addition, they were sequenced independently on different Illumina platforms. The quality of the data available for these different systems varied (Figure 2.2). The isolates from New Zealand were of the highest quality, with a high number of reads covering, on average, >99% of the *M. bovis* genome.

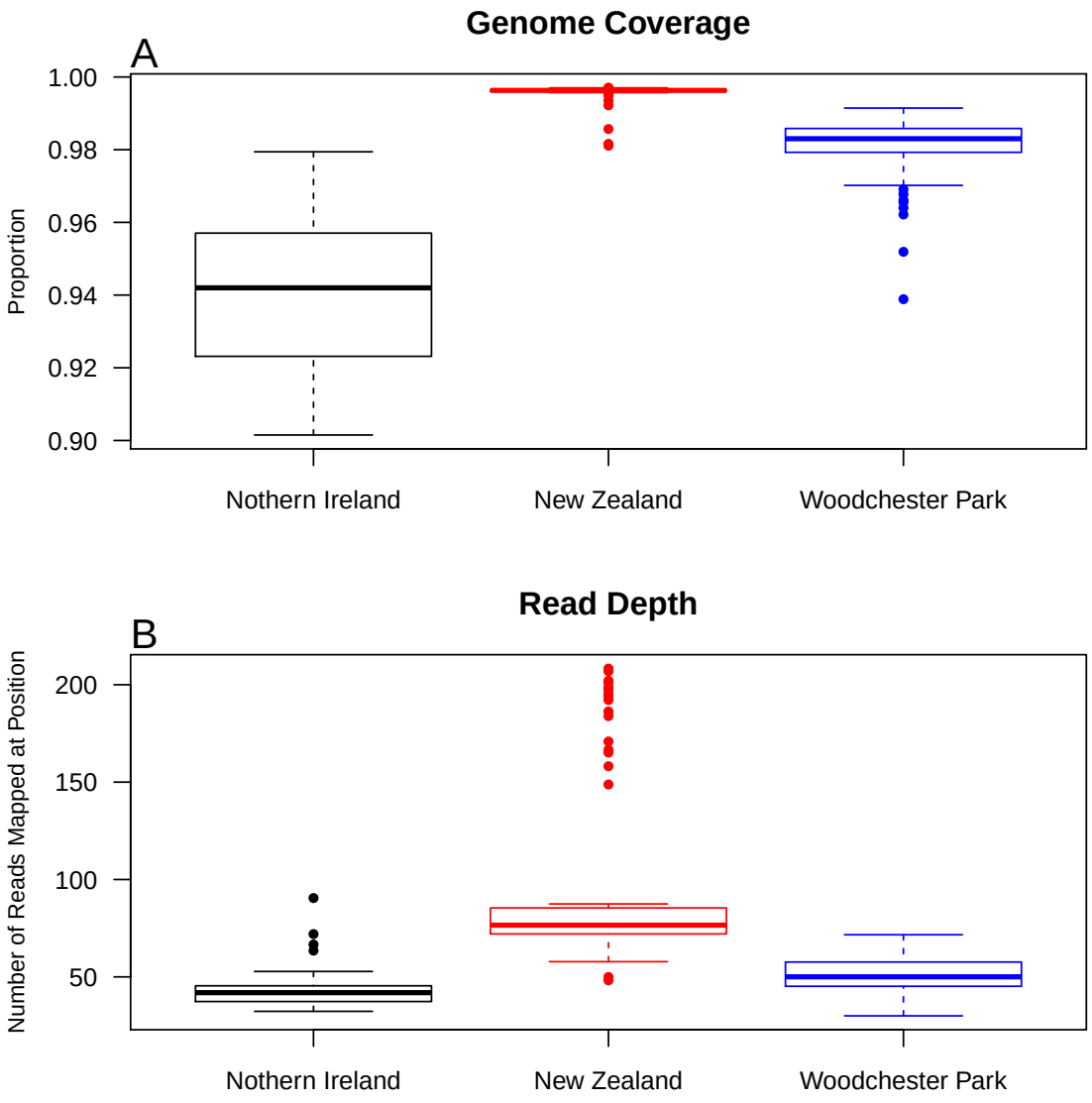

Figure 2.2: The quality, as measured by the genome coverage (A) and the average read depth (B) of the isolates in the different datasets from Northern Ireland (black), New Zealand (red) and Woodchester Park (blue).

The amount of variation in the inter-isolate genetic distances explained by the fitted Random Forest model differed between the datasets analysed (Figure 2.3). The fitted Random Forest models based on the isolates from Woodchester Park were able to explain the most variation (approximately 60%). Varying the High Quality Base Depth or the Site Proximity

filter thresholds had little effect on the variation explained by fitted Random Forest model. In contrast, using values above 30 for Read Depth or above 0.8 for Site Coverage could result in a considerably lower amount of variation explained by the Random Forest models. The variation explained by the fitted Random Forest models based on the WGS data from Northern Ireland, although lower (around 45%), were similarly sensitive to changing the thresholds of the Read Depth and Site Coverage quality filters. The variation explained by the Random Forest models fitted to the New Zealand isolates showed the least amount of variation in response to the changing filter thresholds, especially with lower values.

Figure 2.4 shows how the variance in the variation explained by the Random Forest models change in response to the changing filter thresholds. For the Read Depth and Site Coverage filters, as the filter thresholds are increased the variance in the variation explained by the Random Forest model increases. In contrast, the variance in the variation explained by the Random Forest model is relatively unaffected by changing the thresholds of the High Quality Base Depth and Site Proximity filters.

The filter thresholds used dictated which of the available variant positions were selected to calculate the inter-isolate genetic distances. The extent of this selection was controlled, in part, by the Site Coverage filter threshold. The higher the Site Coverage threshold, the more isolates that were needed to have sufficient coverage (according to the other filters used) at each variant position for it to be selected. In contrast to the variation explained by the fitted Random Forest model, the proportion of sites that were retained was affected by changing any of the filters (Figure 2.5). The most rapid declines in the proportion of sites retained were created by increasing the Read Depth or Site Coverage thresholds. Increasing the thresholds of the High Quality Base Depth or the Site Proximity filters resulted in a more gradual decline in the proportion of sites retained. Similarly to the variation explained, the proportion of sites retained varied least when the different filters were applied to the isolates sourced from New Zealand.

By examining how the proportion of sites that were retained dropped in response to increasing the thresholds of the quality filters, it was possible to define limits for the quality filter thresholds. When the median value, shown by the dark lines in Figure 2.5, dropped below 50% of the proportion of sites retained, a limit was set. Only values below this limit were considered for each of the quality filters. For example, a limit of 40 was set for the Read Depth quality filter for the Woodchester Park dataset. When the Read Depth was above 40, more than half of the filter combinations that used a Read Depth of 40 resulted in less than half of the variant positions having sufficient coverage to be used. Setting the median threshold of 50% was designed to limit the extent of over-fitting, which could result when too many of the variant positions were removed.

Once limits were set for each of the quality filters' thresholds, the Random Forest models that

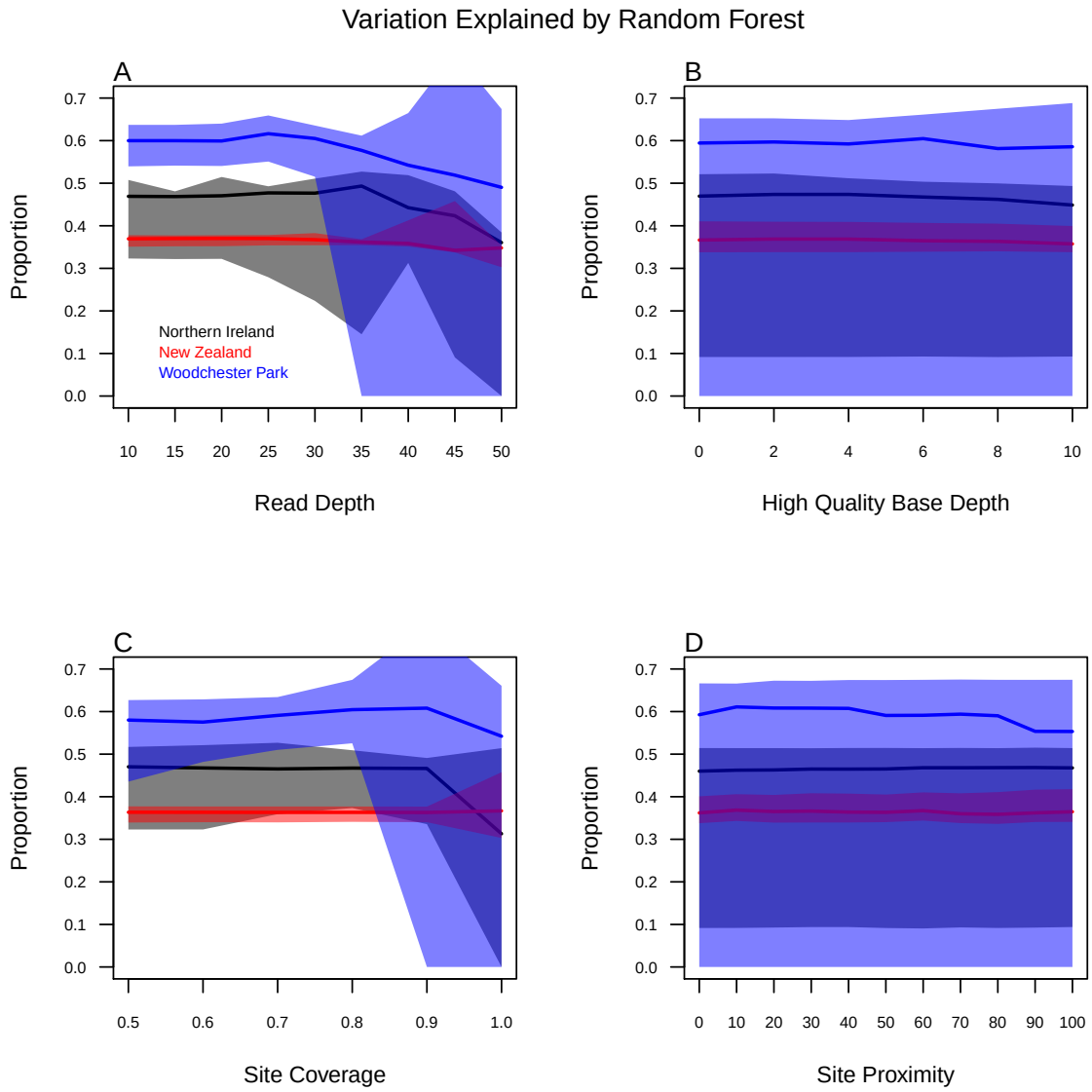

Figure 2.3: The reported pseudo R squared value from a fitted Random Forest model of the inter-isolate genetic distances against their respective epidemiological metrics. The pseudo R Squared value is plotted for each value of the Read Depth (A), High Quality Base Depth (B), Site Coverage (C), and Site Proximity (D) filters whilst varying the thresholds of all the other filters (described in Table 2.3). The results of the analyses based on the different sets of isolates (Northern Ireland (black), New Zealand (red), and Woodchester Park (blue)) are plotted with different colours. A dark line is plotted through the median values, this line is surrounded by a shaded area between the lower 2.5% and upper 97.5% percentiles. The x-axis tick values represent the thresholds used for each quality filter.

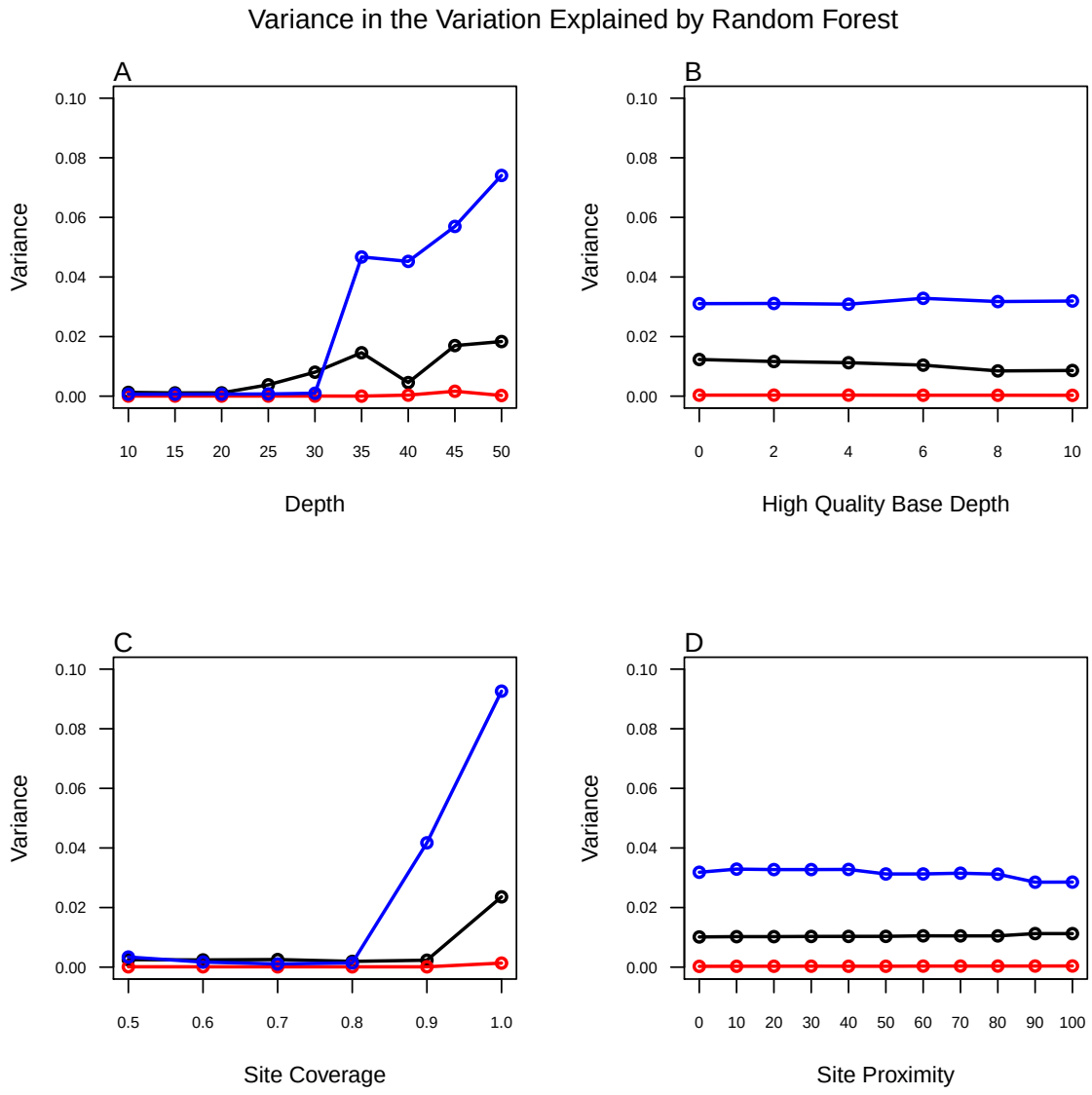

Figure 2.4: Variance in the reported pseudo R squared value from a fitted Random Forest model of the inter-isolate genetic distances against their respective epidemiological metrics. The variance value is plotted for each value of the Read Depth (A), High Quality Base Depth (B), Site Coverage (C), and Site Proximity (D) filters whilst varying the thresholds of all the other filters (described in Table 2.3). The results of the analyses based on the different sets of isolates (Northern Ireland (black), New Zealand (red), and Woodchester Park (blue)) are plotted with different colours. The x-axis tick values represent the thresholds used for each quality filter.

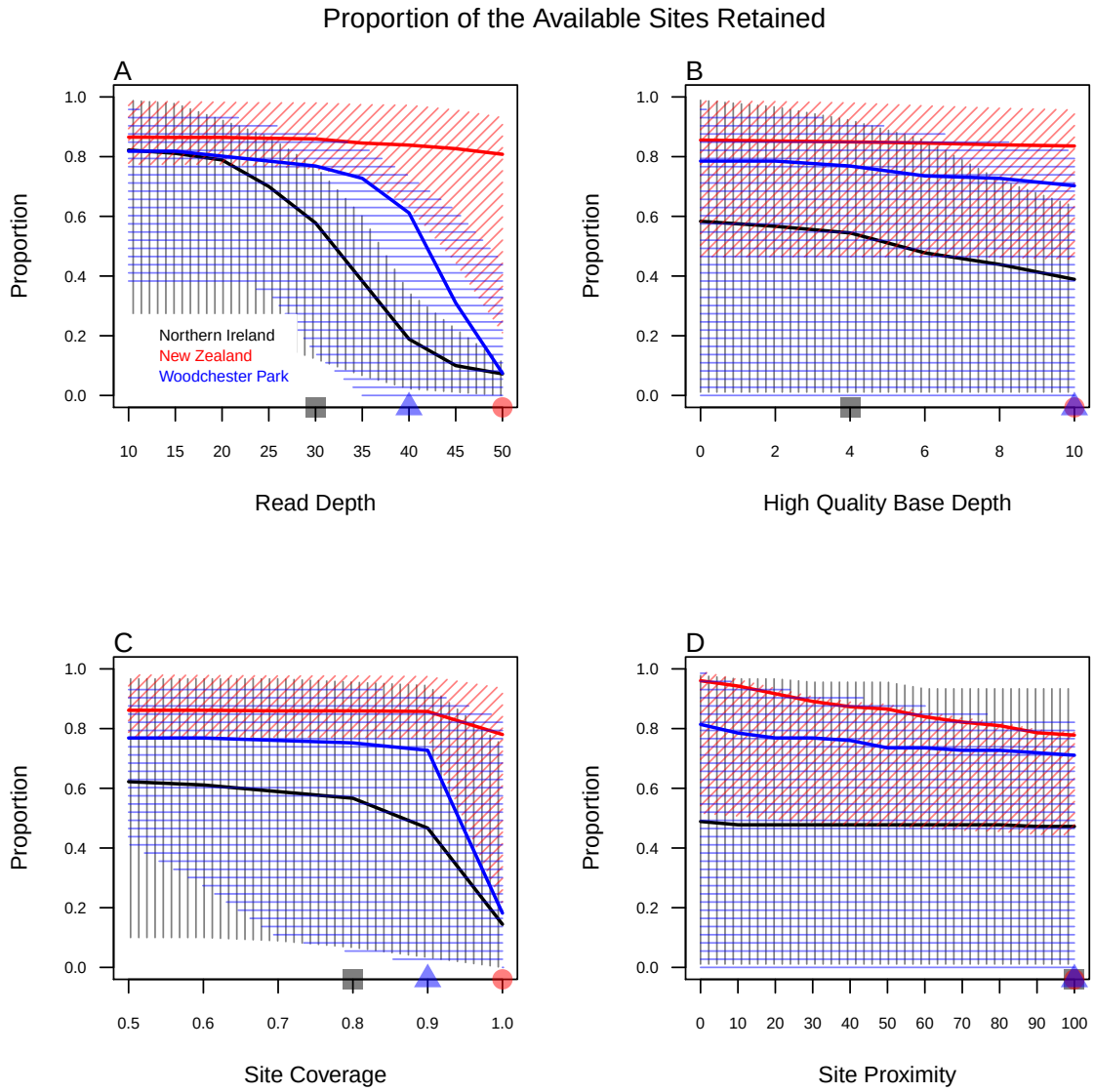

Figure 2.5: The changing proportion of retained variant positions (positions on the genome that vary amongst the isolates) in response to the different thresholds used for the quality filters. The proportion is plotted for each value of the Read Depth (A), High Quality Base Depth (B), Site Coverage (C), and Site Proximity (D) filters for the thresholds of all the other filters (described in Table 2.3). The results of the analyses based on the different sets of isolates (Northern Ireland (black), New Zealand (red), and Woodchester Park (blue)) are plotted with different colours. A dark line is plotted through the median values, this line is surrounded by a cross-hatched area between the lower 2.5% and upper 97.5% percentiles. The angle of the cross-hatching is different for the results based on the datasets from Northern Ireland (vertical), New Zealand (diagonal), and Woodchester Park (horizontal). Points on the x-axis indicate threshold values that were set and used to determine which filter thresholds should be used (black square (Northern Ireland), blue triangle (New Zealand), and red circle (New Zealand)).

were fitted to the remaining filter combinations were examined, to determine which threshold combinations produced genetic data that best agreed with the available epidemiology. The ranges of thresholds for each filter, that produced the top 5% of the fitted Random Forest models, were plotted in Figure 2.6. These ranges were different for each dataset, but the ranges generally overlapped. The ranges were the broadest for the Site Proximity and High Quality Base Depth filters.

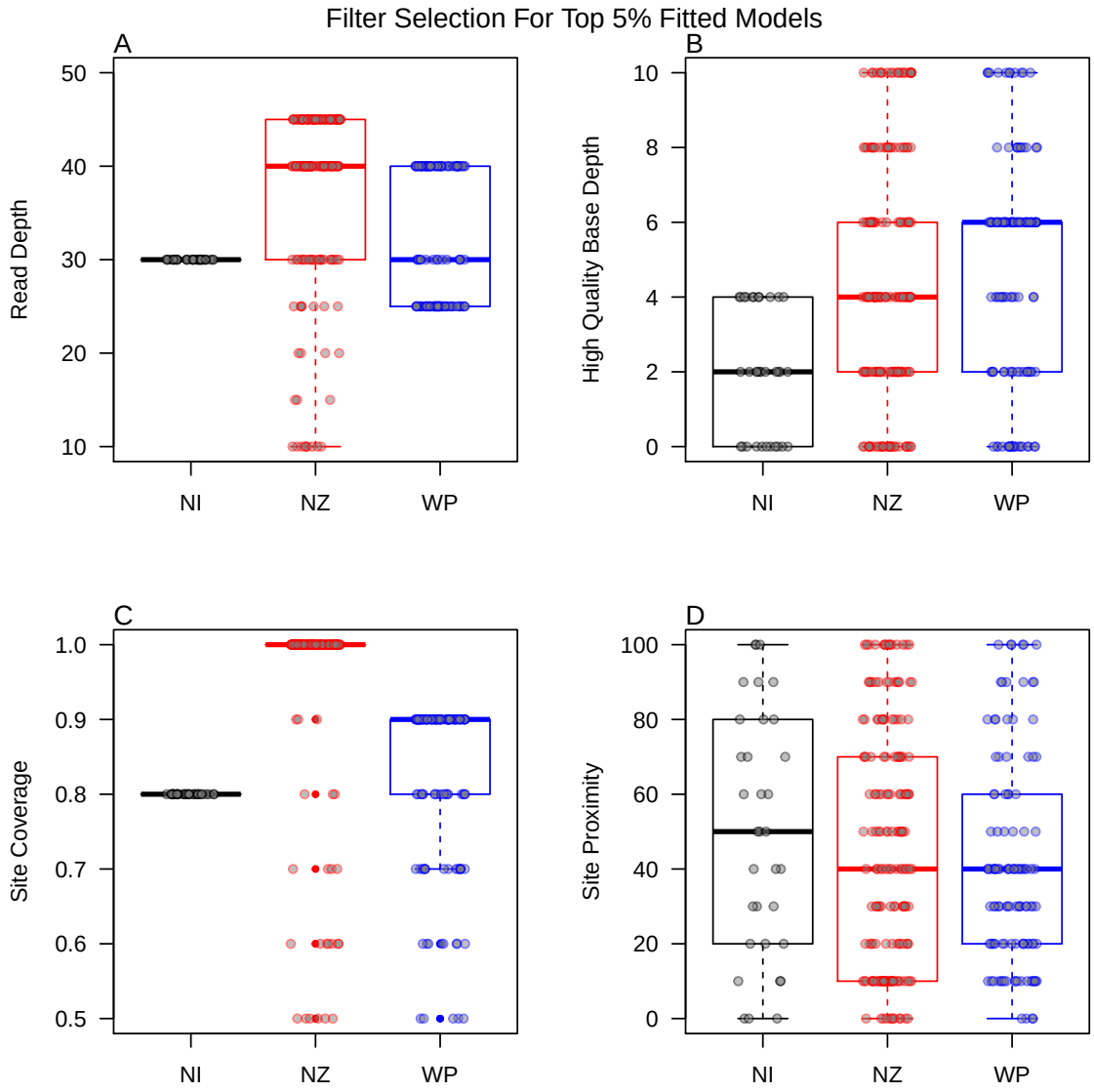

Figure 2.6: The range of threshold values that produced the best fitted Random Forest models, which explained the most variation in the inter-isolate genetic distances (the top 5% of the models). The ranges are shown for each of the different filters: Read Depth (A), High Quality Base Depth (B), Site Coverage (C), and Site Proximity (D). For each filter, its range is shown for the analyses based on the different datasets: Northern Ireland (black), New Zealand (red), and Woodchester Park (blue). The points used to create the respective boxplots are overlaid and jittered along the x-axis.

The fitted Random Forest models were able to explain variation in the inter-isolate genetic

distances using their respective epidemiological metrics (Figure 2.3). Aspects of the epidemiology must, therefore, explain some of the genetic variation observed between isolates. Depending on the information available for the sampling events and the animals sampled, the number of epidemiological metrics available varied (Figure 2.7, Figure 2.9, and Figure 2.11). The Random Forest algorithm measures the informativeness of each predictor variable by randomly permuting each one separately, and measuring the increased difference between the model's predicted inter-isolate genetic distances and the true values. Where spatial, temporal and network based metrics were available, they were all informative in explaining variation in the inter-isolate genetic distances.

When the epidemiological metrics were ranked by their informativeness, these rankings were sensitive to the thresholds used for each quality filter (Figure 2.8, Figure 2.10, and Figure 2.12). However, the changes in rank were mainly restricted to within the spatial, temporal and network based categories and, therefore, were likely the result of many of the epidemiological metrics being highly correlated. The Random Forest algorithm is robust to highly correlated variables, but interpretation of the epidemiological metric rankings should be considered in the light of their sensitivity to the filter thresholds used.

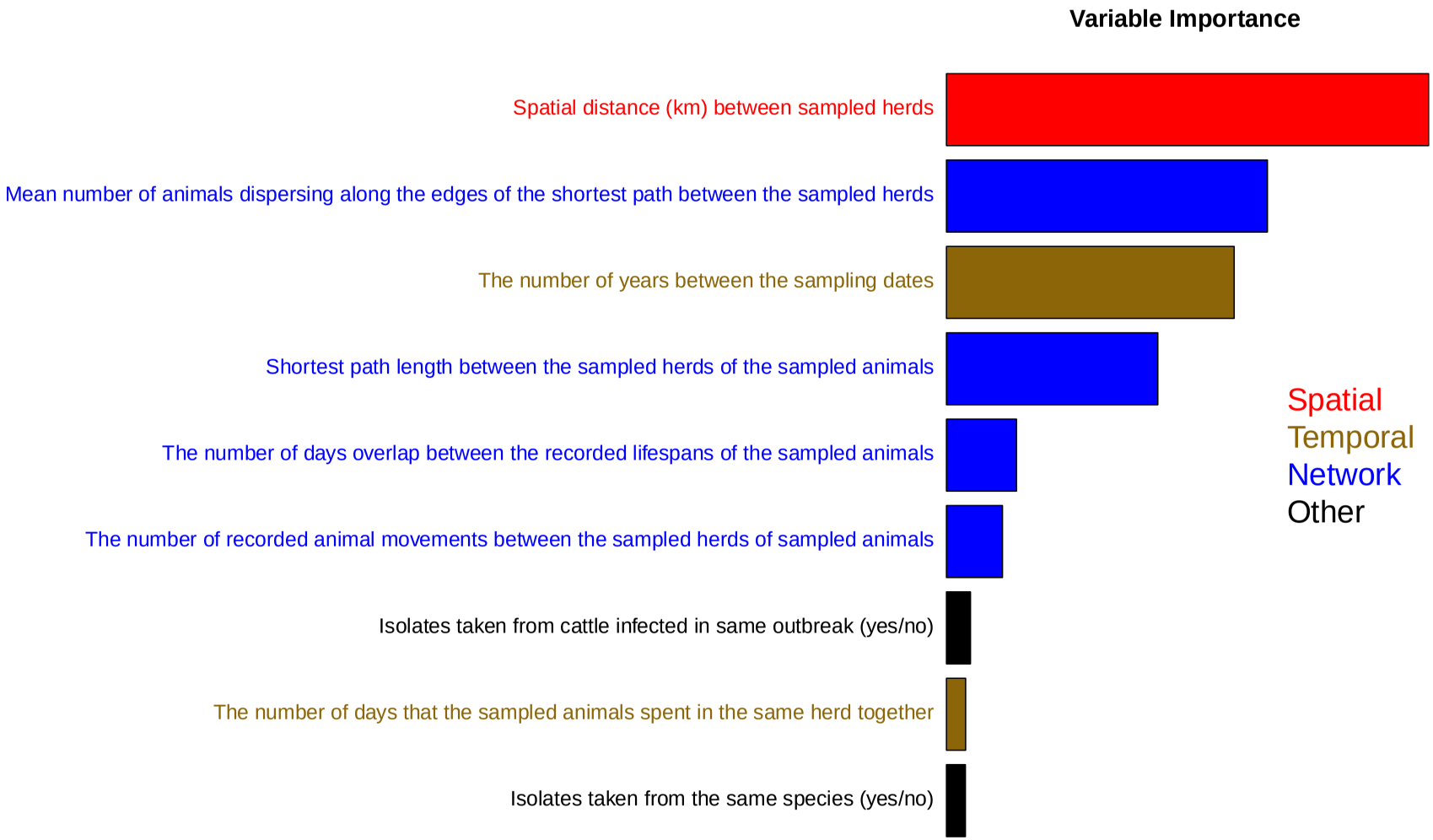

Figure 2.7: The importance of each of the epidemiological metrics used to fit the Random Forest model to the inter-isolate genetic distances for the isolates that were sourced in Northern Ireland.

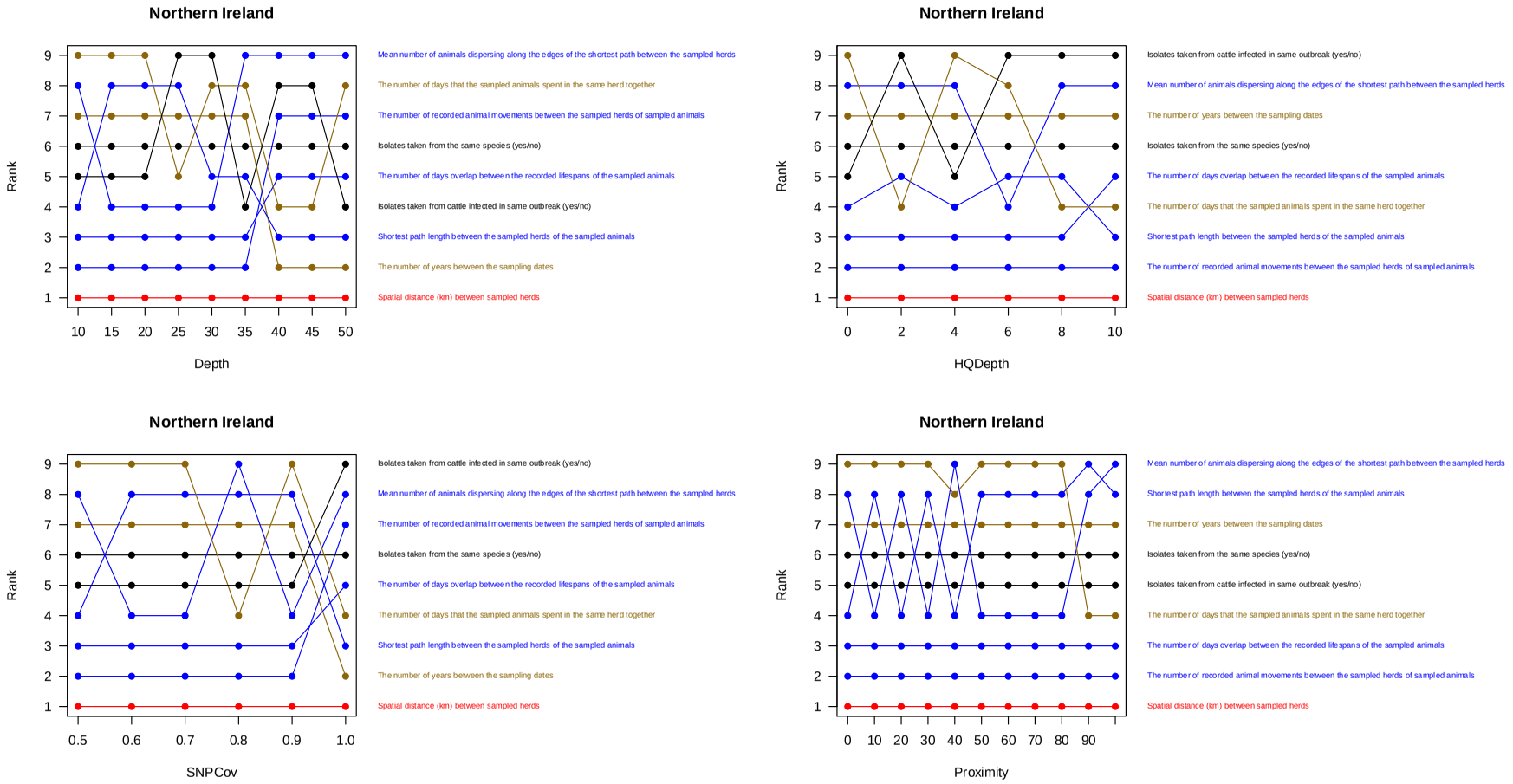

Figure 2.8: The influence of the changing the thersholds of quality filters on the rankings of the epidemiological metrics. The metrics were ranked by their informativeness in the Random Forest model. The ranking was taken for each value of the filter when the values for the other filters were set to their lowest threshold. The epidemiological metrics were coloured according to whether they related to spatial (red), temporal (gold) or network (blue) characteristics.

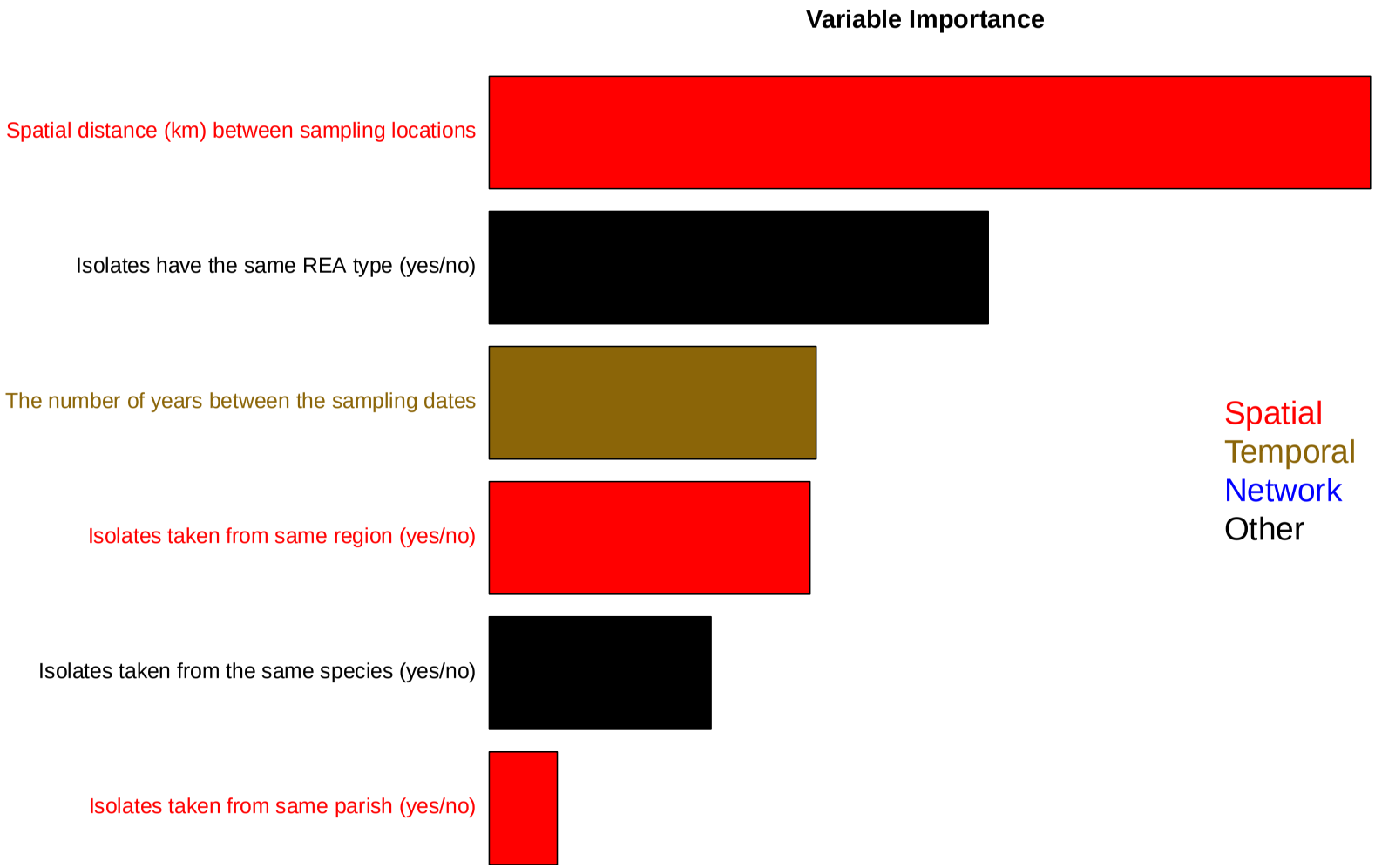

Figure 2.9: The importance of each of the epidemiological metrics used to fit the Random Forest model to the inter-isolate genetic distances for the isolates that were sourced in New Zealand.

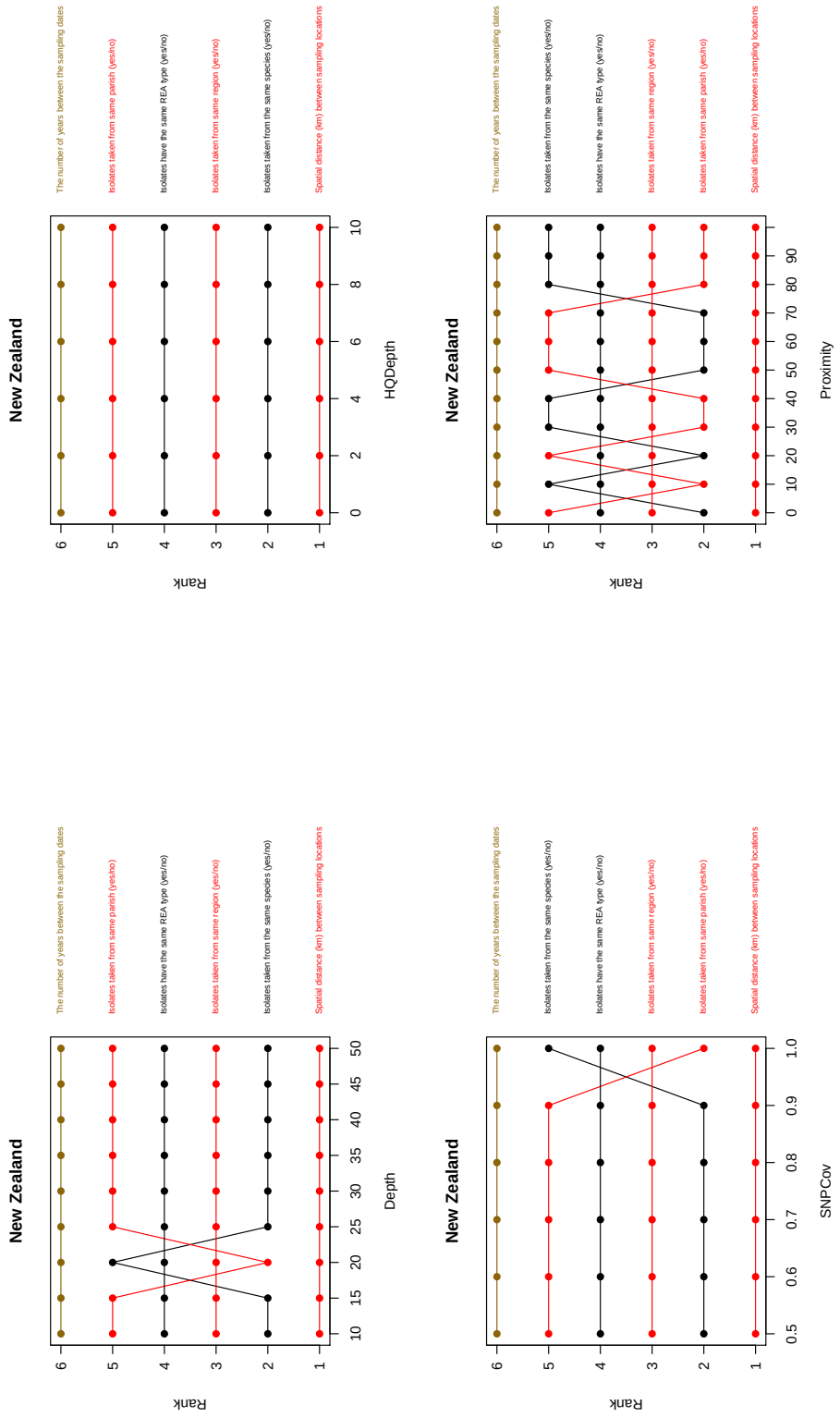

Figure 2.10: The influence of the changing the thersholds of quality filters on the rankings of the epidemiological metrics. The metrics were ranked by their informativeness in the Random Forest model. The ranking was taken for each value of the filter when the values for the other filters were set to their lowest threshold. The epidemiological metrics were coloured according to whether they related to spatial (red), temporal (gold) or network (blue) characteristics.

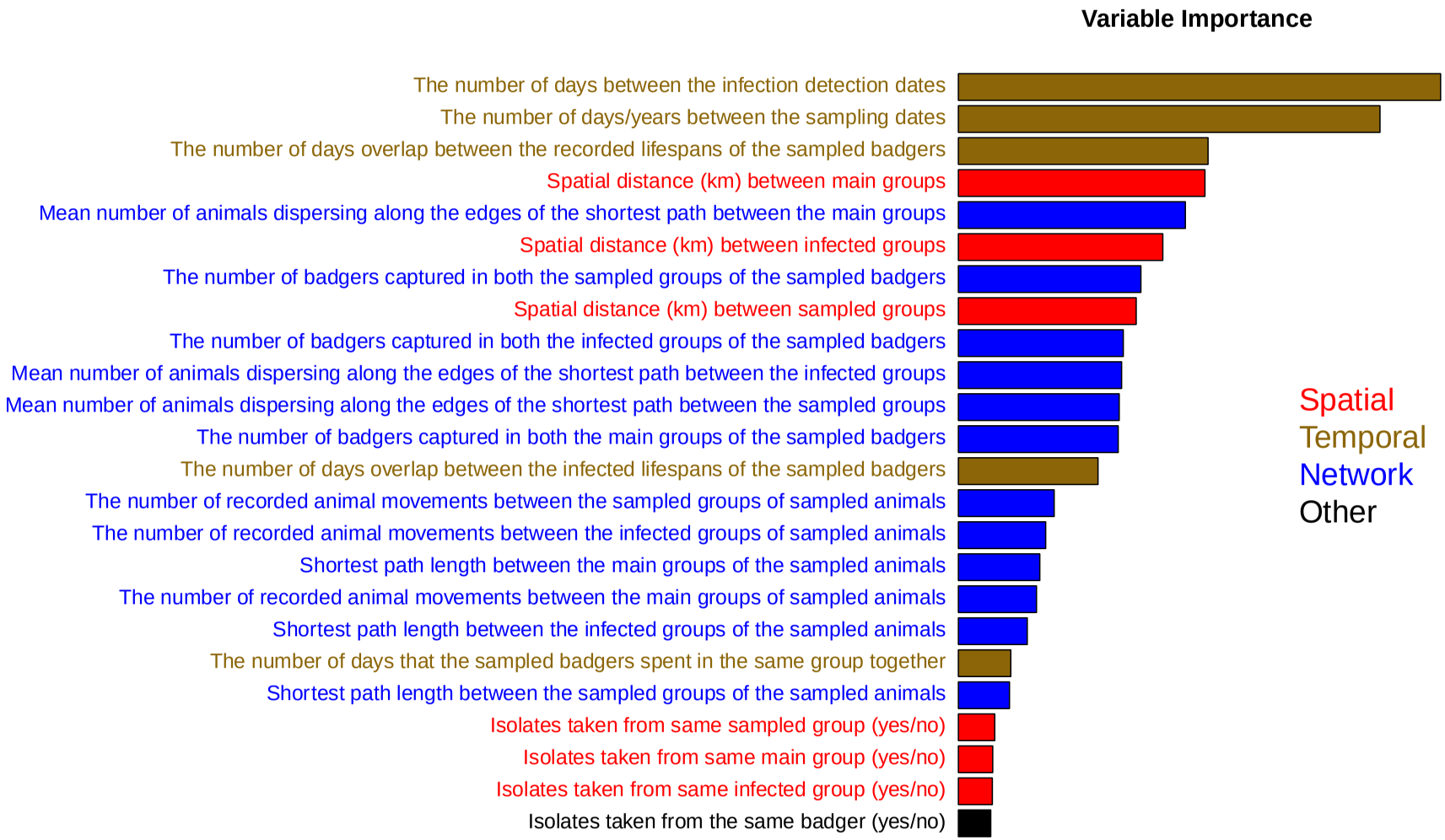

Figure 2.11: The importance of each of the epidemiological metrics used to fit the Random Forest model to the inter-isolate genetic distances for the isolates that were sourced in Woodchester Park.

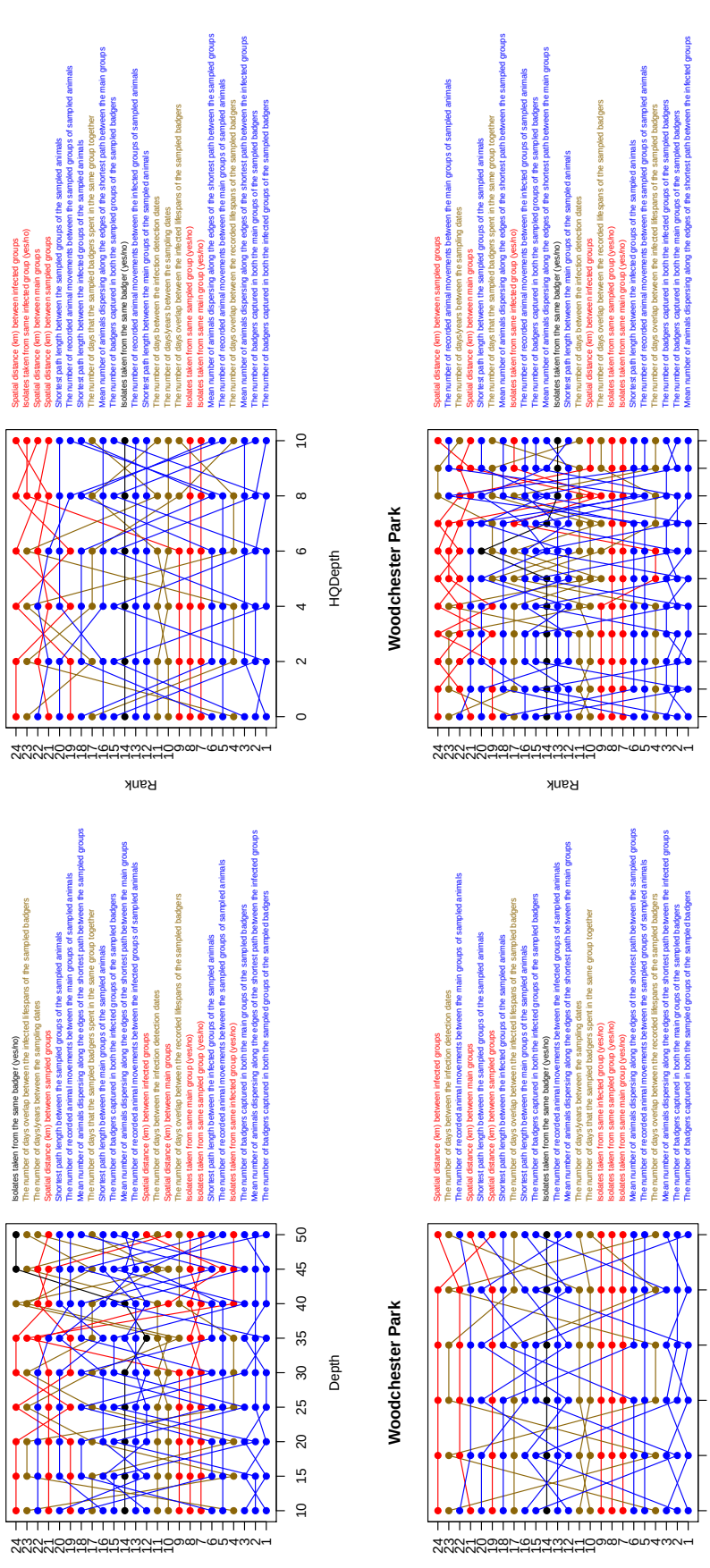

Figure 2.12: The influence of the changing the thersholds of quality filters on the rankings of the epidemiological metrics. The metrics were ranked by their informativeness in the Random Forest model. The ranking was taken for each value of the filter when the values for the other filters were set to their lowest threshold. The epidemiological metrics were coloured according to whether they related to spatial (red), temporal (gold) or network (blue) characteristics.

2.4 Discussion

In the published literature for *M. tuberculosis* and *M. bovis*, where WGS has been used, the quality filters chosen, and the thresholds used for those filters, has varied considerably (Table 2.1). Quality filtering is generally specific to the dataset being analysed as a result of variation in sequencing quality. The current research demonstrates that the process of selecting appropriate quality filters and their thresholds can be informed by epidemiological data assuming that the genetic and epidemiological data are related.

Three different bTB systems were sampled and the *M. bovis* isolates obtained were sequenced using WGS. The quality of these different datasets varied (Figure 2.2), likely as a result of the isolate culturing, DNA extraction, and sequencing being conducted in different places (although the same protocols were used). In addition, the scale of the sampling and population data available differed for each system (Table 2.2). Using the principles of epidemiological concordance, the available epidemiological data for each system provided a means of evaluating the effect of the selection of quality filters and their thresholds.

The Random Forest algorithm was used to evaluate epidemiological concordance. This algorithm was selected as it provides a highly flexible model fitting framework that is insensitive to highly correlated or uninformative predictor variables [197]. Where available, spatial, temporal, and network based epidemiological metrics were informative (Figure 2.7, Figure 2.9, & Figure 2.11). This informativeness demonstrates that the dynamics of *M. bovis* infection in the different sampled systems were captured, at least in part, by the epidemiological data available. The extent of the agreement between the genetic and epidemiological data, as measured by the proportion of the variation explained by the fitted Random Forest models, varied between the datasets (Figure 2.3). The differences in the variation explained for the three datasets, was the result of the different resolutions of sampling information and population data available (Figure 2.3). The WGS data from Woodchester Park, in the southwest of England, were accompanied by detailed life history data regarding the movements and bTB testing histories of any captured badgers within the sampled population. The epidemiological metrics, based on these data, were able to explain approximately 60% of the variation in the inter-isolate genetic distances. In contrast, for the WGS data from New Zealand isolates, for which there were only coarse sampling data available, the variation explained was much lower (approximately 35%).

The degree to which changing the thresholds of each quality filter affected the variation explained by the fitted Random Forest models, varied between the quality filters (Figure 2.3 & Figure 2.4). Different thresholds for the Site Proximity filter had little effect on the agreement between the genetic and epidemiological data. In contrast, different combinations of high values for the Read Depth and Site Coverage filters produced genetic data that was highly variable, in terms of how well it agreed with the epidemiological data (Figure 2.4).

This variability may have resulted from the higher thresholds removing a large proportion of the variant positions (Figure 2.5) and suggests that these extreme values should be avoided. The selection of different thresholds for the Read Depth and Site Coverage filters had more of an impact when the overall sequencing quality of the dataset was low (Figure 2.3). Therefore, even in a low quality dataset, true genetic data relating to the system's epidemiology existed, but were more likely to be removed when quality filter thresholds were increased.

A broad range of filter thresholds that resulted in the best Random Forest models, was observed for the New Zealand dataset (Figure 2.6). Therefore, for the high quality WGS *M. bovis* isolates from New Zealand, the coarse epidemiological data available wasn't very informative for the selection of quality filters or their thresholds. As Figure 2.5 demonstrates, this uninformativeness may be driven by a large proportion of the variant positions being retained across a broad range of quality filter thresholds. For the WGS data from the isolates from Northern Ireland, the detailed epidemiological information was particularly informative suggesting only single thresholds for the Read Depth and Site Coverage filters and a narrow range for the High Quality Base Depth filter. The detailed capture data was informative for selecting thresholds for the selection of Read Depth and Site Coverage filters for the WGS data from Woodchester Park.

The current research demonstrates that the selection of quality filters and their thresholds does impact the genetic data and its agreement with the sampled system's epidemiology. Therefore, where epidemiological information is available, it should be used to inform the selection of quality filters and their thresholds. If epidemiological data isn't available, as is often the case, careful consideration of the filters and thresholds is advised whilst taking into account the quality of the sequencing data available.

Chapter 3

Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand

3.1 Introduction

Control of a disease in a multi-host system is most efficient when the role of the different hosts is understood [123, 198]. Control of bovine tuberculosis (bTB) in domestic cattle herds is motivated by the zoonotic risk of the causative agent *Mycobacterium bovis*, its impacts on animal productivity, and the benefits of TB-free status in international trade [199]. *M. bovis* infection has been successfully combated in many countries [9, 4, 10]. Effective campaigns have relied upon test and slaughter regimes, movement restrictions and abattoir surveillance. Despite success using such regimes, endemic bTB still exists, most notably in areas that have wildlife reservoirs of infection. A broad host range, promoting multi-host bTB systems, is considered to be one means by which *M. bovis* persists in the face of control [13, 200].

In New Zealand, the introduced brushtail possum (Trichosurus vulpecula) has long been recognised as an important maintenance reservoir for *M. bovis* [201, 16]. In addition, deer, pigs, and ferrets are thought to act as key spatial and temporal vectors of infection [16]. Control of bTB in cattle herds uses test and slaughter surveillance; more frequent testing and movement control are employed in Vector Risk Areas (VRAs), where the risk of infection from wildlife is highest [60]. Within VRAs, control methods such as trapping and poisoning are primarily aimed at the possum population so as to limit the potential for intra- and inter-species transmission [58]. The incidence of infected cattle herds has been drastically reduced over the last two decades [7] but complete eradication remains elusive, likely as a result of persistent infection in wildlife populations.

Discriminatory molecular typing tools have been extremely helpful in the study of *M. bovis* infection in livestock, informing the tracking of infection [94, 104, 92] and improving

our understanding of how bTB spreads and persists [62, 202]. Traditionally in New Zealand, Restriction Endonuclease Analysis (REA) typing was used extensively during bTB surveillance. Cattle and wildlife were shown to share the same REA type [21], and importantly, local regionalisation of REA types enabled the distinction between re-infection and introduction [36]. While REA typing is discriminatory, it is technically challenging to perform, interpret and document, and has recently been replaced with Variable Number Tandem Repeat (VNTR) typing [92].

The advent of Next Generation Sequencing has made it increasingly feasible to sequence and compare Whole Genome Sequences (WGS) in order to inform epidemiological analyses. WGS data provide the highest resolution and, therefore, discriminatory power for understanding the sampled system [24, 118]. Recently Glaser *et al*. [26] used WGS data to distinguish outbreaks carrying identical VNTR types, as well as identifying transmission within and between cattle and deer populations. Similar work in New Zealand has demonstrated the utility of WGS as a robust and highly discriminatory typing method (in preparation: Price-Carter *et al*. 2017). Biek *et al*. [24] used WGS methods to examine bTB transmission in Northern Ireland, and demonstrated that badgers and livestock living in close proximity shared very similar *M. bovis* strains, suggesting that multiple inter-species transmission events had occurred.

Our research aimed to refine our understanding of the role of wildlife in the transmission and persistence of bTB across New Zealand and estimate the substitution rate of *M. bovis* in this system. Samples taken from infected cattle and wildlife provided *M. bovis* isolates for which WGS data was generated. In agreement with previous knowledge, wildlife species were implicated in the transmission and persistence of bTB infection in the sampled population. We found evidence of multiple inter-species transmission events and estimated their force and direction. Estimating the transmission direction was found to be influenced by the sampling patterns. The availability of WGS data presented the opportunity to evaluate the use of WGS in routine typing. WGS methods were able to discriminate isolates to a finer resolution than REA typing, and there was good agreement between these typing methods. The utility of WGS techniques depends on the frequency with which mutations are fixed within the population. The estimated substitution rate was higher than those previously estimated for *M. bovis*.

3.2 Materials and Methods

3.2.1 Sampling and Isolate Preparation

As part of the routine bTB surveillance in New Zealand, any cattle or wildlife suspected of *M. bovis* infection undergo a post-mortem examination, and if lesions are discovered a selection are investigated using culture and strain typing. Conventional tests (described in [203]) were used to positively identify *M. bovis* infection. Isolates were REA typed according to previously described methods [90, 94] and cultures were frozen and stored in the strain archive at AgResearch Ltd. Isolates from the archive were selected to provide a representative sample of the *M. bovis* population circulating in cattle and wildlife across New Zealand between 1985 and 2013 (Figure 3.1).

To create this representative sample, groups of isolates, from cattle and wildlife, of the same or closely related REA type from the same geographical region were selected from the Central North Island region, and the West Coast and Northeast regions of the South Island. The groups were selected to include all of the most frequently isolated REA types.

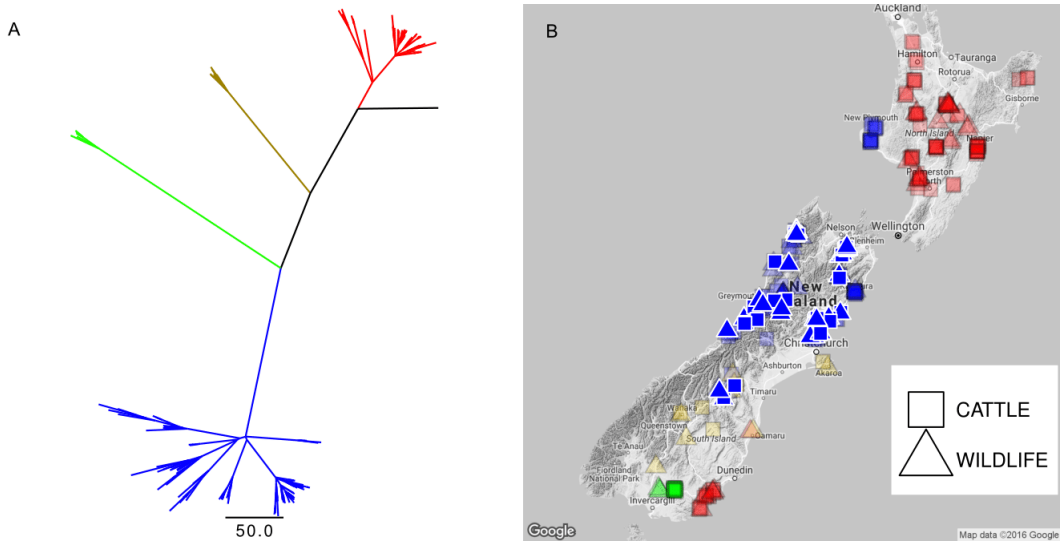

Figure 3.1: (A) An unrooted maximum likelihood phylogenetic tree built using PHYLIP [204] and rooted using PATH-O-GEN [205]. Assigned clades are coloured accordingly: clade 1 = blue, clade 2 = red, clade 3 = gold and clade 4 = green. (B) The sampling locations of the isolates are plotted onto a map of New Zealand. Cattle and wildlife isolates are represented by squares and triangles, respectively. Isolates are coloured by their associated clade in the phylogenetic tree (A). Only isolates from clade 1 (blue) were selected for further analysis (white outline), faded isolates are those not selected. Isolate locations were jittered to ease interpretation.

Selected isolates were re-cultured at AgResearch Ltd. to generate DNA (Deoxyribonucleic Acid) extracts for WGS. Frozen culture stocks were grown to mid log phase (OD600 = 0.4-0.8) in 5ml of Tween/albumin broth [206] and sub-cultured into 100ml of the same media for

5 to 11 weeks until the cultures reached stationary phase. Cell cultures were heat killed and stored at -20°C. Bacterial DNA was specifically separated from the other cellular components with a high salt hexadecyl trimethyl ammonium bromide (CTAB) extraction [183, 91]. DNA was then extracted with a chloroform/isoamyl alcohol method and quantified using Invitrogen Qubit fluorometry.

A selection of 90 DNA isolates were sequenced at the University of Glasgow Polyomics Facility using an Illumina MiSeq platform that produced 2 x 300bp paired end reads per isolate. 17 additional isolates were sequenced at New Zealand Genomics Ltd. on a MiSeq platform that produced 2 x 250bp paired end reads. The remaining isolates (n=204) were sequenced at the Wellcome Trust Sanger Institute using an Illumina HiSeq that produced 2 x 100bp paired end reads.

3.2.2 Processing Sequencing Data

The raw reads for each isolate were examined using FASTQC (v0.11.2 - [189]) to identify poor quality ends that were then trimmed using PRINSEQ (v0.20.4 - [191]). If adapter sequences were present these were removed using TRIMGALORE (v.0.4.1 - [190]). Each isolate's trimmed reads were aligned to the *M. bovis* reference genome, AF2122/97 [102], using the freely available Burrows-Wheeler Alignment tool [192, 176]. The mean coverage (sites with Read Depth (DP) $\geq$20) for the isolates was 99% (2.5% Lower: 96.9, 97.5% Upper: 99.8).

Site information across the isolates was collated to allow the quality of individual sites to be assessed. Sites that fell within Proline-Glutamate (PE) and Proline-Proline-Glutamate (PPE) genes or annotated repeat regions were removed [193]. Thereafter only variant positions, sites for which at least one of the isolates showed variation against the reference genome, were retained.

High quality sites were selected for subsequent analyses based on the Mapping Quality (MQ), High Quality base Depth (HQDP) and Read Depth (DP). The variant positions for each isolate were filtered according to their specific MQ, HQDP, DP and allele support (SUP, the proportion of reads mapped to the position of interest that support the allele called for that position). In addition, the proportion of isolates which had sufficient quality at each variant position (COV) and the number of positions that SNPs had to be apart (Site proximity – PROX) were used as filters. A sensitivity analyses was conducted in order to establish an optimal filter combination.

For each isolate the sampling year, location (region, district, and latitude and longitude data of sampling location), sampled species, and the REA (Restriction Endonuclease Analysis) type was available. This isolate information was used to inform the selection of thresholds

(for the filters described above), based upon the assumption that these data should, to some degree, explain the variation observed in the inter-isolate genetic distance distribution.

Inter-isolate comparisons were made using the discrete (district, region, species, and REA) and continuous (spatial and temporal distances) data that were then fitted to the corresponding inter-isolate genetic distance using a Random Forest regression model [197]. The Random Forest model was fitted in the statistical programming environment R (3.0.1 [194]). A filter combination was selected by maximising the variation explained using the fitted Random Forest model. The following filters were selected: MQ$\geq$30, HQDP$\geq$4, DP$\geq$30, SUP$\geq$0.95, COV$\geq$0.7, PROX=10. Variant positions were filtered based on this filter combination to create a concatenated sequence FASTA file.

3.2.3 Isolate Selection

An early examination of the WGS data revealed that, although most isolates with the same REA type were very similar, several isolates were quite distinct from the others with the same REA type. These "outliers" were further investigated to determine whether they were mislabelled. Although it was not possible to re-examine these isolates with the REA typing method, potentially mislabelled samples were further examined with Variable Number Tandem Repeat (VNTR) assays (conducted by Marian Price-Carter of AgResearch). Specific REA types are known to be associated with specific VNTR types. VNTR assays were conducted (described in [92]) using a subset of VNTR loci that were likely to discriminate the isolates in question. For controls, a selection of isolates with similar sample numbers to the questionable isolates was also re-examined. If the determined VNTR types differed from what would have been expected based on previous analyses of these types [92], the isolate was considered to have been mislabelled. Out of the 28 (14 suspects and 14 controls) isolates that were examined, 15 had VNTR loci that differed from what would have been expected. These 15 mislabelled isolates were removed from any further analyses (Table 3.1), leaving 296 isolates for further investigation.

Using the 296 isolates, a maximum likelihood phylogenetic tree was constructed in the program PHYLIP (v3.695 - Felsenstein 1989) and rooted using the program PATH-O-GEN (v1.4 - [205]). For each isolate the sampling location (including latitude and longitude) and year (of sample submission), sampled species, and REA type were available. Using the maximum likelihood tree and the available sampling information, a selection of spatially and temporally associated isolates were chosen from within clade 1 (Figure 3.1: A).

Although a large number of isolates were available for the current analyses, these isolates fell within highly distinct clades. Isolates from a single clade were selected to ensure a relatively recent common ancestor to the isolates analysed, and limit the effects of biases introduced by examining genetically distinct groups. Unique pairs of cattle and wildlife isolates were

Isolate ID	Type	REA	Exp etrD	Obs etrD	Exp DR1	Obs DR1	Exp NZ2	Obs NZ2	Exp Q26	Obs Q26	Exp Q11a	Obs Q11a	Exp DR2	Obs DR2	Exp 3232	Obs 3232
2122/97	Control	21	5	5	5	5	-	-	-	-	-	-	-	-	-	-
202	Suspect	1/6	4	5	5	5	-	-	-	-	-	-	-	-	-	-
203	Suspect	62	4	4	3	5	-	-	-	-	-	-	-	-	-	-
204	Control	62	4	4	3	3	-	-	-	-	-	-	-	-	-	-
205	Suspect	62	4	3	3	5	-	-	-	-	-	-	-	-	-	-
206	Suspect	1/6	4	4	5	3	-	-	-	-	-	-	-	-	-	-
207	Control	62	4	4	3	3	-	-	-	-	-	-	-	-	-	-
210	Suspect	21	5	4	5	3	-	-	-	-	-	-	-	-	-	-
211	Suspect	300	3/4	4	5	3	5	5	4	4	9	9	15	15	6	8
212	Control	62	4	4	3	3	5	5	4	4	9	9	15	15	8, 9	8
213	Suspect	19	3	4	5	5	-	-	-	-	-	-	-	-	-	-
122	Suspect	21	3/5	4	5	5	5	5	4	4	9	9	15	15	5,7,8	8
123	Control	62	4	4	3	3	5	5	4	4	9	9	15	15	8, 9	8
185	Suspect	93	4	5	6	5	5	5	4	4	10	9	13	15	9	8
186	Control	62	4	4	3	3	5	5	4	4	9	9	15	15	8, 9	8
267	Suspect	16	3/4	5	5	5	5	5	4	4	6/9	9	15	15	7,8,9	8
268	Control	12	4	4	5	5	5	5	1,3,4	3	9	9	15	15	8,9,10,12	8
314	Suspect	6	4	4	5	3	5	5	4	4	9	9	15	15	6	8
315	Suspect	62	4	4	3	5	5	5	4	4	9	9	15	15	8, 9	6
316	Control	6	4	4	5	5	5	5	4	4	9	9	15	15	6,4,8	6
9	Control	113	4	4	6	5	5	5	4	3	10	10	12	12	10	10
10	Control	115	4	4	5	5	5	5	3	3	10	10	12	12	7	7
11	Control	115	4	4	5	5	5	5	3	3	10	10	12	12	7	7
12	Control	115	4	4	5	5	5	5	3	3	10	10	12	12	7	7
18	Control	115	4	4	5	5	5	5	3	3	10	10	12	12	7	7
19	Suspect	113	4	4	6	5	5	5	4	3	10	10	12	12	10	7
20	Suspect	115	4	4	5	5	5	5	3	3	10	10	12	12	7	10
21	Control	151	4	4	5	5	6	6	4	4	10	10	12	12	7	7

Table 3.1: The expected and observed VNTR loci present for a selection of 14 suspect and 14 control isolates. Where the observed doesn't match the expected VNTR loci for an isolate, the cells showing the mismatch and isolate ID were highlighted in grey.

chosen from within strict spatial (40km) and temporal (+/- 3 years) limits to reduce the impact of potential temporal and spatial sampling biases (Figure 3.1: B & Figure 3.2). The spatial and temporal thresholds were chosen so as they were the minimum values necessary to retain a large enough sample size for further analyses. Using the spatial and temporal thresholds described, only Clade 1 had enough spatially and temporally associated isolates to warrant further analyses.

3.2.4 Clustering of Inter-Isolate Genetic Distances

The available data for the isolates – REA type, sampling location (district where the sampling took place) and sampled host - were used to define groups of isolates and the within- and between-group genetic distances were examined to determine whether there was an association. The concatenated sequence of variant positions of each isolate was compared to one another to generate an inter-isolate genetic distance (using the p-distance – defined as the proportion of the sites that differ between two sequences).

The observed difference between the mean intra- and inter-group genetic distances was calculated where the groups were defined, separately, by host species sampled, sampling location and REA type. To determine whether each observed difference could have arisen by chance alone, the isolate data were shuffled and the difference re-calculated. The shuffling was repeated 10,000 times to generate null distributions of observed differences. The associations were considered significant if the observed metric fell outside the lower (2.5%) and upper (97.5%) quantiles of the null distribution. Importantly, any species signature is likely to be nested within a spatial one, since regional localisation of bTB is known. To account for this, only comparisons that were between isolates sampled in the same district were included in the clustering analyses using the host species sampled.

3.2.5 Phylogenetic Analyses

The Bayesian Evolutionary Analysis by Sampling Trees (BEAST v1.8.4 - [207]) software was used for a phylogenetic analysis of the isolates' sequences combined with their sampling years. BEAST was used to estimate the phylogenetic tree topology, substitution rate and date of the Most Recent Common Ancestor (MRCA) for the sampled *M. bovis* population. A BEAST analysis requires the existence of a clock-like substitution process.

Additional analyses, as conducted by Firth *et al.* [208], were used to examine whether a clock-like process could have produced the inter-isolate variation. A linear regression was conducted on the sampling year against the root-to-tip distance for the temporally and spatially matched isolates from clade 1 and the reference (to aid with the tree rooting). A significant relationship was observed (p-value=0.018) with an R^2 value of 0.075. The estimated substitution rate, 0.53 (2.5% Lower: 0.22, 97.5% Upper: 0.94) events per genome per year, was compared to an analysis using the same model structure but the associated sampling dates

were randomly shuffled. The sampling date shuffling was repeated 10 times and BEAST was used to estimate the substitution rate. A substitution rate of $2.85 * 10^{-5}$ (2.5% Lower: $2.81 * 10^{-6}$, 97.5% Upper: $1.24 * 10^{-4}$) events per genome per year was estimated using the shuffled data. These estimations are significantly lower than those estimated on the unshuffled data. The difference between the substitution rates based upon shuffled and true data, in addition to the significant relationship observed between the root-to-tip distances and sampling year, support the presence of a temporal signal to inform the estimation of a substitution rate for the sampled population.

Models selected in a BEAST analysis may significantly impact the results. Care must be taken to select appropriate models for the substitution process [209] and the underlying population dynamics [210]. A series of BEAST analyses were completed in a hierarchical fashion to explore the different models available; for each analysis a chain length of 500,000,000 steps, sampled every 50,000 steps, was used and three replicates were completed. This approach had three levels: 1) a range of substitution models were examined whilst using the simplest clock and population models, 2) once the substitution model had been selected, the available clock models were evaluated using the selected substitution model and the simplest population model, and 3) lastly, the different population models were investigated in combination with the selected substitution and clock models. Following the removal of a 10% burn-in, the posterior distributions were examined to determine which structure of BEAST analysis best described the isolate data. At each level (1, 2, and 3) the analyses were compared based upon log likelihood scores, model convergence and posterior support of parameters (assessed using TRACER v1.6 [211]), path sampling and stepping stone analyses. In addition the biological feasibility of the results was examined for each analysis (Table 3.2).

The selected BEAST analysis used the Hasegawa-Kishino-Yano (HKY) substitution model, a relaxed clock model, drawing from an exponential distribution, and the Gaussian Markov Random Field (GMRF) Bayesian Skygrid population model. The HKY substitution model allows variable base frequencies, transition and transversion rates to be estimated [212]. A relaxed clock model enabled the estimated substitution rate to vary across the branches of the phylogenetic tree; the extent of this variation was modelled using an exponential distribution. The GMRF Skygrid model is a flexible model that is able to estimate changing population dynamics over the course of a phylogenetic history [210]. In a BEAST analysis, population dynamics are estimated based on the structure of the phylogenetic tree according to coalescent theory [213].

An additional Discrete Ancestral Trait Mapping (DATM) analysis [214, 209] for two states was implemented in the BEAST analysis. According to the host species sampled, isolates were assigned either a cattle or wildlife state. Based upon the states of the tips of the phylogenetic tree (the isolates), the DATM estimates the ancestral states in the phylogeny, and as such the most likely sources of infection within the sampled *M. bovis* population. A compar-

Table 3.2: A hierarchical approach to model selection for the BEAST analyses. Each model structure described above (defined by the substitution, clock and population models selected) was repeated three times using a chain length of 500,000,000 with every 50,000 step being sampled. The average likelihood (across the three replicates) is reported for the Path Sampling and Stepping Stone model comparison methods. In addition, whether or not the replicates converged was reported.

Run	Substitution Model	Clock Model	Population Model	Path Sampling	Stepping Stone Sampling	Converged
1	JC	Strict	Constant	-5629980.18	-5629974.54	YES
2	HKY	Strict	Constant	-5435731.8	-5435724.63	YES
3	GTR	Strict	Constant	-5435743.4	-5435740.84	YES
4	HKY	Relaxed-Log	Constant	-	-	NO
5	HKY	Relaxed-Exp	Constant	-5435687.14	-5435685.9	YES
6	HKY	Relaxed-Exp	Logistic	-	-	NO
7	HKY	Relaxed-Exp	Exponential	-5435682.54	-5435681.21	YES
8	HKY	Relaxed-Exp	Expansion	-	-	NO
9	HKY	Relaxed-Exp	Skyline	-5435682.32	-5435681.25	YES
10	HKY	Relaxed-Exp	Skyride	-5435692.02	-5435690.3	YES
11	HKY	Relaxed-Exp	Skygrid	-5435685.72	-5435683.68	YES

ison was made between a symmetric and asymmetric DATM analyses in BEAST using the spatially and temporally matched isolates. The former symmetric analysis refers to the state transition matrix being symmetric; this analysis estimates a single parameter (in a two state analysis), the transmission rate of the pathogen from one state to another. The asymmetric analysis has two inter-state transmission parameters and as such can be used to determine whether there is a directional bias in the exchange; is the pathogen jumping from one population into another more often than in the other direction?

3.2.6 Influence of Prior Selection

The influence of the selection of prior distributions for the parameters estimated in the BEAST analyses, described above, was investigated by running an analysis where the data were removed and only the prior distributions sampled. It was shown that the prior distributions selected were conservative and that the data provided a strong signal for the parameter estimations of our model.

The models selected in a BEAST analyses have set parameters whose estimation requires the specification of prior distributions. Here prior knowledge about the sampled population was used to inform the BEAST analyses.

The prior distributions for the final model were specified as follows:
HKY substitution model

- Allele frequencies were estimated using the sequence data provided.

- The transition-transversion parameter prior distribution was a Log Normal distribution with a log(mean) of 1 and log(SD) of 1.25.
- Four categories, modelled with a gamma distribution, were included to allow the substitution rate to vary across sites. The shape parameter for the gamma distributions was estimated from an Exponential distribution with a mean of 0.5.

Clock Model

- A relaxed clock model drawing from an Exponential distribution with a mean of 0.005 (events per site per year) was used.

Population Model

- The Skygrid model uses a smoothing parameter to avoid large jumps in the population size estimated based upon the structure of the phylogenetic tree. A Gamma distribution (shape and scale values equal to 0.001 and 1000, respectively), was used as the prior distribution for the smoothing parameter.

The prior distributions specified define the space used to estimate the parameters of interest. It is important to investigate the extent to which the estimates resulting from a BEAST analysis are influenced by the specification of the prior distributions. A BEAST analysis, using the prior specifications defined above, was completed using only the sampling year data (for the spatially and temporally matched isolates from clade 1). This prior sampling analysis estimated a substitution rate of 0.004 (97.5% Lower: 0, 2.5% Upper: 0.19) events per site per year. This estimate is orders of magnitude higher than the substitution rate estimated when the genetic data were included. A right-skew on the true substitution rate estimates was not evident; despite the broad and inaccurate prior distribution specified, there was enough of a signal in the data to estimate the substitution rate.

3.3 Results

3.3.1 Structure in the sampled *M. bovis* population

There were four recognisably distinct clades formed by the 321 isolates sampled in New Zealand, which were regionally localised (Figure 3.1). A total of 3449 variant positions were found. Long distance translocation and establishment of new foci of infection was evident when the genetic structure of the population was considered. Clade 2 (Figure 3.1 – red), although mostly found in New Zealand's north island, has an established foci of infection involving both cattle and wildlife on the south island. Clade 1 (Figure 3.1 – blue) isolates were mostly situated on the south island of New Zealand, providing a densely sampled genetically similar set from which to select the spatially and temporally associated isolates for further analysis (Figure 3.1 & Figure 3.2). Clade 3 (Figure 3.1 – gold) included eight wildlife and five cattle isolates, found across a broad spatial range in the southwest of New Zealand's South Island. Clade 4 (Figure 3.1 – green) included thirteen cattle and three wildlife isolates, which were sampled from two locations <20km apart in the south of New Zealand's South Island.

3.3.2 Clustering of Inter-Isolate Genetic Distances

The inter-isolate genetic distance distribution of the spatially and temporally associated isolates from clade 1 was examined. Isolates of the same REA type were, on average, more genetically similar than those of different types. This difference was reflected in lower average within- than between- group genetic distances, when groups were defined by REA types (Figure 3.3: B). In addition, diversity was evident in the within group distances demonstrating the added resolution of WGS data. The observed difference between the mean inter- and intra-group genetic distances was unlikely to have arisen by chance when the isolates were grouped by their REA type or sampling location (Figure 3.3: B & C). In contrast to the lower within- than between-group genetic distances observed when groups were defined by REA type or sampling location (as was demonstrated by the positive observed difference (Figure 3.3: B & C)), when groups were defined by the host species sampled, the within-group distances were, on average, higher than the between-group distances (Figure 3.3: D). These higher within-group distances resulted in a negative observed difference, which was unlikely to have arisen by chance as it fell just outside the 95% bounds of the generated null distribution. This negative difference may be caused by lower within-outbreak distances resulting from sampling local outbreaks (involving cattle and wildlife) that are separated in space.

3.3.3 Substitution Rate Estimation

Using a bootstrapping procedure the posterior distributions resulting from BEAST analyses incorporating different population models were compared (Figure 3.4: B). For each pairwise posterior comparison, a distribution of differences was generated by calculating the differ-

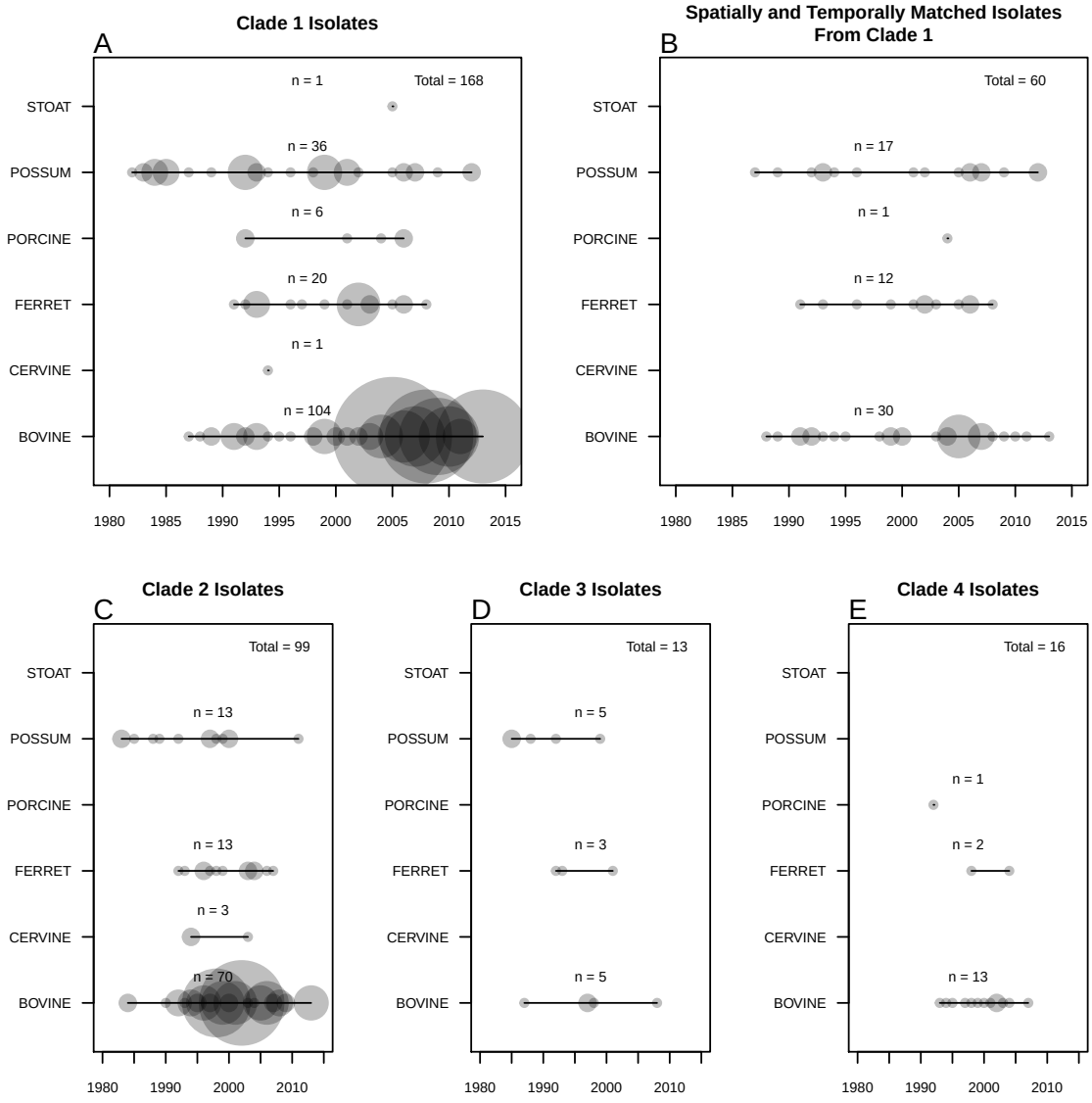

Figure 3.2: Five plots illustrating the temporal range associated with each sampled host species for all the isolates in the different clades (1 (A), 2(C), 3(D), and 4(E)), and the spatially and temporally associated isolates from clade 1 (B). The size of each point is scaled by the number of isolates that were taken from the given species in the given year.

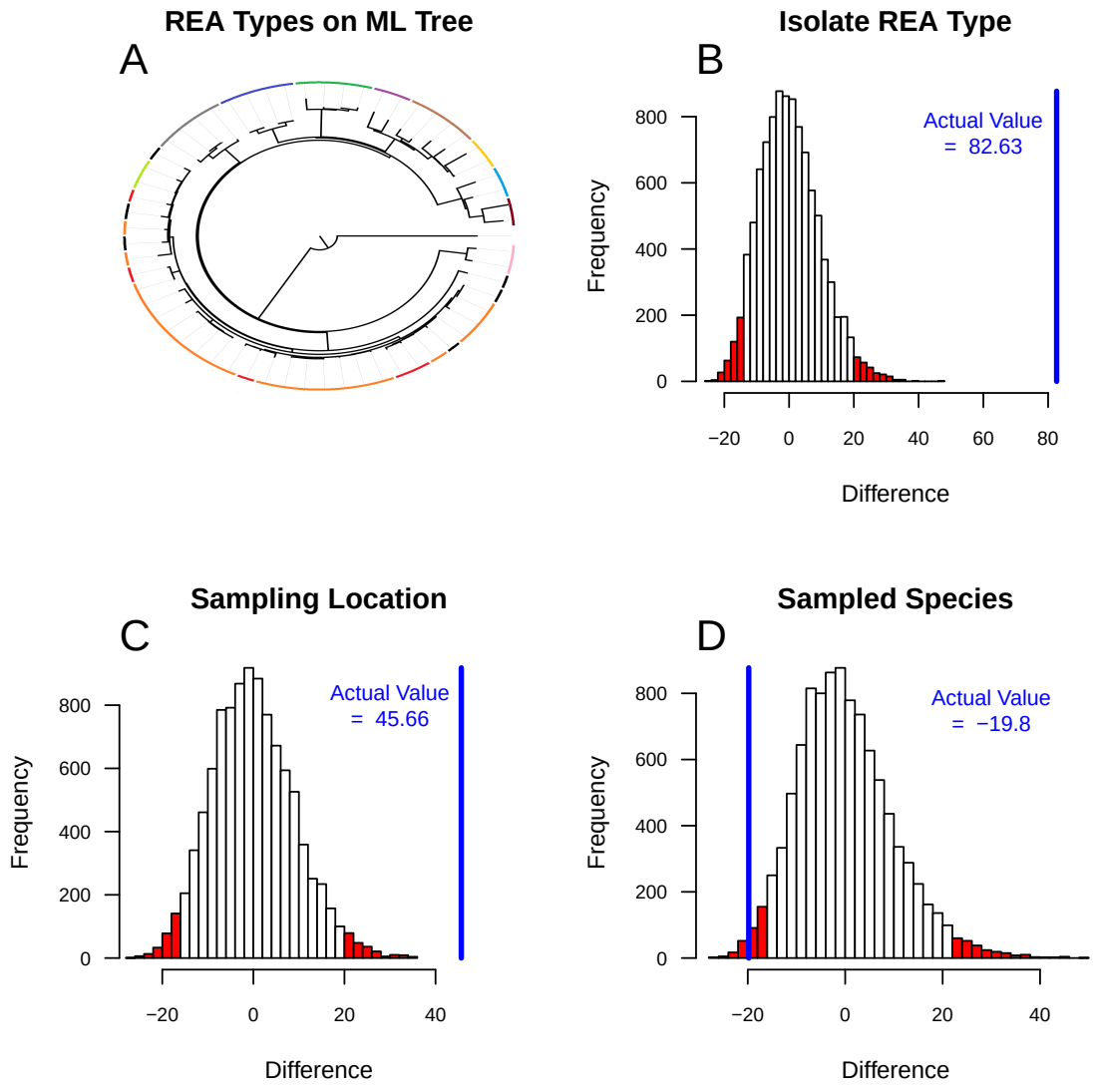

Figure 3.3: Clustering in the inter-isolate genetic distance distribution for the spatially and temporally matched isolates from clade 1. (A) A Maximum Likelihood phylogenetic tree generated using PHYLIP; coloured bars are used to highlight isolates that have the same REA type (note that REA types that are only represented by one isolate are colour in black). (B, C, and D) Three plots showing how the observed difference between the mean inter- and mean intra-group genetic distances, when isolate groups were defined by REA, Sampling Location, and Species (B, C, and D, respectively) compared to null distributions of differences calculated on shuffled sequences. The sampling location was defined as the region where sampling occurred. The difference was calculated for 10,000 independently shuffled sets. Only the spatially and temporally matched isolates from clade 1 were used in this clustering analysis. The blue line shows the observed difference between mean inter- and mean intra-group genetic distances. The area outside of the lower (2.5%) and upper (97.5%) bounds of the null distribution are coloured in red.

ence between single point estimates, that were sampled proportionately from each of the two posterior distributions. If similar distributions are compared using this pairwise comparison, the calculated differences between point estimates drawn randomly from each distribution will be close to zero. The paired posterior distributions were not significantly different; the distribution of calculated differences resulting from each pairwise comparison overlapped with zero. The Skygrid population model, which had a high likelihood in the model selection procedures (Table 3.2) and agreed well with the other population models used (Figure 3.4), estimated the substitution rate of the sampled *M. bovis* population to be 0.53 (2.5% Lower: 0.22, 97.5% Upper: 0.94) events per genome per year.

Using the Skygrid population model, the MRCA to the sampled *M. bovis* population was estimated to have been circulating in 1859 (2.5% Lower: 1525, 97.5% Upper: 1936). Binney *et al.* [215] recently established that a large number of cattle were imported into New Zealand in the 1860s, mostly originating from Australia and the United Kingdom. The structure of clade 1 and of the full maximum likelihood phylogeny (Figure 3.1: A) aren't indicative of a single introduction event into New Zealand. The sampling window used in this study was narrow, relative to the phylogenetic history of the sampled population. Via simulation it was shown that estimates of the substitution rate were robust to the shortening of the sampling window; estimates increasingly lacked precision but retained accuracy (Appendix A).

3.3.4 Ancestral Traits Analysis

Different selections of isolates from clade 1 (all isolates, temporally and spatially matched isolates, or 30 random cattle and wildlife pairs) were used in separate DATM BEAST analyses. According to the path sampling likelihood values, the symmetric model (equal rates from cattle to wildlife and vice versa) was favoured for the matched isolates (Figure 3.5: A). The asymmetric (different rates) was favoured when the DATM analysis was based on all the clade 1 isolates or the randomly paired isolates (Figure 3.5: A). The DATM analyses were able to provide an estimate for the overall state transition rate (Figure 3.5: B). For the analyses based on all the clade 1 isolates or randomly matched isolates the asymmetric model provided directional estimates of the state transition rates (Figure 3.5: C & D). Using all isolates or the random cattle and wildlife pairs, a dominant direction of transmission from wildlife to cattle was estimated (Figure 3.5: C & D). Using the spatially and temporally matched isolates the symmetric model out-performed the asymmetric model and directional estimates weren't available. The difference in the support for the symmetric versus asymmetric model was a result of the isolates selected and therefore shows a strong influence of sampling. Without knowing which sample set is most representative it is difficult to have confidence in the directional state transition rates.

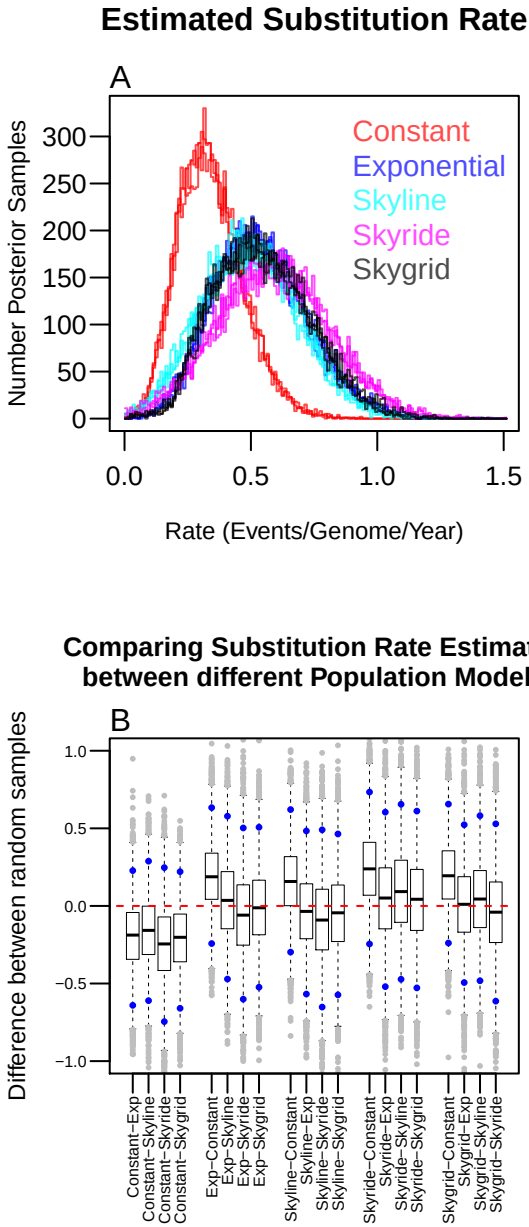

Figure 3.4: The estimated substitution rate of the sampled *M. bovis* population. (A) The sampled (n=9000, 10% burn-in removed) posterior distributions of the substitution rate estimated by BEAST analyses using different population models. Each analysis in BEAST was repeated 3 times and replicates plotted with the corresponding colour for the population model. (B) Pairwise comparisons of the posterior distributions resulting from analyses based upon different population models. Each boxplot summarises the distribution of differences produced by calculating the difference between 10000 random samples of the posteriors being compared. The blue points represent the upper and lower bounds of distribution of differences. Outliers of the difference distributions are coloured in grey.

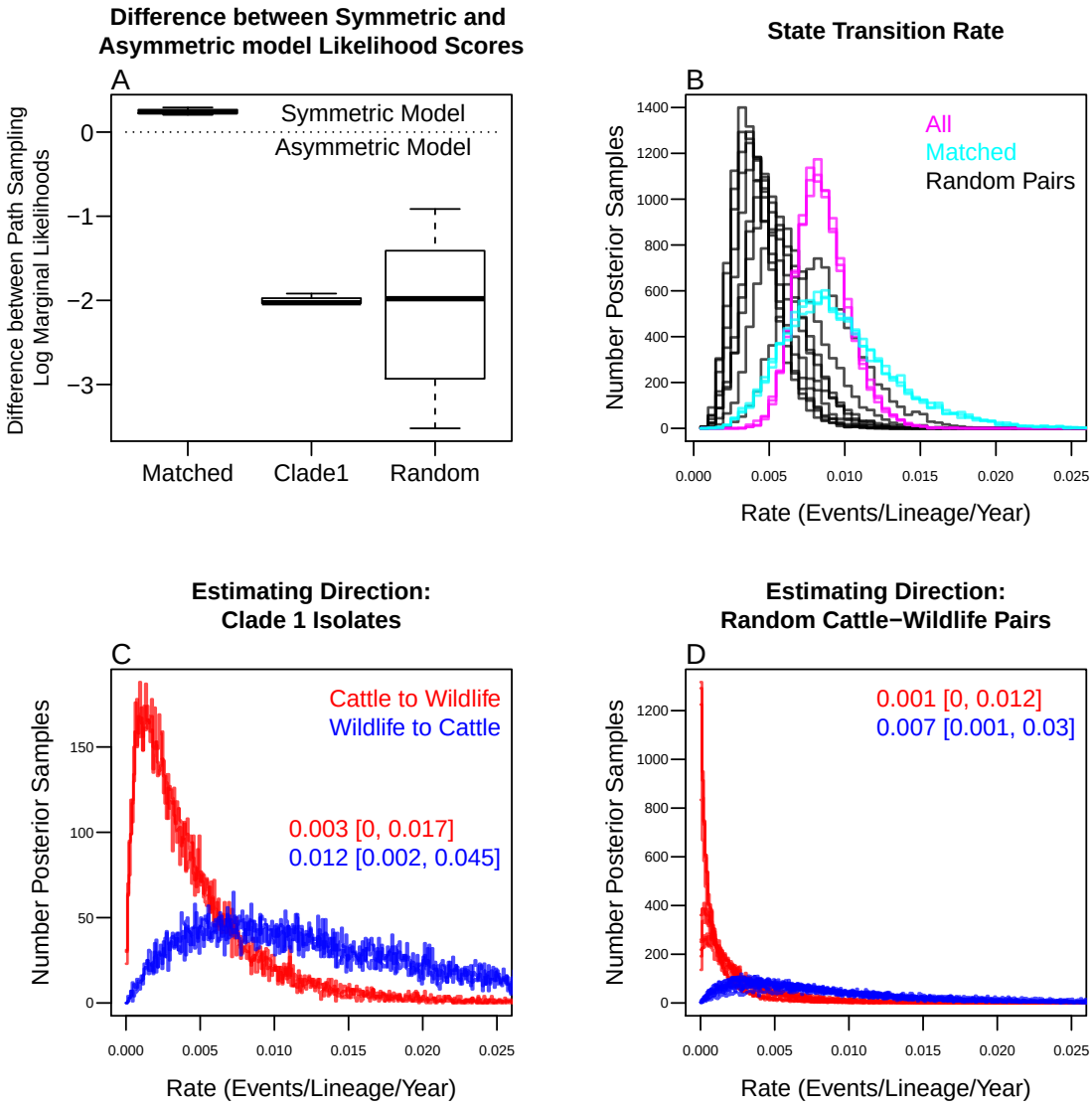

Figure 3.5: The state transition rates estimated by a discrete traits analysis in BEAST on isolates selected from clade 1. States were defined as either Cattle or Wildlife. BEAST analyses were completed using all the clade 1 isolates (3 replicates), only spatially and temporally matched isolates (3 replicates) and 30 randomly matched cattle and wildlife isolates (10 replicates). (A) A box plot of the difference between the likelihoods (estimated using path sampling) of the symmetric and asymmetric models for the different sampling sets. The symmetric model was favoured for the matched isolates and the asymmetric model for the analyses based on all the clade 1 isolates and randomly matched cattle and wildlife. (B) The sampled (n=10,000) posterior distributions of the estimated overall transition rate between Cattle and Wildlife based on the three sampling sets. Plots C and D show the posterior distributions of the transition rates from Cattle to Wildlife (Red) and Wildlife to Cattle (Blue) resulting from BEAST analyses completed on the clade 1 isolates and the randomly matched isolates. The median and 95% Credible Intervals are stated for the distributions.

3.4 Discussion

The current research suggests that *M. bovis* infection was being transmitted between the sampled wildlife and cattle populations. In Northern Ireland, where the role of badger populations is under investigation, WGS data with only a few wildlife isolates has been used in an attempt to elucidate the mechanisms of persistence of bTB in cattle herds [24, 25]. High genetic similarity suggested recent transmission links between badger and cattle populations. Similarly, Glaser *et al*. [26], used WGS data to reveal exchange within and between cattle and deer populations in Minnesota.

By the end of June 2015, there were 39 infected cattle herds in New Zealand [60]. Whilst cattle movements remain a recognised cause of newly detected herd infections, wildlife are thought to be the main contributors. In the current research, isolates originating from cattle and wildlife sources were indistinguishable suggesting a high degree of exchange. A high degree of exchange was supported by the estimations of an overall inter-species transmission rate. With New Zealand's low prevalence of bTB in livestock, despite being unable to estimate inter-species transmission direction in the current research, it would seem highly likely that wildlife populations are acting as maintenance reservoirs and as such should remain the target of the eradication campaign.

When investigating any epidemiological process using genetic data of a pathogen, the relative speed of that epidemiological process compared to the rate of change of the sampled pathogen must be considered [121]. Ideally, the sampling of a system of interest should reflect the underlying epidemiological processes, and not produce additional noise or biases. For example, if isolates are too distantly related (both genetically and epidemiologically) noise may dominate the signal of the epidemiological events of interest, making the estimation of these events difficult. Here, the inter-species transmission rate was estimated. The difficulty encountered when estimating the direction of inter-species transmission may be a reflection of a high rate of exchange of *M. bovis* between cattle and wildlife estimated using a slowly evolving pathogen sampled from a broad genetic distribution.

The role of wildlife in the maintenance of bTB in New Zealand could provide an explanation for why the substitution rate estimated here was relatively high, in comparison to previously published estimates of the substitution rate for the *M. tuberculosis* complex (Table 3.3). This difference will enhance the utility of genomic data for routine epidemiological investigations because it will allow for better estimates of the time of introductions of new infections into herds and wildlife populations and thus aid in the identification of likely sources of infection.

Most brushtail possums suffer an extensive *M. bovis* infection if exposed, and many will die within 6 months [216, 16]. In contrast, the majority of humans, cattle, and badgers suffer

Table 3.3: A comparison between estimates of the substitution rates (events per genome per year) taken from WGS analyses on *M. tuberculosis* and *M. bovis*, with the results of this study inserted in the final row.

Published Source	Bacteria Species	Mean/Median	Lower	Upper	Host Sampled	Country
Walker *et al.*, [116]	*M. tuberculosis*	0.5	0.3	0.7	Human	UK
Bryant *et al.*, [117]	*M. tuberculosis*	0.3	NA	NA	Human	Netherlands
Roetzer *et al.*, [118]	*M. tuberculosis*	0.4	0.3	0.7	Human	Germany
Biek *et al.*, [24]	*M. bovis*	0.15	0.04	0.26	Cattle/Badger	UK
Trewby *et al.* [25]	*M. bovis*	0.2	0.1	0.3	Cattle/Badger	UK
Current Research	*M. bovis*	0.53	0.22	0.94	Cattle/Possum	New Zealand

a localised latent TB infection [217, 31, 65]. Given that herd breakdowns in New Zealand are thought to be mainly the result of spill-over events from wildlife vectors [60, 36], the higher levels of replication during the more extensive infection in possums could result in an increased accumulation of mutations for the sampled *M. bovis* population.

Colangeli *et al.* [120] demonstrated that the likely lower rates of replication occurring during latent *M. tuberculosis* infection, in humans, resulted in significantly lower accumulation of mutations when compared to active infection. This research supports the theory that replication rates impact substitution rates and is consistent with other observations on host-level variability [116]. However, Ford *et al.* [114] were unable to find an effect of latency on the substitution rate of *M. tuberculosis* in infected Macaque monkeys, in an experimental setting, and so this area requires further study. Alternatively, the higher substitution rate could be the result of a lineage specific trait, which has been demonstrated for *M. tuberculosis* [218, 219].

The patterns of sampling and their influence on results of any analysis are an important consideration. Broad credible intervals were estimated around the substitution rate and date of the MRCA for the sampled population. The isolates analysed in the current research were sampled between 1987 and 2013; relative to the estimated root height (1859 [1525, 1936]), this sampling window is narrow. *M. bovis* is likely to have been circulating in New Zealand since the mid-1800s [7], therefore sampling early in this outbreak wasn't possible. Analyses based on simulated epidemics sampled using an increasingly late and narrow window, demonstrated that a narrow sampling window had a pronounced effect on the precision of estimates but, importantly, little effect on the accuracy of parameter estimation in BEAST (See Appendix A).

In the DATM analysis a temporal bias was evident in the original set of clade 1 isolates (Figure 3.2), with dense sampling of wildlife in early years and of cattle in later years, resulted in a dominant direction of spread from wildlife to cattle being estimated. Using the current data it wasn't possible to determine whether this dominance exists, and the sampling

patterns are a true reflection of New Zealand's bTB system, or the directionality observed was an artefact of the sampling patterns.

The WGS data provided added resolution to the examination of bTB in New Zealand, distinguishing isolates sharing an identical REA type (Figure 3.3: A). The declining cost, added resolution, good agreement with REA typing, and evidence of a strong spatial signature all act to endorse the use of WGS typing in routine surveillance.

The utility of any typing method lies in its molecular clock speed; too quick and noise masks important events, too slow and important events could be missed [111, 112]. Both human- and bovine-TB are caused by slowly evolving, genetically conserved, bacteria [102, 62]. With such a recognisably slow rate of change it is unlikely that infection dynamics within or between individuals will result in significant genetic signatures. In contrast herd level signatures are likely to be present and of use in routine surveillance that targets herds [24, 121, 26].

In the current research it was shown that knowledge of epidemiology combined with WGS data can provide a means for in-depth investigations of bTB dynamics, shedding light on important and as yet unquantified features, such as the extent of inter-species transmission and the substitution rate. A caveat though, the influence of the sampling strategy used, should be thoroughly examined as to its potential impact on any findings. Targeted control of wildlife populations is part of New Zealand's eradication strategy [58] and wildlife were implicated in the current research. Identifying local persistence or introduction is the focus of bTB surveillance in New Zealand and regional localisation of isolates makes this possible. WGS data, despite the low substitution rate of *M. bovis*, adds resolution, decreasing the scale at which persistence versus introduction can be evaluated. In addition, an estimate of the substitution rate of *M. bovis* in New Zealand, however broad, will inform these evaluations. For routine surveillance, the resolution gained by using WGS data must be weighed against any increased costs, a decision that will be aided by the decreasing price of sequencing technologies.

Chapter 4

Evidence of inter-species transmission between cattle and badger populations residing within and surrounding Woodchester Park

4.1 Introduction

Mycobacterium bovis infection of cattle populations remains a problem across large parts of the United Kingdom [220]. Surveillance of bTB in the UK uses a cattle test and slaughter regime, combined with movement controls and abattoir surveillance [57]. Despite similar regimes being successful elsewhere [9, 4], including in Scotland [10], in parts of Northern Ireland, Wales and England *M. bovis* infection of livestock remains endemic, and regular surveillance is necessary to control spread and reduce health and economic burdens [4, 199, 57]. Control is particularly necessary in the southwest of England, a high prevalence endemic bTB area whose range is expanding and where prevalence is increasing [220].

Although cattle are the principal host for *M. bovis*, it has an extremely broad host range and a large number of mammals in the UK are susceptible to infection [11]. This broad host range enables multi-host systems to develop and these greatly hinder eradication programs [12, 13, 221, 222, 126, 58]. A multi-host system, a system in which a pathogen infects multiple host species, is important when the different hosts play a role in the spread and persistence of the pathogen [123]. If the pathogen can survive in a multi-host system, control operations in one species may be hindered by re-infection through spill-over of infection from a different species. When each species is capable of maintaining infection independently, control operations in one species can be rendered ineffective as a result of spill-over events. In the UK, the European Badger (*Meles meles*), a protected species [66, 17], is susceptible to *M. bovis* infection [11, 144]. In order for control operations in badger populations to be a necessary part of bTB control in the UK, badger populations need to be capable of both maintaining *M. bovis* infection independently, and passing the infection into cattle populations. Understanding the role of this species in the spread and persistence of bTB in cattle populations

has been the aim of a number of sanctioned culling trials with often apparently contradictory outcomes [157, 158, 159, 161]. While the outcomes of these trials have been controversial, these trials were able to demonstrate that, where re-population of an area was restricted, reduction in badger population densities resulted in decreased bTB prevalence in associated cattle herds. The current use of badger control operations in the national bTB control scheme of the Republic of Ireland is supported by the results of the culling trials that were completed in this country [33, 160].

A variety of molecular typing methods have further informed the investigations into the role of badger populations in bTB epidemiology in the UK and Republic of Ireland. Sympatric cattle and badger populations have been shown to carry similar *M. bovis* strains [223, 224, 22]. This similarlity was demonstrated, using methods such as Spoligotyping [225] and Variable Number Tandem Repeat typing [103], which use annotated repeat regions (one or multiple, respectively) of the *M. bovis* genome to characterise strains. Unfortunately the resolution of these methods is such that it is difficult to quantify the direction of inter-species transmission, since isolates taken over large spatial and temporal ranges can be indistinguishable according to these techniques [112]. Next Generation Sequencing technologies have allowed Whole Genome Sequencing (WGS) to become a feasible epidemiological tool to study bTB inter-species transmission [111, 121]. Biek *et al.* [24] were able to demonstrate high genetic similarity between spatio-temporally proximate infected cattle and badgers, and inferred recent transmission between these populations. However, despite the high resolution of WGS data, it wasn't possible to assign a direction of transmission, since only a small number of samples from the badger population were available. Earlier this year, Glaser *et al.* [26] identified transmission between cattle and deer populations in Minnesota by combining WGS with detailed epidemiological data. Although evidence of within- and between-species transmission was provided, the extent and direction of inter-species transmission wasn't quantified.

A naturally infected badger population residing in Woodchester Park, a National Trust owned park in the southwest of England, presented the ideal opportunity to investigate inter-species transmission of bTB. Badgers in Woodchester Park have been routinely captured for more than 30 years [185]. This densely sampled longitudinal dataset provides a rich source of data regarding the dynamics of the host population. A similar resolution of data was available for the cattle population in Great Britain, in the Cattle Tracing System (CTS) and SAM (system for recording bTB testing results) databases, which provide the movements and bTB test history of every cow in Great Britain, respectively [226]. The current research investigates WGS data from isolates sourced from badgers and cattle living within and around Woodchester Park that were found to be infected with *M. bovis*. In contrast to previous WGS research of bTB in the UK, at least 100 isolates were available from both cattle and badger populations living in close proximity. In addition, these data were accompanied with detailed life history information describing both the sampled cattle and badger populations.

The high resolution genetic data were examined in the context of each sampled animal's life history, to determine whether the sampled high density badger population was capable of maintaining *M. bovis* infection independently. In addition, the current research aimed to reveal and determine the direction of inter-species transmission events between the sampled cattle and badger populations.

4.2 Materials and Methods

4.2.1 Woodchester Park Surveillance

A long term field study has been on-going at Woodchester Park since 1977 [185]. Woodchester Park is home to a high density badger population that is naturally infected with *M. bovis*. The study at Woodchester Park involves trapping operations, within annually delimited badger group territories, four times a year. All badgers captured are given a unique identifying tattoo. In addition, each time a badger is captured, samples are taken of blood, urine, faeces, tracheal and pharyngeal aspirates, and swabs of any wounds or abscesses are collected. If, during the capture operations, a dead badger is found, a post-mortem examination is carried out and samples are collected from major organs and lymph nodes. Wilkinson *et al.* [146] estimated the annual badger capture rate to be 85% of the population, therefore the capture data represents an exceptionally dense record of the population and infection dynamics within Woodchester Park. The blood, urine and sputum samples are tested to ascertain *M. bovis* infection using the tests described in Table 4.1.

Table 4.1: A description of the different tests used on the samples collected from badgers, which are captured during the on-going field study at Woodchester Park. The test sensitivity and specificity estimates were taken directly from the literature [187, 20, 185]

Test	Description	Sensitivity	Specificity	Year Implemented
Bacteriological culture	Urine and sputum samples are cultured for the presence of *M. bovis* bacteria. If present, the bacteria were VNTR typed and spoligotyped [225, 103].	8%	99.8%	1982
Brock Test ELISA	An *in vitro* blood test, blood samples are exposed to an *M. bovis* antigen to test for for the presence of specific antibodies [187].	37%	98%	1982
γ-IFN test	An *in vitro* blood test, blood samples are exposed to either avian or bovine tuberculin and the amount of γ-IFN produced is quantified using an Enzyme-Linked Immunosorbent Assay (ELISA) with relative reseponses to the two types of tuberculin indicating infection status [42].	79.9%	95%	2006
Stat-Pak test	A serological test utilising Lateral Flow Immunoassay (LFI) technology. Specific antigens bound to a medium visibly react to specific antibodies present in a blood sample [188].	79.9%	95%	2006

The capture dataset from Woodchester Park, describing the timings and locations of trappings as well as each badger's age, sex and test results, were used to assess the infection status of each sampled badger. Two hundred and thirty badger isolates were cultured from samples taken from 84 infected badgers, captured during routine trapping operations.

4.2.2 Cattle Population Surveillance

In the southwest of England, where Woodchester Park is situated, cattle herds are tested for bTB on an annual basis as part of routine bTB surveillance, using the Single Intradermal Comparative Cervical Tuberculin test (SICCT). For the SICCT, an *M. bovis* derived tuberculin and an *M. avium* derived tuberculin are injected, and the hypersensitivity reactions are compared to establish whether the cow is infected. In addition to the SICCT, abattoir surveillance is part of routine bTB surveillance in the UK [57]. Each animal tested or examined (with the SICCT or during abattoir surveillance, respectively) is recorded in the SAM database. The

SAM database was established last year and provides a computerised version of the bTB cattle testing data, previously recorded in VETNet - the national disease database. In addition to recording testing histories, the movements of every cow in the UK are recorded in the CTS database, which was established in 1998 [226]. Cattle herds surrounding Woodchester Park have a history of bTB (Figure 4.2), and 100 cattle isolates were cultured from 100 different infected cattle found on 58 farms within 15km of Woodchester Park. The movement data in the CTS database isn't complete, movements are sometimes missed or recorded incorrectly. In addition, 42 of the cattle were sampled before 2005, when the cattle movement data was less complete [226]. Figure 4.1 shows the range of herds sizes on herds around Woodchester Park.

For each cattle farm in the UK, a single point location (Eastings and Northings) was available for the farm buildings. The point location information was used to identify cattle farms surrounding Woodchester Park. Although specific land parcel data (describing the outlines of each field associated with a cattle farm) does exist, it wasn't available for the current analyses. Therefore, the current analyses assume that these point locations are an accurate representation of where the cattle associated with each farm reside.

4.2.3 Sampling and Isolate Preparation

4.2.3.1 Badger Isolates

All isolates that were cultured from samples taken from infected badgers, captured in Woodchester Park, were archived at the Food and Environment Research Agency (FERA) in York. Two hundred and thirty of these *M. bovis* isolates were successfully re-cultured. Re-culturing and DNA extraction were carried out at the the Agri-Food and Biosciences Institute in Northern Ireland (AFBNI). The archived isolates were grown on Löwenstein-Jensen medium (LJ) slopes to form single colonies. DNA was extracted from these single colonies using standard high salt hexadecyl trimethyl ammonium bromide (CTAB) and solvent extraction protocols [183, 91]. The extracted DNA of each isolate was sequenced at the Glasgow Polyomics facility using an Illumina MiSeq platform, which produced 2 x 300bp paired end reads.

4.2.3.2 Cattle Isolates

The 100 cattle isolates originated from cattle, which were determined to be infected either through use of the SICCT or during abattoir surveillance (Figure 4.2). These isolates were selected between 1997 and 2012; over this period there were 1012 confirmed breakdowns[1] involving 8750 reactors (of which 2156 were confirmed to be *M. bovis* with culture) on 369 (of 749) cattle farms within 15km of Woodchester Park. Each of these reactor cattle were subject to a post-mortem examination following slaughter, and *M. bovis* was cultured from suspect

[1] A herd breakdown is when a cow on the herd has reacted positively to the SICCT, and its infection has been confirmed during a post-mortem exminmation or through *M. bovis* culture.

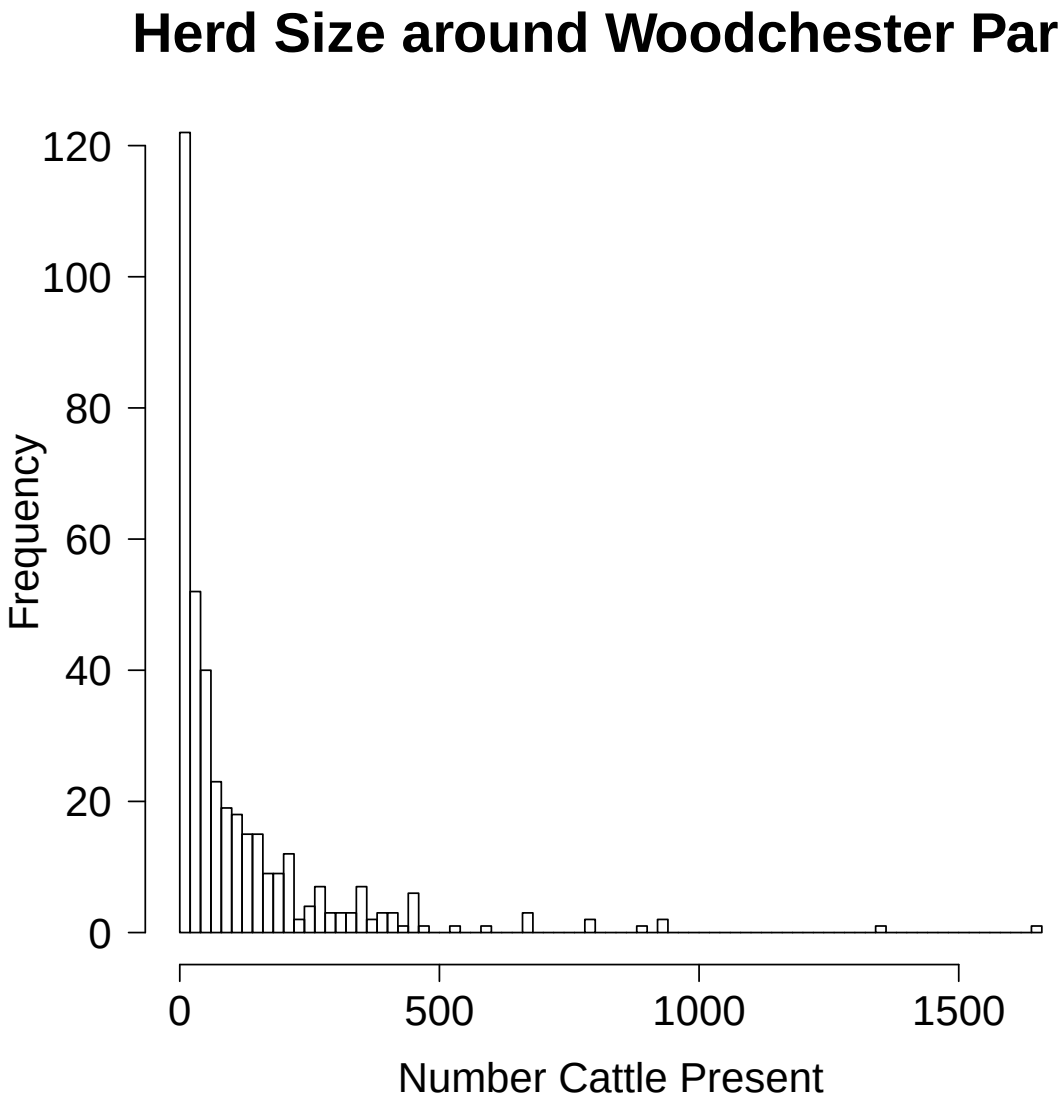

Figure 4.1: The number of animals present on herds with 15km of Woodchester Park in 2012.

granulomatous tissue. Cultures were grown and confirmed to be *M. bovis* using the same standard protocol described above. Cultured isolates were then archived by the APHA. The 100 cattle isolates used for the current research were selected from the archive, re-cultured and then DNA was extracted from the cultures. Re-culturing and DNA extraction were conducted, as above, by APHA in York. The extracted DNA was sequenced at APHA's central sequencing unit in Weybridge, on an Illumina HiSeq platform producing 2 x 150bp paired end reads. The isolates were selected to be from cattle that were infected on beef and dairy herds within 15km of Woodchester Park between 1997 and 2012. The selection procedure aimed to have samples from herds surrounding Woodchester Park over as broad a timeframe as possible.

4.2.4 Processing Whole Genome Sequences

The raw reads of all the isolates sourced from infected badgers and cattle were examined using FASTQC (v.0.11.2 - [189]). Poor quality ends that were identified in FASTQC, were trimmed using PRINSEQ (0.20.4 - [191]). Adapter sequences, where present, were removed using TRIMGALORE (v0.4.1 - [190]). The trimmed reads were aligned to the *M. bovis* reference genome, AF2122/97 [102], using the Burrows-Wheeler Alignment tool [192, 176]. Information for each site on the reference genome was collated across the isolates. Any sites that were found within the regions encoding the Proline-Glutamate (PE) and Proline-Proline-Glutamate (PPE) genes or annotated repeat regions were removed [193]. Only sites that varied amongst the isolates (variant positions) were considered in further analyses.

High quality variant positions were selected based on the Mapping Quality (MQ), High Quality base Depth (HQDP) and Read Depth (DP). Filters were created using those quality metrics as well as the support for the allele called (SUP), the site coverage across the isolates (COV), and the minimum number of positions separating variant positions (PROX). The following filters were used: MQ$\geq$30, HQDP$\geq$4, DP$\geq$30, SUP$\geq$0.95, COV$\geq$0.1, and PROX=10.

During a preliminary investigation of the isolates sourced from badgers, it was determined that five were mislabelled. The extent of the mislabelling was examined, in the context of the detailed capture data available for the sampled badgers. Eight additional isolates were identified as being potentially mislabelled. The 5 mislabelled and 8 potentially mislabelled isolates were removed, and not considered in any further analyses. Please refer to Appendix B for a description of the investigation undertaken to determine the presence and extent of the mislabelling. Isolates from badgers that failed the quality control levels (genome coverage $\geq$90%) established for the bioinformatics pipeline, were removed (n=54) from any further analyses. It wasn't possible to include these isolates in the mislabelling investigation and, therefore, it wasn't possible to determine whether these were mislabelled. Only those isolates sourced from cattle that had <10% of the variant positions with sufficient coverage were removed (n=19). A lower threshold was used for the cattle isolates to retain as many of them

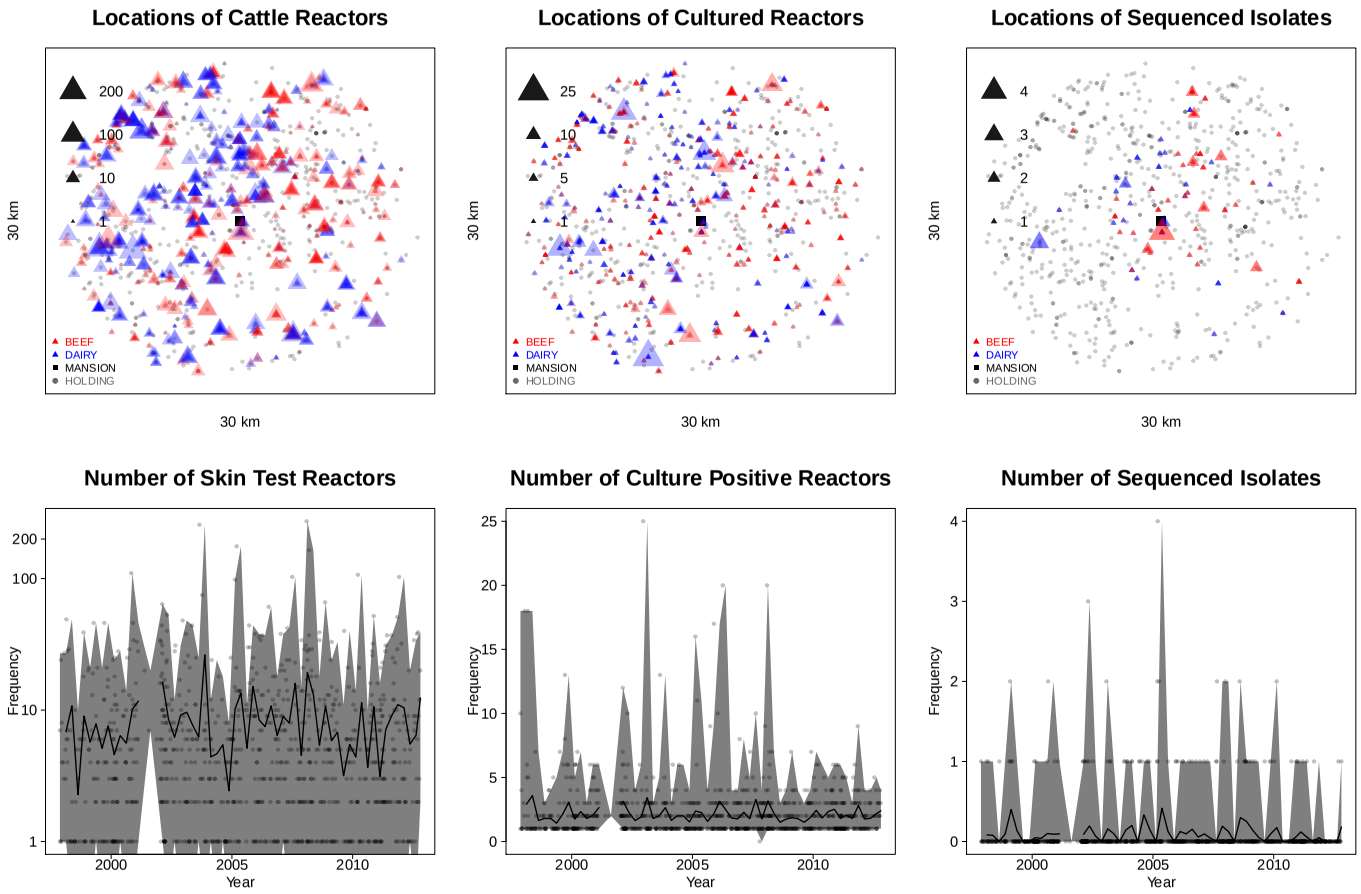

Figure 4.2: Figures describing where, when and how many reactors (LEFT), confirmed (cultured) reactors (MIDDLE) and sequenced isolates (RIGHT) were present on the cattle farms surrounding Woodchester Park from 1997 to 2012. These data were taken from the SAM database and reflect the confirmed breakdown records. The number of reactors and cultured reactors represent a cumulative total over the time course of a confirmed breakdown. TOP: The spatial locations of confirmed breakdowns. BOTTOM: When the confirmed breakdowns began against their eventual size. The mean breakdown size during a 3 month window is plotted as a dark line. A grey polygon spans the upper and lower limits within each 3 month window.

as possible. For the remainder of the current chapter, a sampled animal refers to an animal from which *M. bovis* was cultured, sequenced and selected for the current analyses.

4.2.5 Intraspecies Investigation of Inter-Isolate Genetic Distances of Isolates from Badgers

This section describes an analysis that compares the inter-isolate genetic distances to epidemiological metrics describing the relationships between the sampled badgers based upon spatial, temporal and network patterns. If the sampled badger population, present in Woodchester Park, was capable of maintaining *M. bovis* infection, then the infection must be spread through the movements and contact patterns of the badgers. We therefore expected there to be a relationship between the inter-isolate genetic distances and sampled animal's life histories. For example, *M. bovis* isolates from two badgers that have lived their entire lives in the same badger group at the same time might be expected to be more similar than isolates taken from badgers that have never lived in the same group and weren't alive at the same time.

For each inter-isolate comparison a genetic distance (the number of sites that differ) was calculated, and the isolates were compared using the following epidemiological metrics:

1. Isolates taken from the same badger (yes/no).

2. Isolates taken from badgers with the same main/sampled/infected group (yes/no). The main group was defined as the group that the badger spent the majority of it's recorded life in. The sampled group was defined as the group that the badger was captured in when it was sampled for the current isolate. The infected group was defined as the group that the badger was captured in when it's infection was first detected.

3. Spatial distance (km) between the main/sampled/infected groups of the sampled badgers for the isolates being compared.

4. The number of badgers that have dispersed between their main/sampled/infected groups of the sampled badgers, for the isolates being compared.

5. Shortest path length between the main/sampled/infected groups of the sampled badgers, for the isolates being compared. The shortest path was calculated on a weighted adjacency matrix that recorded the number of badgers that dispersed between social groups. The shortest path, on this adjacency matrix, between two groups was calculated using Dijkstra's algorithm [196].

6. The mean number of badgers dispersing along the edges of the shortest path between main/sampled/infected groups.

7. The number of badgers captured in both of the main/sampled/infected groups of the sampled badgers, for the isolates being compared.

8. The number of days overlap of the lifespans of each isolate's sampled badgers.
9. The number of days overlap of the infected lifespans of the sampled badgers associated with the isolates being compared. A badger's infected lifespan started on the day of its first positive test and ended with its last capture.
10. The number of days that the sampled badgers, of the isolates being compared, spent in the same group together.
11. The number of days between the infection detection dates of the sampled badgers, for the isolates being compared.
12. The number of days between the sampling dates for the isolates being compared.

The network, upon which the network based epidemiological metrics were calculated, was built using the dispersal events that were recorded in the capture data. In this network, badger social groups were treated as nodes. When a badger was captured in a group that was different to the group it was previously captured in, this was treated as a dispersal event and an edge between the social groups involved was added to the network. The number of dispersal events across each edge was also recorded.

Genetic distances between the high quality isolates sourced from badgers, were calculated and compared to epidemiological metrics. The comparison between genetic and epidemiological data was completed using a Random Forest regression model [197], which fitted the epidemiological metrics to the corresponding genetic distances. The same fitting procedure was also completed using a Boosted Regression model [227]. The model fitting was completed in the statistical programming environment R (v3.2.1 [194]). Genetic distances <15 SNPs were used for these comparisons to avoid fitting very large genetic distances, which wouldn't relate to the epidemiological information available, since these data correspond to the recent dynamics in Woodchester Park. The 15 SNP threshold was chosen to select the inner part of the bimodal inter-isolate genetic distance distribution, thereby retaining within-, and ignoring between-, clade distances.

The Random Forest and Boosted Regression models are tree-based machine learning algorithms. In each, a large number of decision trees are built using explanatory variables (epidemiological metrics) and a response variable (inter-isolate genetic distances) [197, 227]. Each node of a decision tree splits the data based upon an explanatory variable and its threshold. The choice of which explanatory variables and their thresholds at each node, are done to minimise the predictive error of the tree (the difference between actual and predicted genetic distances). To model the continuous inter-isolate genetic distances, regression models were necessary. In the current research the decision trees represented regression trees. Regression trees, instead of defining response categories at the tips of the tree, as a classification tree would, provide a single predicted response value at each tip of the tree. These single

predicted values are calculated as the mean of the response values that remain at each tip, following the splits based upon the criteria at the previous internal nodes. Decision tree based algorithms were selected in the current research because of their flexibility and insensitivity to uninformative or highly correlated explanatory variables [197, 227]. The explanatory variables considered here, the epidemiological metrics, were often highly correlated and highly variable in terms of their informativeness.

Although both the Random Forest and Boosted Regression algorithms are based upon the same decision tree structures, they differ in other respects. The Random Forest algorithm builds a forest of random, independent, decision trees [197]. The final predictive Random Forest model aggregates the results of each of the trees in its forest. The Boosted Regression algorithm also builds a large number of decision trees but, in contrast to the Random Forest model, these trees aren't independent - each tree is built upon the residuals of the previous trees [227]. This iterative process is designed such that the model is continually improved, by targeting what is being poorly predicted by the current combination of decision trees. Both the Random Forest and Boosted Regression algorithms were used for the current research so as to increase the robustness of the analyses by using two different approaches.

4.2.6 Interspecies Phylogenetic and Epidemiological Investigation of Cattle and Badger Isolates

Following the examination of each isolate's genome sequence, labelling and sequencing quality, 163 and 81 isolates sourced from infected badgers and cattle, respectively, were available. Fifty-four of the 81 cattle isolates had genome coverage $\geq$90%. A maximum likelihood phylogenetic tree was constructed in the program PHYLIP (v3.695 - [204]) and rooted against the reference sequence - AF2122/97. Using this tree it was possible to identify potential inter-species transmission events by examining which isolates were closest to the root of each clade. For example, if a clade contained closely related isolates from badgers and cattle, and an isolate from a cow was closest to root of the clade then that was taken as an indication of a potential transmission event from cattle into badgers. Clusters of highly genetically related isolates were defined surrounding the potential inter-species transmission events (Figure 4.6). For each cluster, the life histories of the associated sampled badgers and cattle were interrogated to investigate whether there was additional evidence available to determine the direction of transmission by providing information on the potential temporal sequence of events that led to infection.

For each of the clusters defined in the phylogenetic tree, the life histories of the sampled animals were interrogated. Here, life history refers to the recorded data regarding the capture/movement history of cattle/badgers surrounding/within Woodchester Park and the dates and results of their bTB tests. The sampled animals were linked to their respective databases (CTS for cattle, and Woodchester Park capture data for badgers) via an eartag number (for

cattle) or a unique tattoo (for badgers). An animal's observed lifespan started when they were first recorded in their respective databases and ended on the date of their last observation. Over the course of each sampled animal's observed lifespan, the dates of bTB testing and sampling events were examined. In addition to interrogating the life histories of the sampled animals, the histories of in-contact animals were investigated. An in-contact animal was defined as an animal that lived in the same herd/social group at the same time as a sampled animal.

The bTB tests (Table 4.1) conducted on badgers in Woodchester Park resulted in a positive or negative reaction. Cattle tested using the SICCT were defined as reactors (positive), inconclusive reactors or negative. Inconclusive reactors were those animals where a difference between the hypersensitivity reactions to injected bovine and avian tuberculins was evident, but not above the threshold used to define reactors.

4.3 Results

4.3.1 Intraspecies Investigation of Inter-Isolate Genetic Distances of Isolates from Badgers

Both the Random Forest and Boosted Regression models were able to accurately predict the inter-isolate genetic distances using the available epidemiological metrics. A Pearsons correlation between the predicted and actual distances produced a value of 0.85 for the Random Forest model and 0.91 for the Boosted Regression model.

The Random Forest and Boosted Regression models ranked each of the epidemiological metrics, by their informativeness in the fitting procedure. The Random Forest algorithm measures the informativeness of predictor variables by how much the prediction error of the model increased, when the values of each of the predictor variable were randomly permuted [197]. The Boosted Regression model measures the informativeness of predictor variables by examining the number of times each variable is chosen at each node of each decision tree to split the data, and by how much the model prediction is improved as a result of that variable being used [227]. There was good agreement between the predictor variable rankings by the Random Forest and Boosted Regression models (Figure 4.3).

Almost all the epidemiological metrics were informative in describing the variation in the inter-isolate genetic distances (Figure 4.3). The temporal variables, describing the difference in days between the sampling and infection dates, were the most informative. The next best variables were spatial and network based metrics: the spatial distance between and the number of badgers that lived on both of the main groups of sampled badgers for the isolates being compared. The least informative variables were those with binary answers, regarding whether the isolates came from the same badger or whether the sampled badgers were infected or sampled in the same group. In general, variables that referred to the sampled badger's main group were more informative than those that referred to the infected or sampled groups. The main group for each badger was defined as the social group that the badger spent the most time in. The variables that counted the number of badgers dispersing between social groups were less informative than those that noted how many badgers lived on both the groups being compared. By examining the number of badgers that lived on both, rather than the number that dispersed between, badger groups, both indirect and direct movements were accounted for. The variables that included the mean number of badgers dispersing along the shortest path between badger social groups, in comparison to those that only measured the length of the shortest path, were more informative.

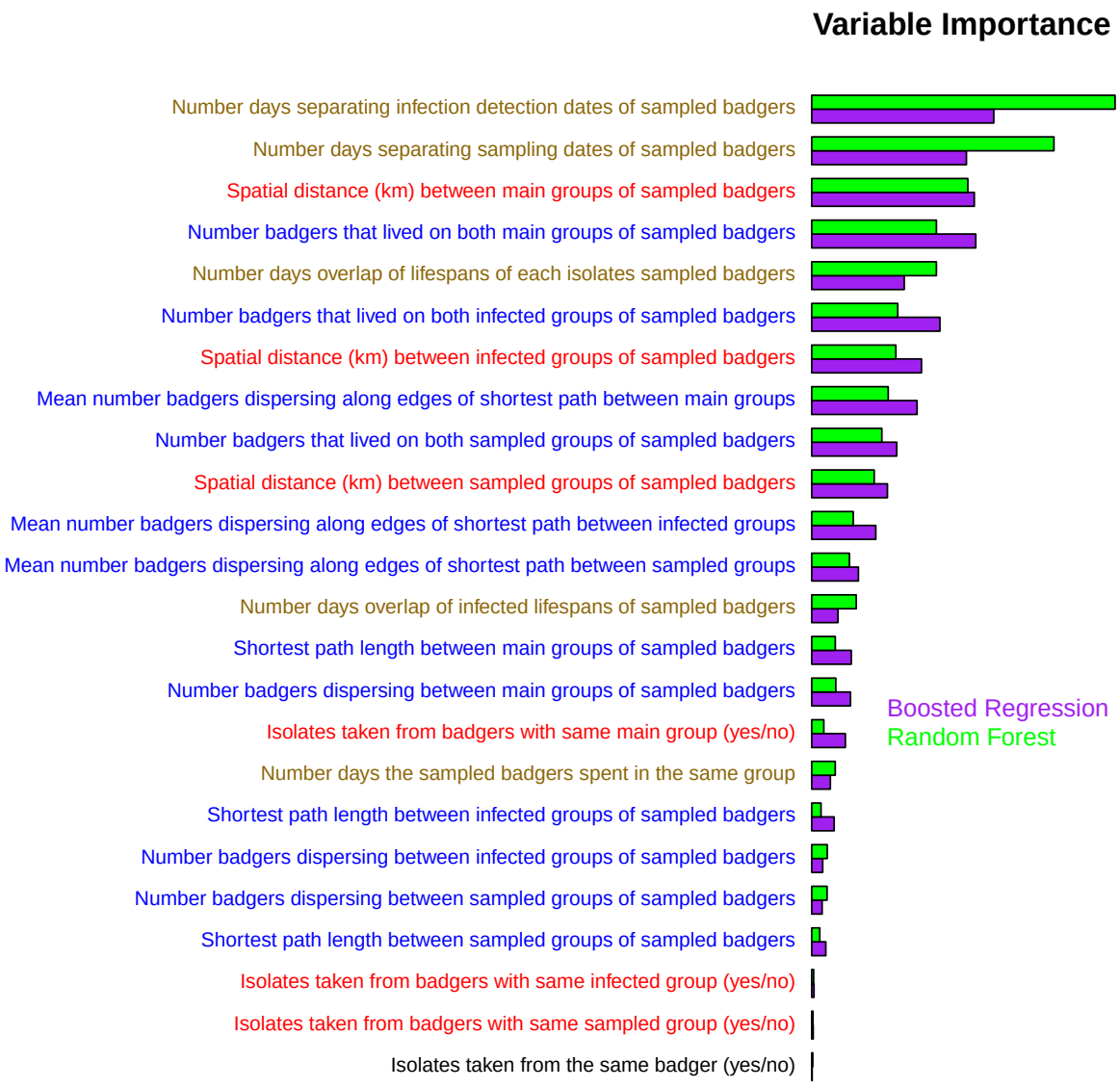

Figure 4.3: The relative importance of the epidemiological metrics used in the Random Forest (green) and the Boosted Regression (purple) models. The metrics were coloured if they were based on temporal (gold), spatial (red) or network (blue) data, otherwise they were left black.

4.3.2 Interspecies Phylogenetic and Epidemiological Investigation of Cattle and Badger Isolates

When clades were defined by a 15 SNP threshold, 10 separate clades could be defined (Figure 4.4). These clades demonstrated a limited degree of spatial clustering. The cattle associated with the red clade (n=8), which contained 131 badger isolates, were found on farms that were <5km from Woodchester Park. Similarly, the pink clade (contained 15 badger and 3 cattle isolates) was very closely associated with Woodchester Park. In contrast, for the blue, orange and cyan clades, which contained only cattle isolates (n=6, n=7, and n=3, respectively), no cattle were sampled on farms closer than those sampled for the isolates from the pink and red clades. The isolates sourced from cattle were considerably more diverse than those from badgers. The majority of the isolates sourced from badgers were situated in a dense highly genetically similar clade - red in Figure 4.4. In contrast, the isolates from cattle were found in several highly distinct clades.

Using the phylogeny it was possible to identify clusters where inter-species transmission events may have recently occurred, by identifying cattle and badger isolates that were highly genetically similar (Figure 4.5). A number of the lower quality isolates also fell within the defined clusters. The presence of these isolates within the clusters should be stable but the length of the branches joining these isolates to the phylogeny must be considered unknown. In clusters 0, 1, and 2, the presence of a number of cattle isolates towards the root of these clusters suggested that the outbreaks represented by these sampled clusters may have begun in the cattle population. In contrast, a larger number of badgers were associated with clusters 3 and 4, and the isolates sourced from cattle were fewer and nested within groups of isolates from badgers. A zoomed in version of the clusters defined in Figure 4.5 is shown in Figure 4.6.

4.3.3 Interrogating the Sampled Animals in the Defined Clusters

The recorded movements and capture events of the sampled and in-contact cattle and badgers associated with each cluster were interrogated. In addition, the bTB testing histories for each sampled animal were examined. A number of metrics were recorded during the interrogation of the sampled animals involved in each cluster, these were recorded in Table 4.2. This table provides a summary of the figures that follow it.

In cluster 0 (Figure 4.7), all four badger isolates came from a single animal during its only capture event. The two cattle isolates came from two different cows. For one of the sampled cattle, it wasn't possible to link their eartag to any movements in the CTS database, although test data was available. Both cows were present and infected after the sampled badger. Four hundred and eighty three unsampled in-contact cattle were associated with the sampled cow for which movement data were available. Fifty-one of these in-contact cattle reacted to the SICCT test (Table 4.2). The two sampled herds (herds that the sampled cattle were taken

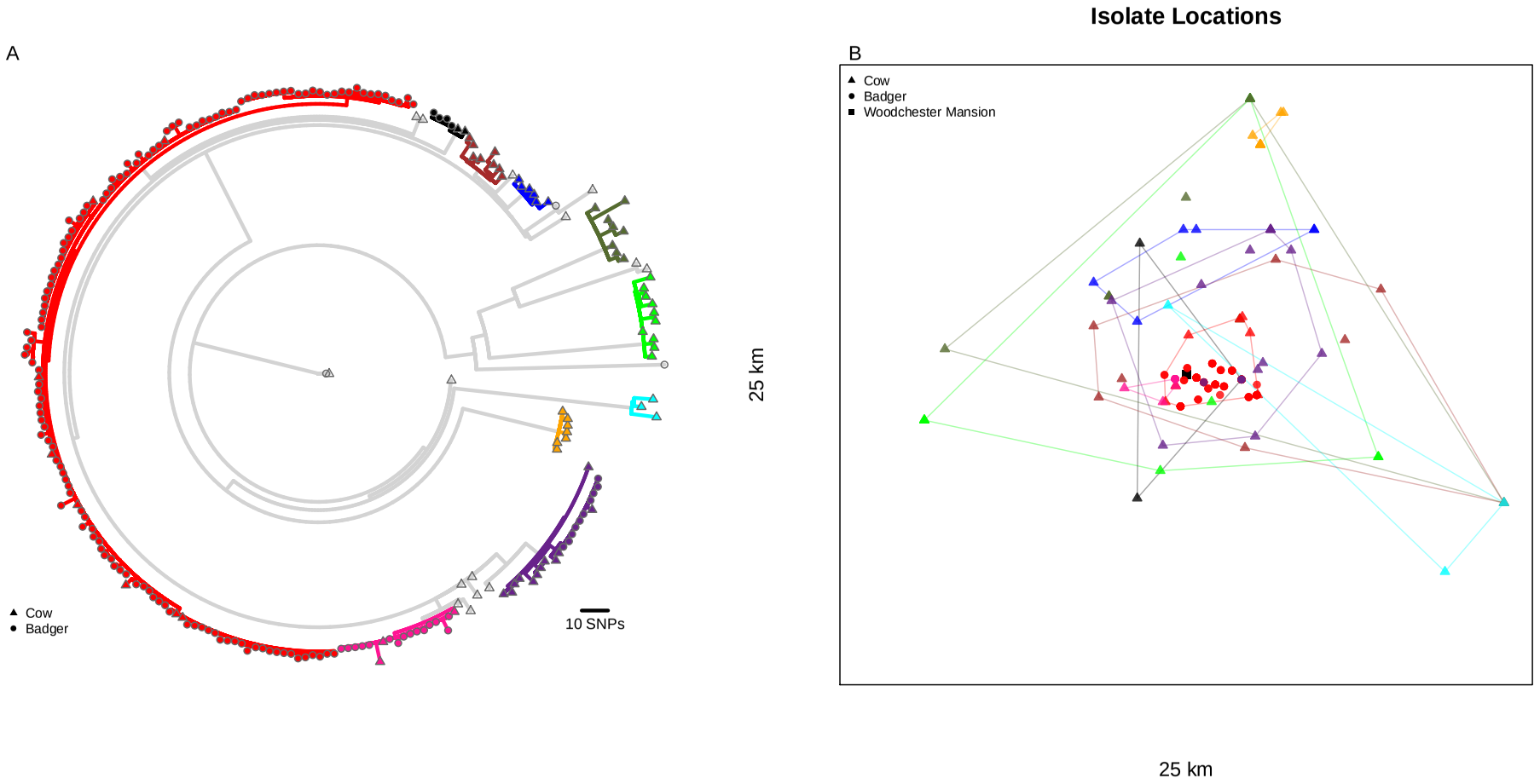

Figure 4.4: A: Maximum likelihood phylogenetic tree built using PHYLIP [204], rooted against the *M. bovis* reference genome, AF2122/97 [102]. Clades containing more than 3 isolates that were less than 15 SNPs from their root node were coloured. B: The spatial locations of the isolates with their colours relating to the clades they're found in. A polygon was drawn around each coloured set of isolates such that they were all contained within or on it. Isolates from cattle and badgers are represented by triangles or circles.

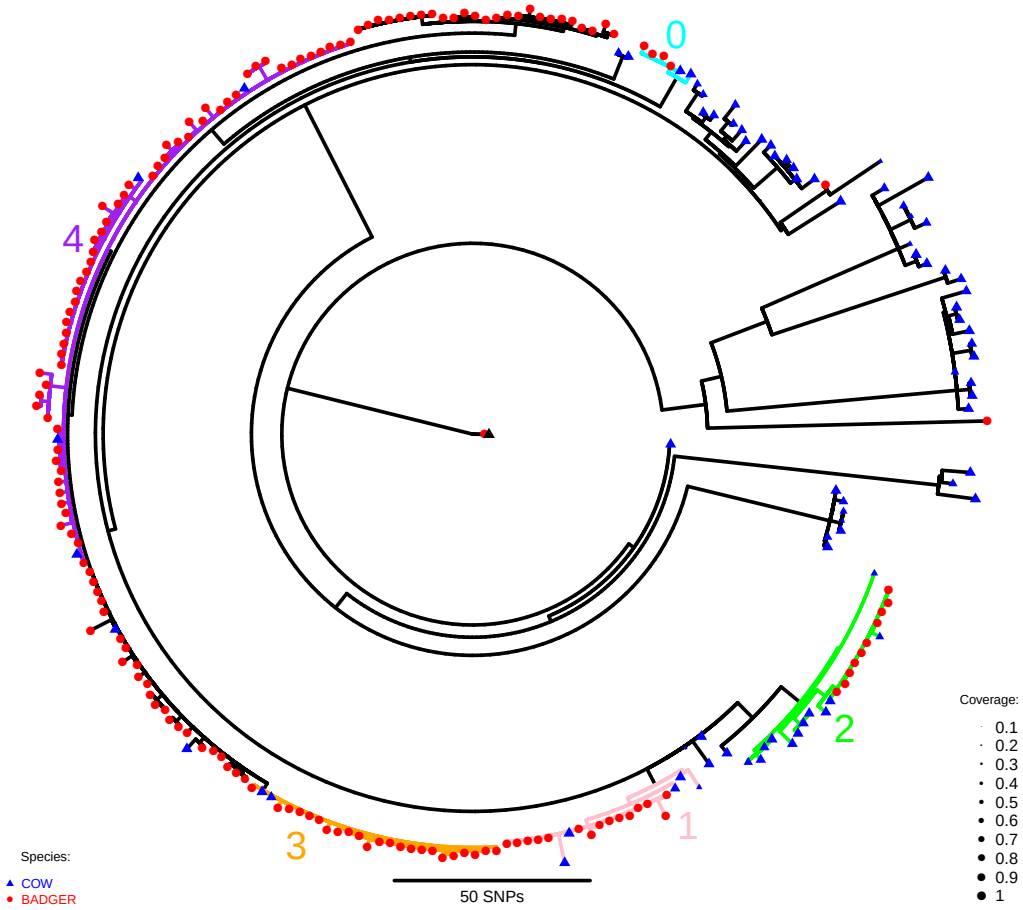

Figure 4.5: A maximum likelihood phylogenetic tree built using PHYLIP [204], rooted against the *M. bovis* reference genome, AF2122/97 [102]. Assigned clusters were coloured as follows: cluster 0 = cyan, cluster 1 = pink, cluster 2 = green, cluster 3 = orange and cluster 4 = purple. The reference sequence was represented by a black triangle. The sizes of the tips on the phylogeny were scaled according to the sequencing quality as determined by the proportion of the reference genome that had sufficient coverage (DP≥20). Isolates from cattle and badgers are represented by blue triangles or red circles.

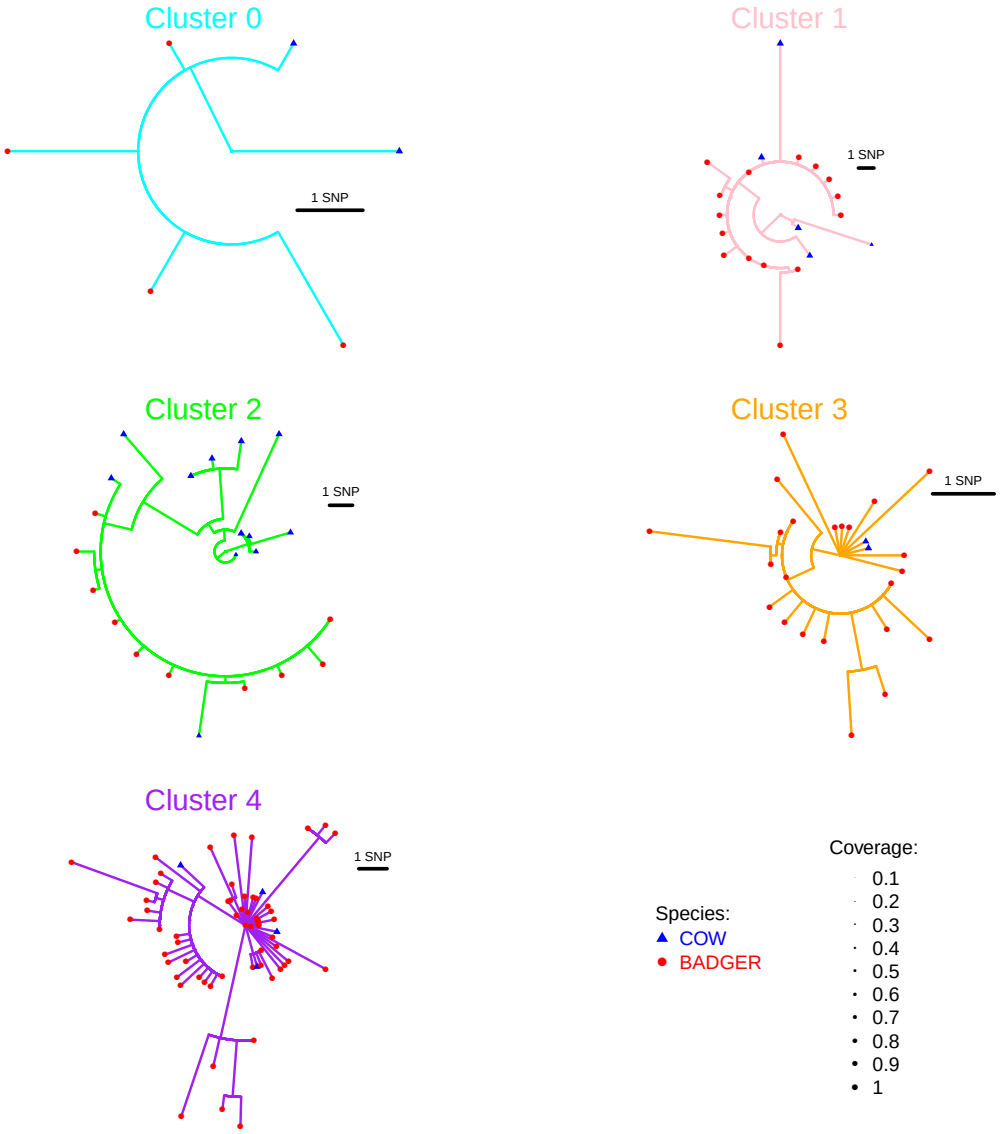

Figure 4.6: The phylogenetic relationships between the *M. bovis* isolates in the clusters described in Figure 4.5. The sizes of the tips on the phylogeny were scaled according to the sequencing quality as determined by the proportion of the reference genome that had sufficient coverage (DP≥20). Isolates from cattle and badgers are represented by blue triangles or red circles.

Table 4.2: A summary of the defined clusters described in Figure 4.5. Where necessary, values were rounded to 2 significant digits. The rows are coloured by whether they refer to temporal (gold), genetic (grey), spatial (red) or network (blue) characteristics.

	Cluster-0	Cluster-1	Cluster-2	Cluster-3	Cluster-4
Number of badgers sampled	1	6	5	13	36
Number of cattle sampled	2	5	12	2	4
Number of in-contact badgers that tested positive	3	50	18	72	151
Number of in-contact cattle that tested positive	51	91	96	11	20
Number of in-contact badgers that NEVER tested positive	2	28	19	78	132
Number of in-contact cattle that NEVER tested positive	434	789	938	33	311
Earliest date that a sampled badger tested positive	19-9-2000	24-5-2000	9-12-2002	1-11-2000	9-9-1997
Earliest date that a sampled cow tested positive	5-2-2002	2-4-2007	2-7-2002	7-5-2002	12-3-2002
Earliest date that an in-contact badger tested positive	9-12-1997	1-8-1989	4-12-1995	31-1-1995	26-11-1991
Earliest date that an in-contact cow tested positive	27-9-2004	30-3-2004	1-7-2002	9-8-2005	7-5-2002
Minimum patristic distance (SNPs) of the sampled badgers to the MRCA of cluster	1.89	3.41	5.39	0.47	0.0
Minimum patristic distance (SNPs) of the sampled cattle to the MRCA of cluster	1.89	1.42	0.47	0.47	1.14
Mean spatial distance (KM) from the sampled herds to Woodchester Park	6.0	2.52	5.01	2.9	1.94
Mean number of movements of sampled cattle to or from the sampled herds	1.5	1.6	1.67	1.0	1.0
Mean number of movements of in-contact animals that tested positive to or from the sampled herds	51.0	35.2	11.0	9.5	0.75
Mean number of movements of in-contact animals that NEVER tested positive to or from the sampled herds	433.0	303.0	126.89	31.0	55.75

from) were approximately 6km from Woodchester Park (Figure 4.8).

In cluster 1 (Figure 4.9) the isolates were taken from six badgers and five cattle that lived between 1998 and 2011. Infection was detected in three badgers before it was detected in the sampled cattle. There were 880 in-contact cattle that encountered the sampled cattle, 91 of these cattle reacted positively to the SICCT (Table 4.2). Those in-contact cattle that reacted to the SICCT, did so between 2002 and 2014. The sampled badgers encountered 53 (of 81) unsampled in-contact badgers that reacted positively to bTB testing at some point in their lives and the infection was detected in 30 of these in-contact badgers before 2002. The sampled cattle with strains associated with cluster 1 were sampled in five different herds within 5 km of Woodchester Park (Figure 4.10). These sampled herds were associated with an average of 2-3 recorded movements of the sampled cattle associated with cluster 1 (Table 4.2). Three of these sampled herds were directly linked by recorded cattle movements. The sampled herds were, on average, involved in almost 60 recorded movements of unsampled in-contact cattle that reacted positively to the SICCT. The sampled badgers lived in 3 different social groups, two of which were directly connected through the recorded movement of sampled animals.

In contrast to those animals associated with cluster 1, in cluster 2 (Figure 4.11) infection was first detected in a sampled cow in July of 2002. The lifespans of the five badgers and 12 cattle associated with cluster 2, spanned from 2001 to 2011. 96 of the 1031 in-contact cattle, which encountered the sampled cattle from cluster 2, reacted positively to the SICCT (Table 4.2). 47 of the 96 in-contact cattle had reacted positively to the SICCT by the end of 2004. The sampled cattle from cluster 2 were sampled on nine different farms that were an average of 5km from Woodchester Park (Figure 4.12). These nine herds were involved with an average of five recorded movements of sampled cattle, and 33 recorded movements of in-contact animals that reacted positively to the SICCT. 18 of the 37 in-contact badgers found, tested positive for bTB and two of these badgers were present in 1997. The sampled

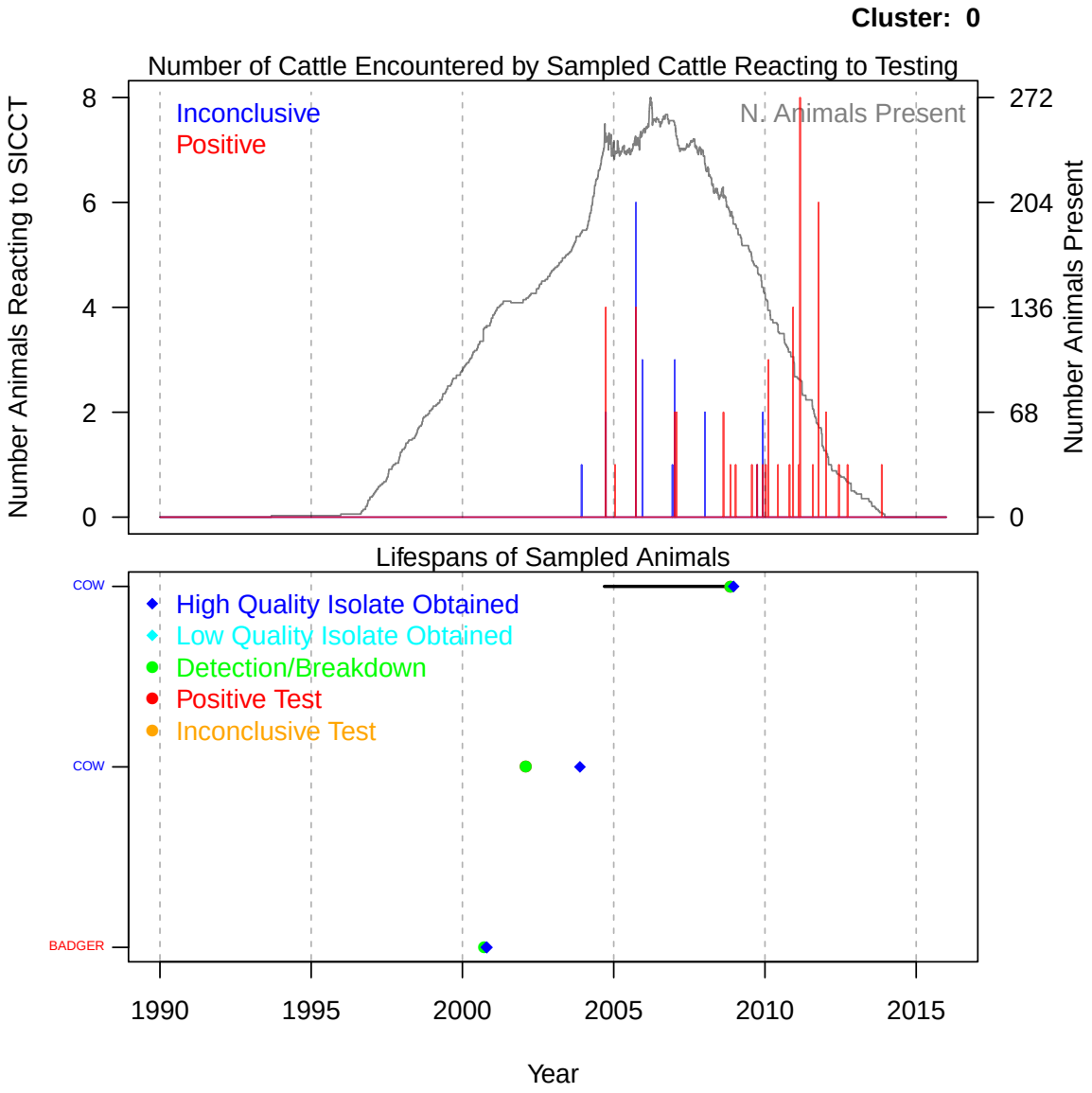

Figure 4.7: The observed lifespans and testing history of the sampled cattle and badgers from cluster 0 and of the associated unsampled in-contact animals. BOTTOM: The observed periods available for each sampled animal and the dates of their bTB testing and sampling events. MIDDLE: A summary of the unsampled in-contact cattle. The grey line represents the number of unsampled in-contact cattle alive on each date between 1990 and 2015. The red and blue lines show the numbers of test reactors (inconclusive in blue and reactors in red) on each date. TOP: A summary of the unsampled in-contact badgers. The total number of unsampled in-contact badgers alive between 1990 and 2015 is represented by a grey line, the number of those in-contact badgers that tested positive is shown by the red line.

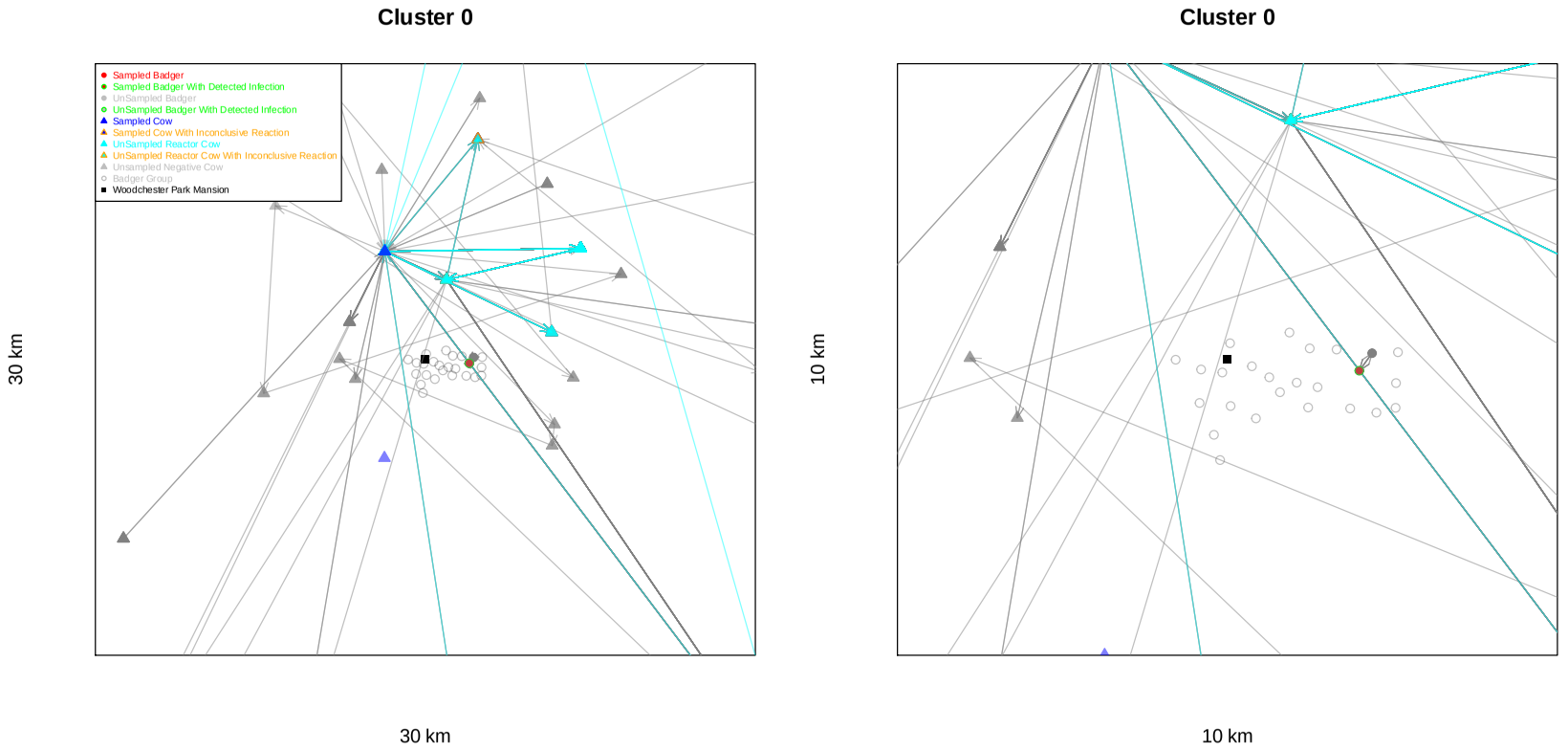

Figure 4.8: The recorded movements (arrows) of the sampled cattle and badgers from cluster 0 and of the associated unsampled in-contact animals, occurring on herds within 15km (LEFT) and 5km (RIGHT) of Woodchester Park. Cattle are represented by triangles coloured grey (in-contact cattle that never reacted to the SICCT), cyan (in-contact cattle that reacted positively to the SICCT during their observed lifespan), or blue (sampled cattle). A cow's triangle is surrounded by an orange line if it reacted inconclusively to an SICCT in its current herd. Badgers are depicted using circles coloured red (sampled badgers) or grey (unsampled in-contact badgers). Badger's circles are surrounded by a green line once they reacted positively to a bTB test. The locations of badger social groups are shown as empty grey circles.

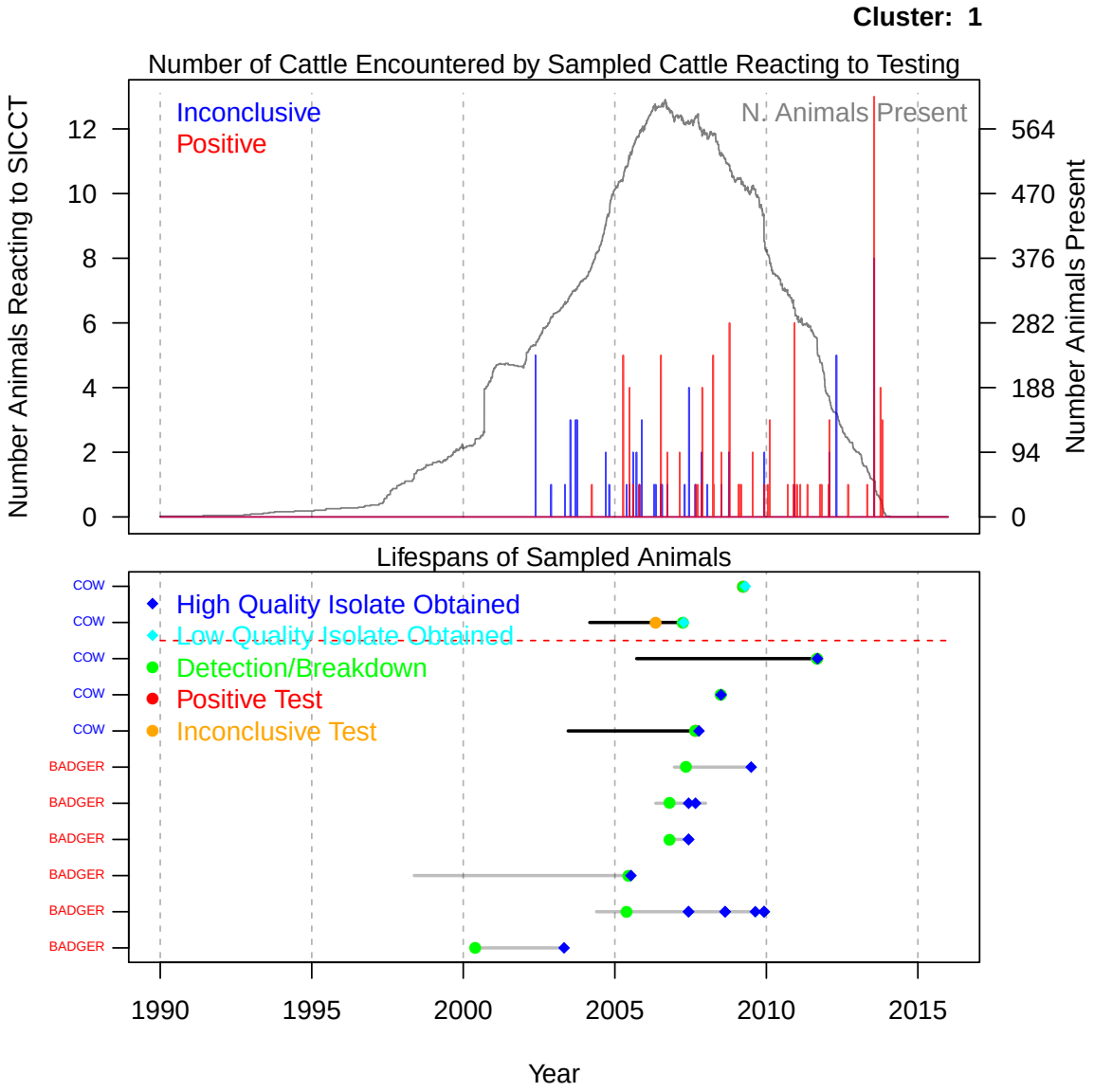

Figure 4.9: The observed lifespans and testing history of the sampled cattle and badgers from cluster 1 and of the associated unsampled in-contact animals. BOTTOM: The observed periods available for each sampled animal and the dates of their bTB testing and sampling events. The horizontal red dashed line separates the isolates with high genome coverage (>=90%) from those with poor coverage. MIDDLE: A summary of the unsampled in-contact cattle. The grey line represents the number of unsampled in-contact cattle alive on each date between 1990 and 2015. The red and blue lines show the numbers of test reactors (inconclusive in blue and reactors in red) on each date. TOP: A summary of the unsampled in-contact badgers. The total number of unsampled in-contact badgers alive between 1990 and 2015 is represented by a grey line, the number of those in-contact badgers that tested positive is shown by the red line.

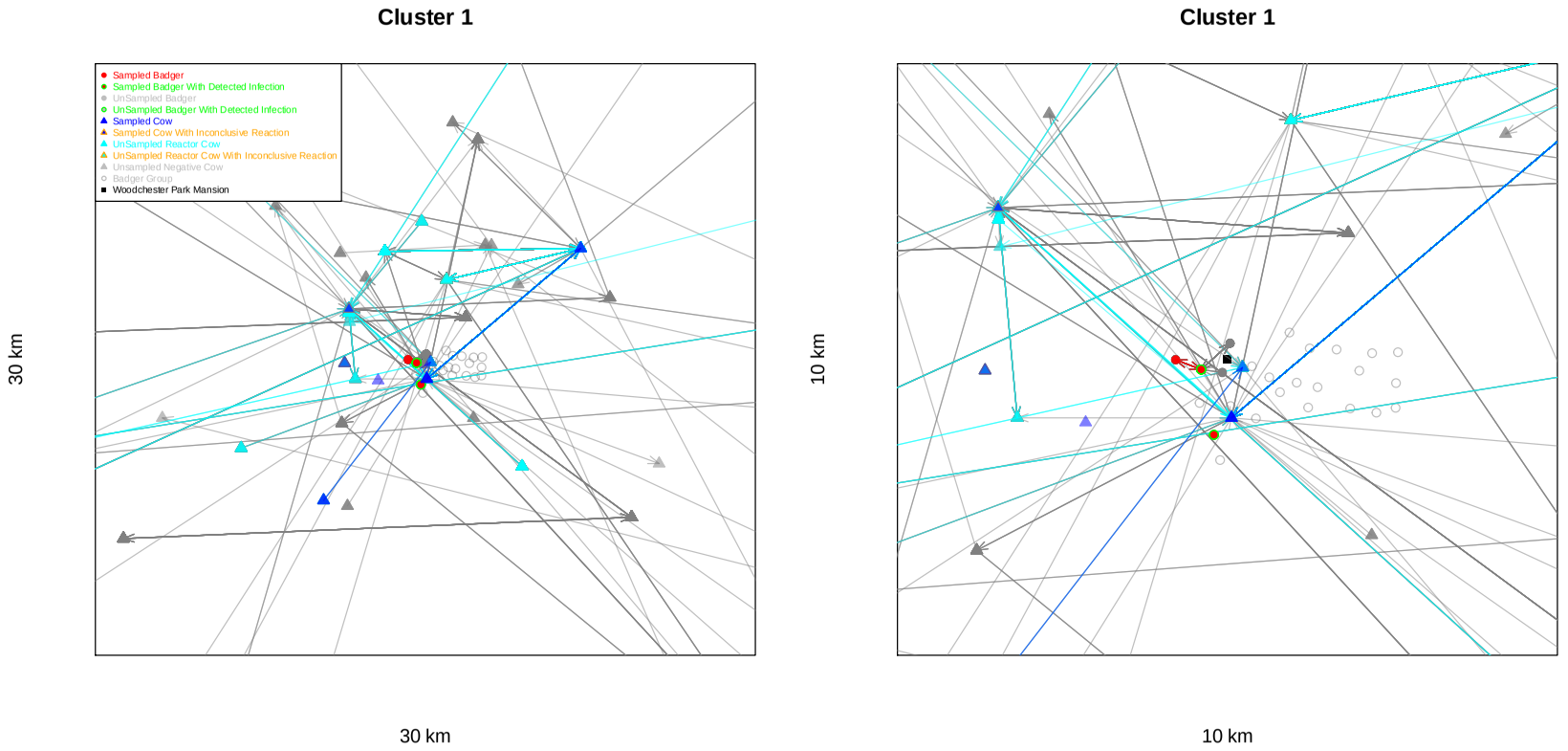

Figure 4.10: The recorded movements (arrows) of the sampled cattle and badgers from cluster 1 and of the associated unsampled in-contact animals, occurring on herds within 15km (LEFT) and 5km (RIGHT) of Woodchester Park. Cattle are represented by triangles coloured grey (in-contact cattle that never reacted to the SICCT), cyan (in-contact cattle that reacted positively to the SICCT during their observed lifespan), or blue (sampled cattle). A cow's triangle is surrounded by an orange line if it reacted inconclusively to an SICCT in its current herd. Badgers are depicted using circles coloured red (sampled badgers) or grey (unsampled in-contact badgers). Badger's circles are surrounded by a green line once they reacted positively to a bTB test. The locations of badger social groups are shown as empty grey circles.

badgers lived in 3 different social groups, two of which were directly connected through the recorded movements of sampled animals.

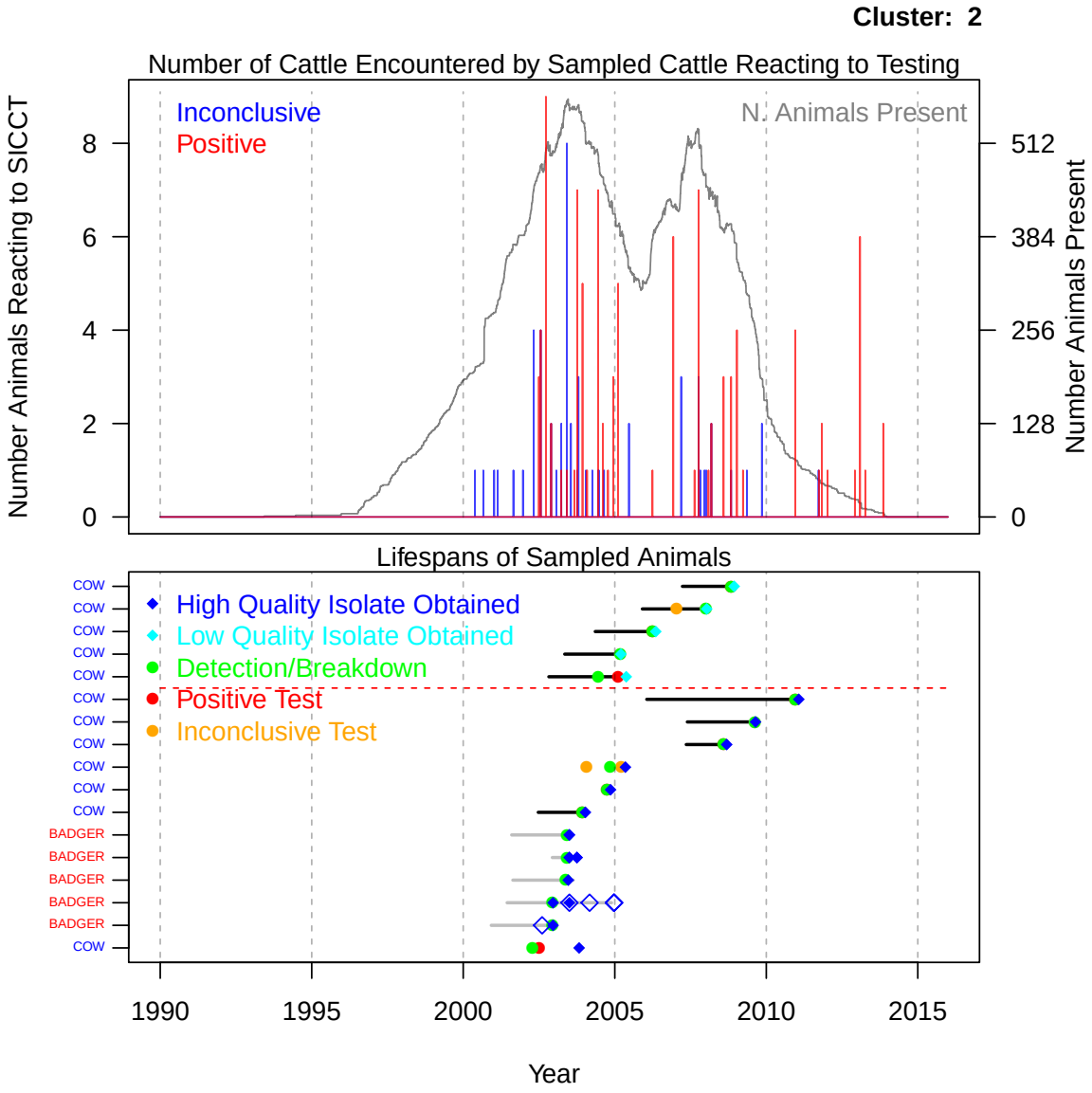

Figure 4.11: The observed lifespans and testing history of the sampled cattle and badgers from cluster 2 and of the associated unsampled in-contact animals. BOTTOM: The observed periods available for each sampled animal and the dates of their bTB testing and sampling events. The horizontal red dashed line separates the isolates with high genome coverage (>=90%) from those with poor coverage. An empty blue diamond is used to show when a sequence associated with a different cluster was sampled. MIDDLE: A summary of the unsampled in-contact cattle. The grey line represents the number of unsampled in-contact cattle alive on each date between 1990 and 2015. The red and blue lines show the numbers of test reactors (inconclusive in blue and reactors in red) on each date. TOP: A summary of the unsampled in-contact badgers. The total number of unsampled in-contact badgers alive between 1990 and 2015 is represented by a grey line, the number of those in-contact badgers that tested positive is shown by the red line.

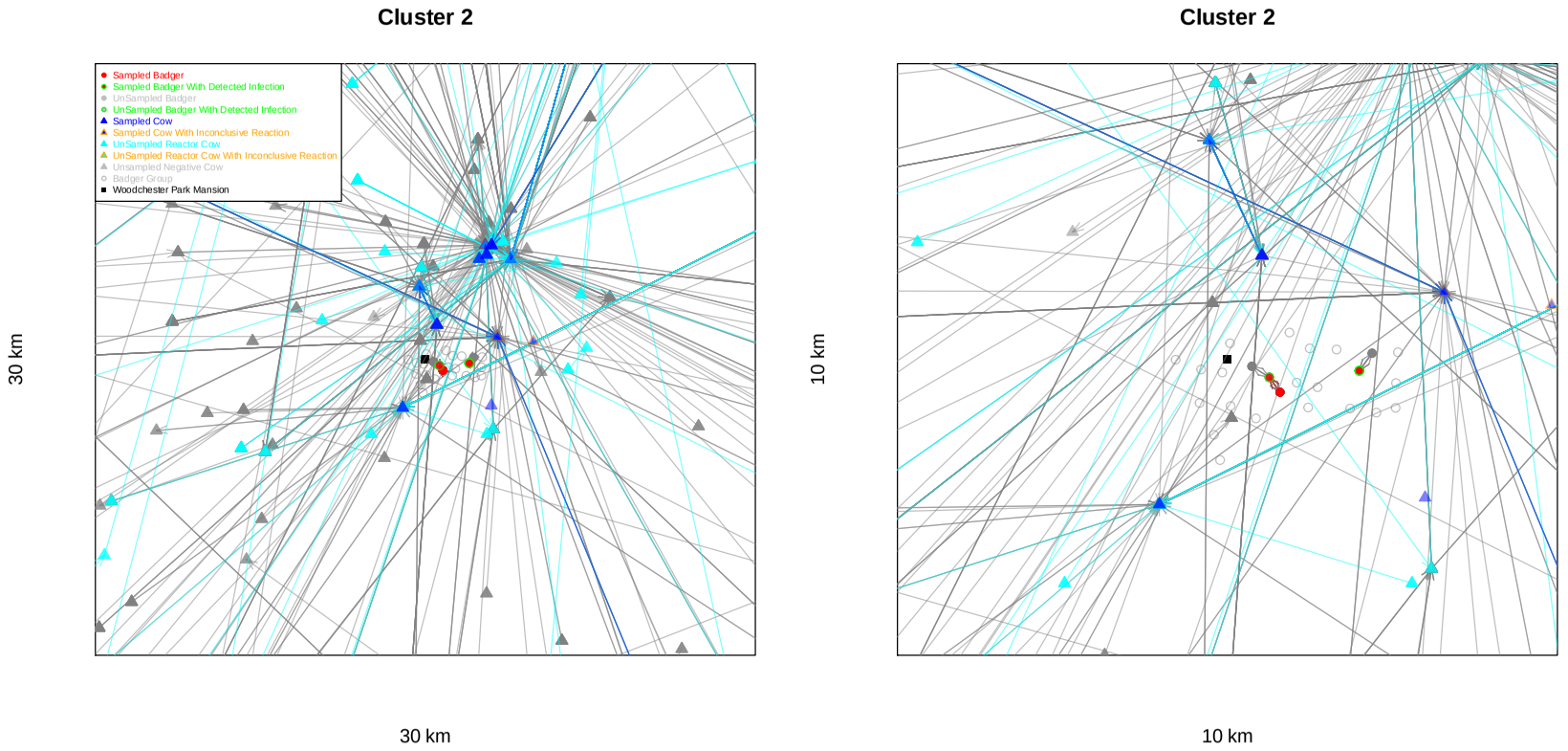

Figure 4.12: The recorded movements (arrows) of the sampled cattle and badgers from cluster 2 and of the associated unsampled in-contact animals, occurring on herds within 15km (LEFT) and 5km (RIGHT) of Woodchester Park. Cattle are represented by triangles coloured grey (in-contact cattle that never reacted to the SICCT), cyan (in-contact cattle that reacted positively to the SICCT during their observed lifespan), or blue (sampled cattle). A cow's triangle is surrounded by an orange line if it reacted inconclusively to an SICCT in its current herd. Badgers are depicted using circles coloured red (sampled badgers) or grey (unsampled in-contact badgers). Badger's circles are surrounded by a green line once they reacted positively to a bTB test. The locations of badger social groups are shown as empty grey circles.

Sequences in cluster 3 originated from two cattle and 13 badgers (Figure 4.13). Infection was detected in two of the sampled badgers before it was detected in the sampled cattle. The lifespans of the sampled animals spanned 15 years from 1995 to 2010. The sampled badgers encountered 153 in-contact badgers and 75 of these in-contact animals tested positive for *M. bovis* infection (Table 4.2). By 1996 four in-contact badgers infected with *M. bovis* were alive. Movement data were only available for one of the sampled cattle, this cow encountered 44 in-contact cattle. All of the 11 in-contact cattle that reacted positively to the SICCT did so after 2005. Two herds, one within Woodchester Park's grounds, were associated with the sampled cattle from cluster 3 (Figure 4.14). The 13 sampled badgers lived in eight different social groups that were all directly connected via the recorded movements of sampled and in-contact badgers.

Cluster 4 contained sequences from four cattle and 36 badgers (Figure 4.15). Infection was detected in the cattle five years after it was detected in the sampled badgers (Table 4.2). The observed lifespans of the sampled animals began in 1993 and ended in 2010. All 20 of the 331 in-contact cattle that reacted positively to the SICCT did so after or during 2002. 162 of the 307 in-contact badgers tested positive for *M. bovis* infection and 72 of these badgers had tested positive before infection was detected in any sampled cattle (March 2002). There were four different sampled herds that were within 3km of Woodchester Park (Figure 4.16), three were within the grounds. These herds weren't connected by any recorded movements of sampled or in-contact cattle. 13 of the 17 badger social groups that the sampled badgers lived on, were directly connected via the recorded movements of sampled or in-contact badgers.

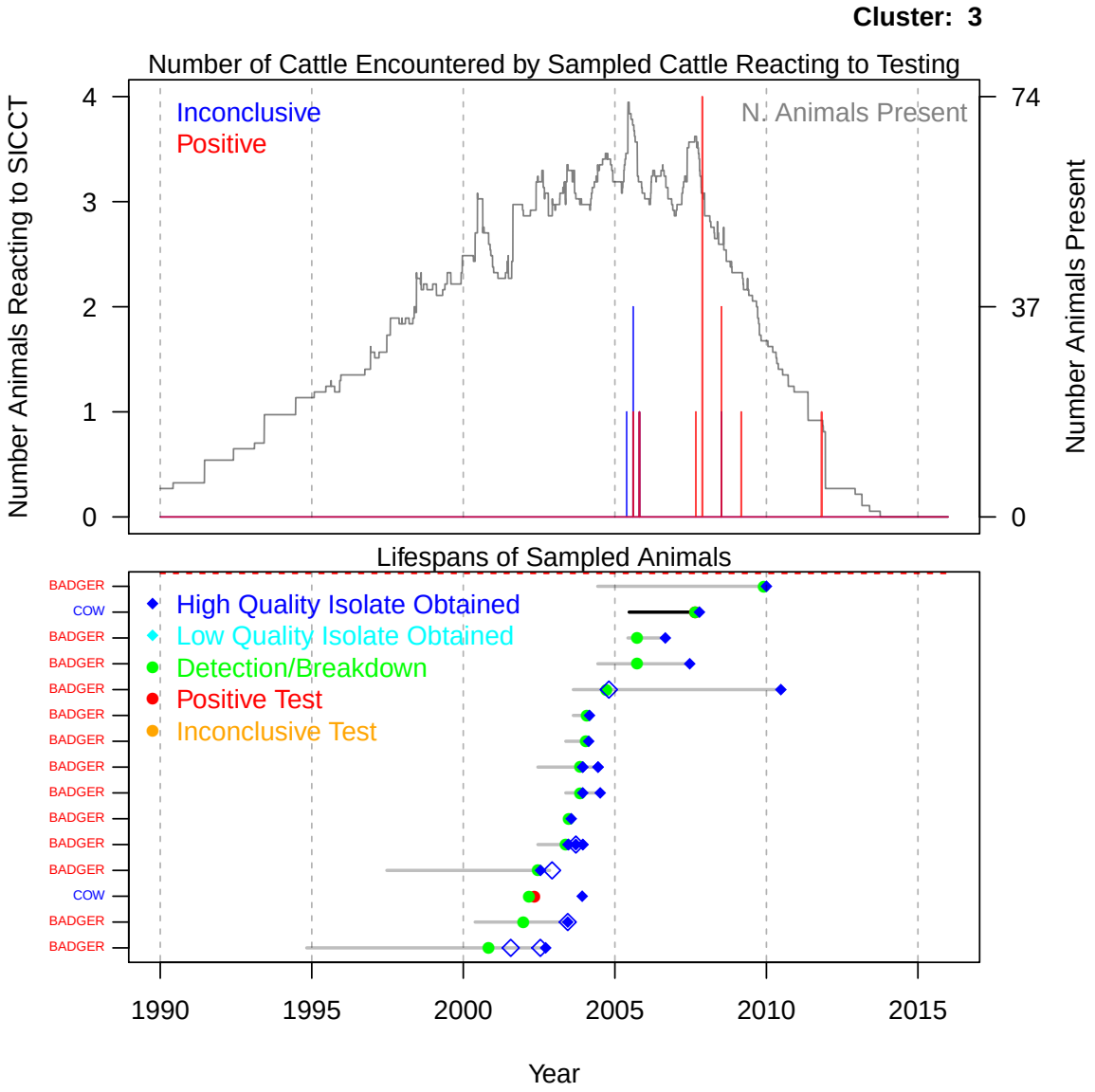

Figure 4.13: The observed lifespans and testing history of the sampled cattle and badgers from cluster 3 and of the associated unsampled in-contact animals. BOTTOM: The observed periods available for each sampled animal and the dates of their bTB testing and sampling events. An empty blue diamond is used to show when a sequence associated with a different cluster was sampled. MIDDLE: A summary of the unsampled in-contact cattle. The grey line represents the number of unsampled in-contact cattle alive on each date between 1990 and 2015. The red and blue lines show the numbers of test reactors (inconclusive in blue and reactors in red) on each date. TOP: A summary of the unsampled in-contact badgers. The total number of unsampled in-contact badgers alive between 1990 and 2015 is represented by a grey line, the number of those in-contact badgers that tested positive is shown by the red line.

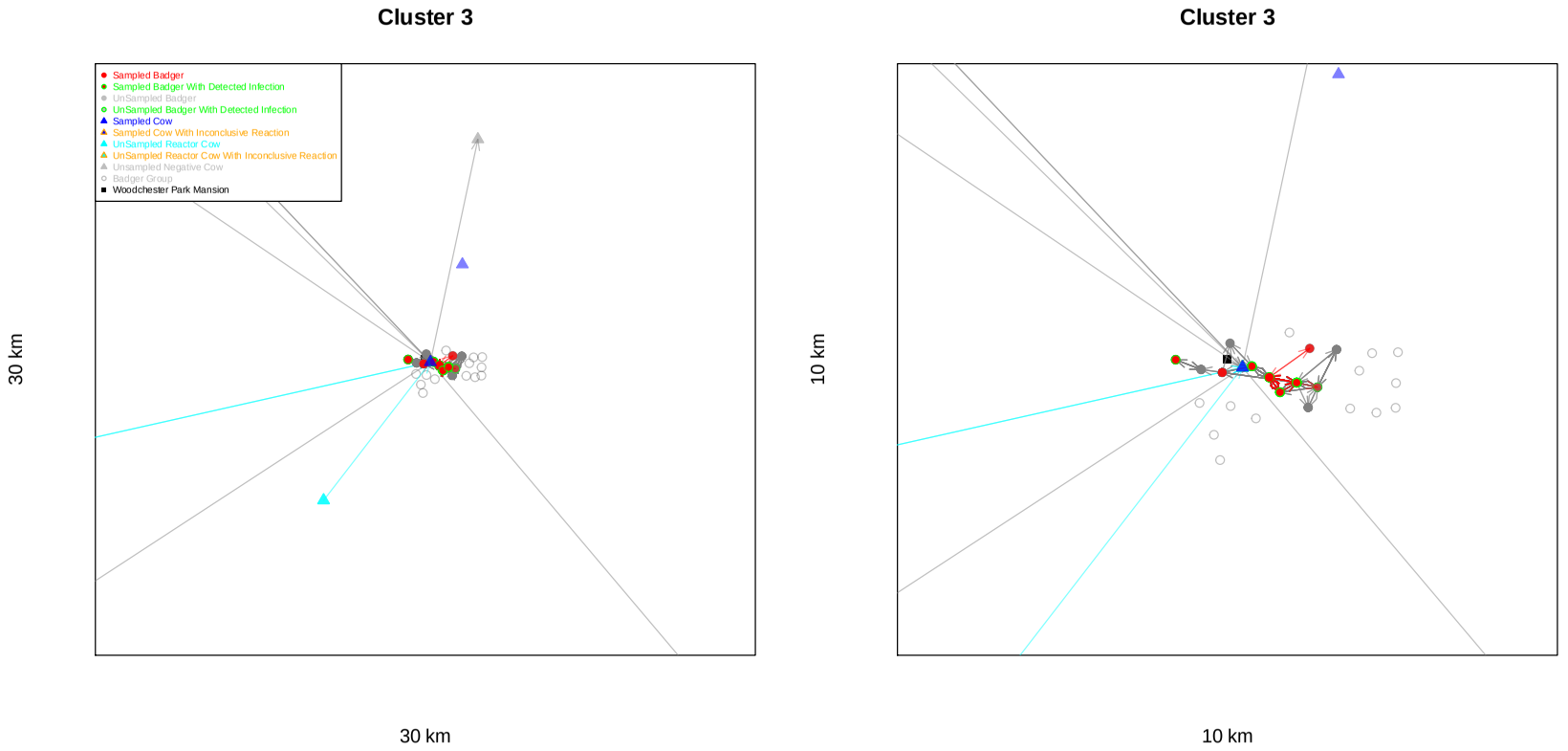

Figure 4.14: The recorded movements (arrows) of the sampled cattle and badgers from cluster 3 and of the associated unsampled in-contact animals, occurring on herds within 15km (LEFT) and 5km (RIGHT) of Woodchester Park. Cattle are represented by triangles coloured grey (in-contact cattle that never reacted to the SICCT), cyan (in-contact cattle that reacted positively to the SICCT during their observed lifespan), or blue (sampled cattle). A cow's triangle is surrounded by an orange line if it reacted inconclusively to an SICCT in its current herd. Badgers are depicted using circles coloured red (sampled badgers) or grey (unsampled in-contact badgers). Badger's circles are surrounded by a green line once they reacted positively to a bTB test. The locations of badger social groups are shown as empty grey circles.

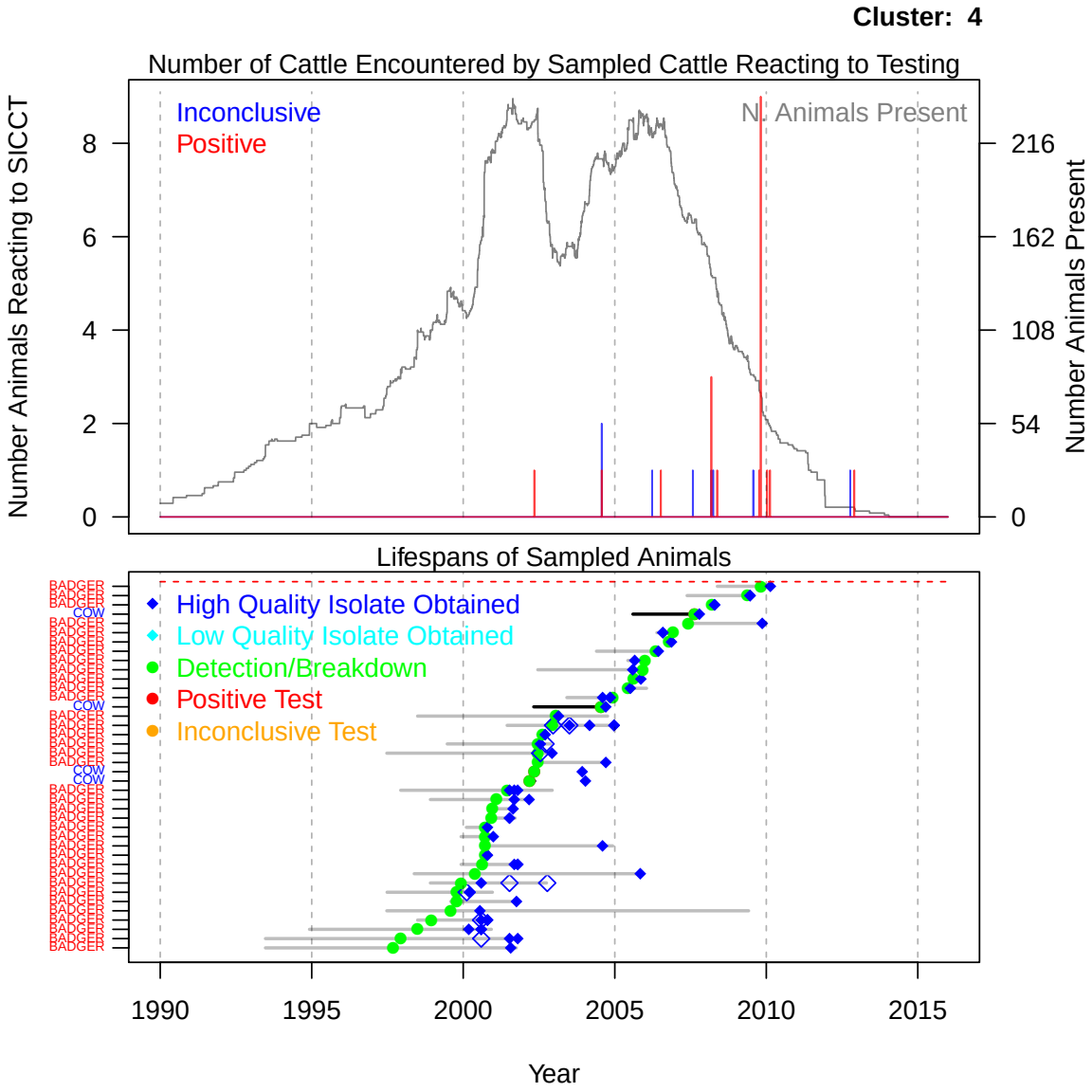

Figure 4.15: The observed lifespans and testing history of the sampled cattle and badgers from cluster 4 and of the associated unsampled in-contact animals. BOTTOM: The observed periods available for each sampled animal and the dates of their bTB testing and sampling events. An empty blue diamond is used to show when a sequence associated with a different cluster was sampled. MIDDLE: A summary of the unsampled in-contact cattle. The grey line represents the number of unsampled in-contact cattle alive on each date between 1990 and 2015. The red and blue lines show the numbers of test reactors (inconclusive in blue and reactors in red) on each date. TOP: A summary of the unsampled in-contact badgers. The total number of unsampled in-contact badgers alive between 1990 and 2015 is represented by a grey line, the number of those in-contact badgers that tested positive is shown by the red line.

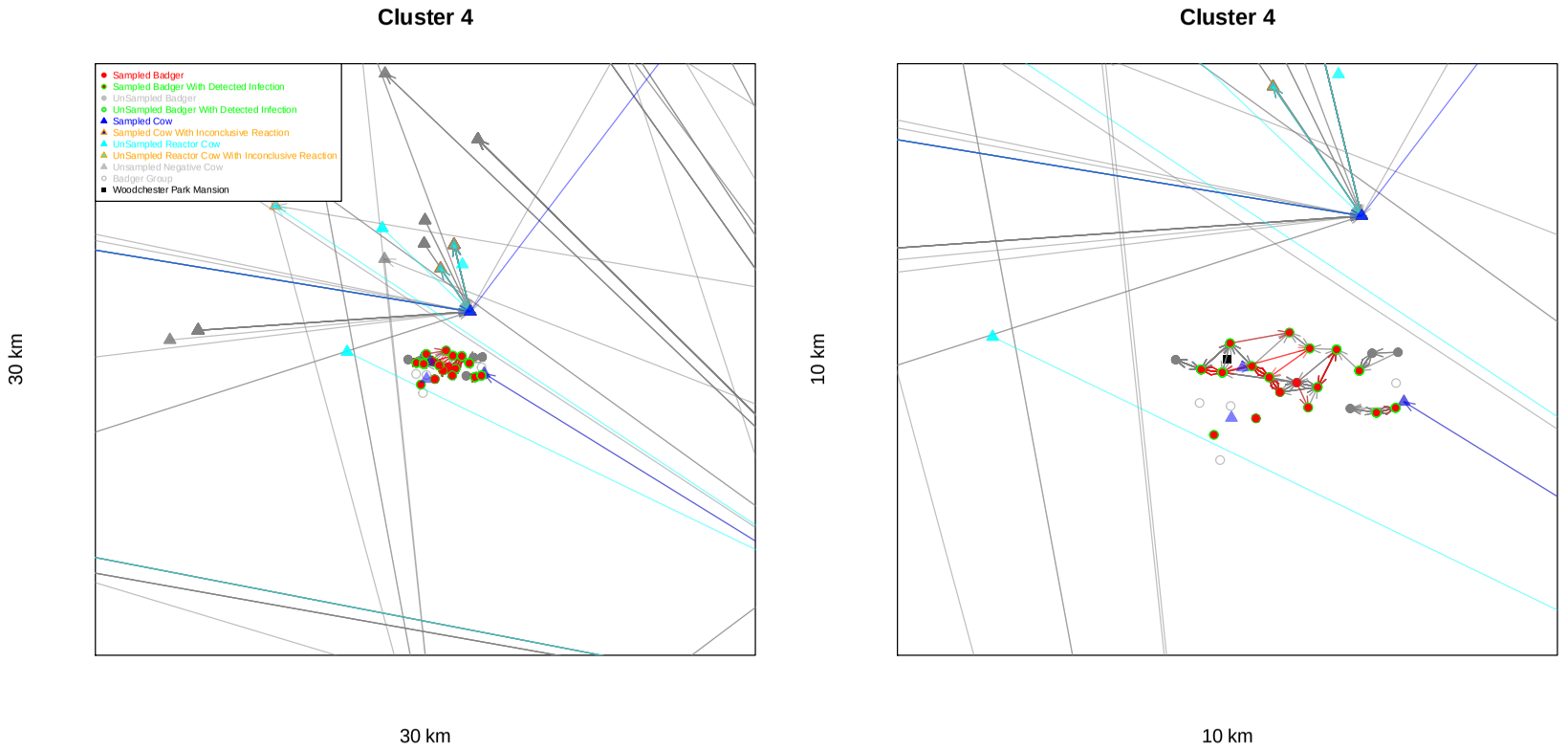

Figure 4.16: The recorded movements (arrows) of the sampled cattle and badgers from cluster 4 and of the associated unsampled in-contact animals, occurring on herds within 15km (LEFT) and 5km (RIGHT) of Woodchester Park. Cattle are represented by triangles coloured grey (in-contact cattle that never reacted to the SICCT), cyan (in-contact cattle that reacted positively to the SICCT during their observed lifespan), or blue (sampled cattle). A cow's triangle is surrounded by an orange line if it reacted inconclusively to an SICCT in its current herd. Badgers are depicted using circles coloured red (sampled badgers) or grey (unsampled in-contact badgers). Badger's circles are surrounded by a green line once they reacted positively to a bTB test. The locations of badger social groups are shown as empty grey circles.

4.4 Discussion

Woodchester Park, in the southwest of England, is home to a naturally infected high density badger population and is surrounded by cattle herds. The prevalence and extent of *M. bovis* infection in the southwest of England has been increasing, and badger populations are heavily implicated [57, 151]. Woodchester Park represents an ideal location to investigate the transmission of bTB within and between badger and cattle populations. The current research provides evidence that a sampled badger population was capable of maintaining *M. bovis* infection independently, and that multiple transmission events occurred between this population and the sampled cattle, in both directions.

A high proportion (approximately 70%) of the variation in the genetic distances between the badger isolates can be explained by a statistical model incorporating spatial, temporal and network-based metrics. The informativeness of the temporal, spatial and network based epidemiological metrics suggests that *M. bovis* infection is being transmitted via the movements and contact patterns of the badgers within the sampled population. The dominant red clade in Figure 4.4 contained 139 badger isolates and eight cattle isolates, all of which were less than 10 SNPs apart. There were infected badgers involved in clusters 3 & 4 from Figure 4.5 between 1997 and 2010 (Figure 4.13 & Figure 4.15). Since these clusters were from within the red clade, defined in Figure 4.4, this suggests that the dominant strain at Woodchester Park has been circulating for over a decade. A population can act as a maintenance reservoir when infection is able to persist within it through time, without continued introductions from outside the population. Given the agreement between the epidemiological and genetic data, the high genetic similarity in the badger isolates, and the evidence of prolonged outbreaks in the sampled badger population, these data suggest that the badger population within and surrounding Woodchester Park is capable of acting as a maintenance reservoir for bTB.

For badger populations to be important in the spread and persistence of bTB in UK cattle populations, badger to cattle transmission must occur. The observed lifespans and bTB testing histories of the badgers and cattle associated with clusters 3 & 4 suggest that the represented strains of *M. bovis* originated in the sampled badger population. In cluster 4, infection was detected in half of the sampled badgers before any of the sampled cattle (Figure 4.15), was circulating within 13 (of 17) connected badger social groups (Figure 4.16), was found on cattle farms that weren't connected but were within 3km of Woodchester Park, and the 20 infected in-contact cattle involved, were infected late in the outbreak (Table 4.2). In cluster 3, the evidence was less substantial but the sampled badgers were circulating in eight connected social groups, infection was first detected in the sampled badgers, and only two sampled cattle were involved, one of them lived on a farm that was within Woodchester Park's grounds. No recorded cattle movements linked the two farms that the sampled cattle lived on. However, it wasn't possible to link one of the sampled cows to any movement data.

Transmission from cattle to badgers presents a mechanism by which the maintenance potential of local badger populations could be linked to the long distance mobility of cattle. There was evidence that the strains represented by cluster 2 originally circulated in the sampled cattle population, and spread into the sampled badgers. In cluster 2, infection was first detected in the sampled cattle. These sampled cattle encountered 96 in-contact cattle that tested positive for bTB - 18 of which were infected before any sampled badgers. In addition, the sampled cattle lived on highly connected herds that were an average of 5km from Woodchester Park, and were closest to the ancestor of the cluster. The five sampled badgers involved, lived in groups that weren't connected via the recorded movements of sampled or in-contact badgers. In addition, Figure 4.5 shows that clusters 0, 1, and 2 are distinct (approximately 20 SNPs) from the main strain circulating in Woodchester Park. These clusters may represent strains that were introduced and are becoming established. In the case of cluster 2, the WGS and epidemiological data suggest that its ancestor originally resided in cattle.

The WGS data available for the sampled cattle and badger population surrounding and within Woodchester Park provides evidence, both of the maintenance ability of a high density badger population, and of inter-species transmission of bTB between cattle and badgers. However, there are problems with the data that was analysed. The badger population, although densely sampled, was only sampled within a limited spatial range. The 100 cattle isolates represent a large under-sampling for the cattle population (Figure 4.2). Figure 4.2 demonstrates that farms around Woodchester Park were more densely sampled, and that these included slightly more beef than dairy farms. In addition, the potential to detect *M. bovis* infection in both cattle and badgers is limited, as a result of the low sensitivities of the diagnostic tests available, which means that there may be a large number of undetected infected animals whose impact is difficult to quantify. The evidence for badger to cattle transmission, and the maintenance role of the sampled badger population, should be unaffected by these data issues, since the badger population was densely sampled and cattle were found carrying the dominant strain present in the badger population. The evidence for cattle to badger transmission may be affected if badgers outside of Woodchester Park are found to carry similar strains to the cattle.

Despite the limitations of the current dataset the results and observations have important implications. The sampled high density badger population appears to be acting as a maintenance reservoir providing a mechanism for local persistence of bTB within Woodchester Park. A wealth of literature demonstrates that badgers and cattle interact [148, 147, 149, 151, 228] and the WGS data available provides strong evidence of badger to cattle transmission. Recurrence of *M. bovis* infection on cattle farms is a problem across the UK [81]. In Ireland, badger populations are recognised as a means of local persistence and a driver of breakdown recurrence [84, 82]. Surveillance and control of infection in a wildlife maintenance reservoir is necessary to limit its spill-over potential. Despite its success in Ireland, badger culling has been shown to be ineffective in the south of England [224, 229]. Vaccination trials in badger

populations in England are on-going [170], but the current vaccine does has limited efficacy [230, 169]. A Test, Vaccinatate, or Remove (TVR) trial, which combines animal-side bTB testing, vaccination, and culling, is on-going in Northern Ireland and may provide a publicly acceptable and efficient means of controlling bTB in badger populations [172].

The current research presents evidence of inter-species transmission from cattle to badgers. As Figure 4.2 demonstrates, there has been a large number of breakdowns in the herds surrounding Woodchester Park. *M. bovis* infection is present in cattle populations across England, although bTB is most prevalent in the southwest [57]. The high genetic diversity of the cattle isolates may reflect links to the broader England-wide *M. bovis* population (Figure 4.5). Whilst the dispersal of badgers is generally under 10km [231], cattle movements regularly span hundreds of kilometres [226]. These long distance cattle movements are recognised as an important driver of the spread and persistence of bTB in the UK cattle population [73]. Importantly, if cattle to badger transmission of bTB does occur, as the current research suggests, it could be a mechanism for creating and linking local reservoirs.

Woodchester Park provided an opportunity to investigate the transmission of bTB between cattle and wildlife. WGS provides the highest possible genetic resolution to inform this investigation. Cattle and badgers from Woodchester Park were sampled and the isolated *M. bovis* was sequenced. The genetic relationships between the isolates suggested that the badgers were maintaining *M. bovis* infection and that inter-species transmission was occurring in both directions, although a single host species dominated in each observed cluster. The detailed epidemiological data available for the sampled populations was interrogated and corroborated the patterns in the genetic data. These findings have important implications for the control of bTB in the UK. Surveillance operations of local badger populations may be necessary to identify and limit the spill-over risk of local maintenance reservoirs. In addition, reducing the amount of cattle moving out of high prevalence areas will be necessary to limit the long-distance spread of bTB.

Chapter 5

General Discussion

5.1 Summary of Findings

Mycobacterium bovis is the bacterium responsible for bovine tuberculosis (bTB). It is well adapted to surviving in multiple hosts and exploits this ability in order to persist in situations where a single-host pathogen could be eradicated. Control of bTB in livestock populations, deemed necessary due to the health and economic burdens of the disease, is often hindered when wildlife reservoirs are present. In the United Kingdom (UK), the European badger (*Meles meles*) is thought to act as a reservoir for infection [18], whereas the brushtail possum (*Trichosurus vulpecula*) plays this role in New Zealand [16]. The presence of these wildlife hosts greatly complicates bTB control in these countries. The overall aim of this thesis was to describe an investigation into the transmission of bTB within and between wildlife and livestock populations by exploiting Whole Genome Sequenced (WGS) *M. bovis* isolates taken from bTB systems in New Zealand and the UK.

The continuing advances of available sequencing technologies, in speed, accuracy and price, now means that WGS could be incorporated into routine surveillance as a molecular typing tool, replacing methods such as Variable Number Tandem Repeat (VNTR) typing and spoligotyping. This transition will require the tools necessary to handle and process the vast quantities of data that WGS produces. Chapter 2 of this thesis describes an automated procedure to inform the selection of quality thresholds used in the processing of WGS data. The selection of quality thresholds is usually informed by expert knowledge and published literature. This chapter examined WGS data taken from different systems in the UK and New Zealand in relation to spatial, temporal, and network data and demonstrated that the epidemiological data available for the sampled bTB systems could additionally be used to inform this selection process.

If WGS data is to be incorporated into the routine surveillance of bTB in livestock it should, ideally, have good agreement to other molecular typing methods that are in use.

A good agreement will ease the transition from previous methods to WGS. In Chapter 3, there was a high level of agreement between the Restriction Endonuclease Analysis (REA) types and the phylogenetic relationships between the WGS *M. bovis* isolates (Figure 3.3). Within each REA type, isolates could be further distinguished using variation present in their genomes. One of the most exciting parts of the current surge in the use of WGS methods, is learning how these high resolution genetic data can be used. The aim of the research described in Chapter 3 was to use WGS to evaluate the role of the wildlife reservoir species found in New Zealand - the brushtail possum (*Trichosurus vulpecula*). Using the data available it was only possible to show that *M. bovis* infection was passing between cattle and wildlife populations, but not to determine the direction of this transmission. Although the analyses of the current data available weren't able to quantify inter-species transmission rates, for reasons discussed below, given the current situation in New Zealand with less than 40 herds across the whole country infected with *M. bovis* [60], the current predominant direction of transmission is likely to be from possums to cattle.

The inability to provide directional estimates of inter-species transmission between wildlife and cattle in New Zealand, based on the WGS dataset available for this system, was partly due to sampling biases. The extent of the effect of these sampling biases was examined using a simple simulation model, which is described in Appendix A. Biased sampling of the pathogen across time and multiple host populations produced potentially misleading estimates of inter-population transmission rates. Therefore, interpretation of the inter-species rate estimates when such biases exist (such as is the case with the New Zealand dataset) must be bolstered by a thorough examination of these biases as was done here.

The WGS data available from New Zealand provided the opportunity to estimate the substitution rate for *M. bovis* in New Zealand. This rate can be used as a prior for future investigations into the separation times of distinct epidemiological populations. An additional simulation study demonstrated that the broad credible intervals surrounding this estimate may have resulted from the narrow sampling window, relative to the potential number of *M. bovis* generations involved for the sampled population (Appendix A). Importantly, Chapter 3 demonstrates that WGS is most powerful when the sampling is conducted appropriately. Sampling of a pathogen in multiple populations should be conducted relative to the incidence of the pathogen in each population. For the particular pathogen it is important to consider how quickly genetic variation is generated. The temporal and spatial sampling ranges should be broad enough to observe genetic variation at the same scale as that of the epidemiological questions being asked. For example, if you were interested in the within-herd dynamics of bTB, densely sampling a herd across a single year would result in a large number of almost identical *M. bovis* strains that aren't very informative. Whereas, sampling the same herd over a number of years would generate considerably more genetic variation and potentially provide insights into the within-herd dynamics. In this case, since the aim of Chapter 3 was to examine inter-species transmission of bTB it may have been more appropriate to retain a

broad sampling timeframe but focus the sampling efforts on a smaller spatial scale.

In the course of the analyses of the WGS data from the New Zealand isolates, evidence of mislabelling was uncovered. Isolates of the same REA type, which were expected to be genetically similar, were highly genetically distinct. Additional VNTR assays were conducted (by Marian Price-Carter, a collaborator in New Zealand) and the mislabelled isolates identified were removed from any further analyses. Using spoligotyping data, a similar pattern revealed mislabelling in the WGS dataset from Woodchester Park (Appendix B). The spoligotyping information revealed mislabelled isolates. The detailed epidemiological data available for the sampled badger population provided an additional means of investigating the extent of the mislabelling. For any mislabelled isolates, it would be expected that the relationship between the proposed epidemiological metrics and the observed genetic distances should be damaged, so long as the set of metrics include the most important factors describing the epidemiological distances between the isolates. Therefore, isolates whose pairwise genetic distances were poorly described by the epidemiological metrics were removed as these were most likely to have been mislabelled. Unfortunately this means that correctly labelled isolates that don't relate well to the available epidemiological data, as a result of interesting and unanticipated interactions, may be removed. The high level of agreement between the genetic and epidemiological data, overall, suggested the extent of the mislabelling was limited and unlikely to impact any conclusions drawn. Mislabelling is an unfortunately common occurrence but the current research shows that if WGS and epidemiological data, which here included other molecular typing data, are interpreted in combination, inconsistencies can be revealed and removed.

The cattle and badger populations surrounding and within Woodchester Park provided an ideal system to investigate inter-species transmission at a local scale (Chapter 4). WGS data from isolates, sampled across 15 years, were accompanied by extremely detailed records of the sampled population dynamics. These data were analysed to determine the direction and extent of inter-species transmission in this bTB system. Unfortunately, sampling biases were present in this dataset with considerably more isolates from badgers sequenced than from cattle. In addition, the badgers were sampled from within a smaller spatial area. Keeping these biases in mind, there was considerable evidence available to suggest that inter-species transmission was occurring between the sampled cattle and badger populations, in both directions. In addition, the detailed epidemiological data available for the sampled badger population was used to demonstrate that genetic signatures in the badger isolates could be tied, using machine learning approaches, to the recorded spatial, temporal, and network dynamics of the sampled population (Figure 4.3). These strong genetic signatures of the badger population dynamics in Woodchester Park, alongside the evidence of outbreaks persisting for over a decade (Figure 4.13 & Figure 4.15) suggested that the sampled high density badger population was acting as a maintenance reservoir. If, as the current data suggests, inter-species transmission does occur in both directions and badger populations can maintain infection in-

dependently, the inter-species transmission provides a mechanism by which the long distance mobility of UK cattle populations can be linked to the maintenance role of local badger populations.

To understand the role of a wildlife reservoir in the spread and persistence of bTB in cattle populations, quantifying the extent of transmission between livestock and wildlife is essential. Chapter 3 & Chapter 4 provide a detailed description of some of the methods used to investigate inter-species transmission between livestock and wildlife in the United Kingdom and New Zealand. These analyses demonstrate the relevance of wildlife reservoirs in the control of bTB in cattle populations, but also highlight how sampling biases can influence results.

5.2 Limitations, Opportunities and Future Directions

This thesis describes a series of methods and analyses conducted to investigate inter-species transmission of bTB using WGS. It is important to understand the results of these chapters in the context of the limitations of the data used and analyses conducted. The future progression of the research described in this thesis will involve addressing and accounting for these limitations, as well as incorporating new data and analyses.

Chapter 2 details an automated procedure to inform the selection of quality criteria to be used in a processing pipeline for WGS data. This procedure relies upon available epidemiological data, and therefore the degree to which it can be informative is limited by the quality of the epidemiological data available. A further limitation of these analyses was effectively avoiding over-fitting. In some cases if too strict quality filtering was implemented, the variation explained by the fitted Random Forest model improved. Simply selecting the quality thresholds that resulted in a fitted Random Forest model explaining the highest amount of variation in the inter-isolate genetic distances would have meant using very strict filter thresholds, keeping very high quality, and likely true, genetic data but discarding a lot of genetic data that may have proved informative. In Chapter 2 this was addressed by using the number of sites removed by the filtering as an indication of when the filtering was too harsh. An improvement to the analyses could be the development of a method that penalises quality filtering by automatically accounting for the number of sites that the filtering removes.

The analyses in Chapter 2 were limited to only six quality filters. Additional quality filters could be incorporated. For example, the general quality score - QUAL - that has been used previously on WGS *M. bovis* data (Table 2.1) could be examined. In addition, the threshold ranges for the quality filters investigated could be broadened to allow for the effects of stricter thresholds to be examined.

The filter sensitivity analysis could be further improved by examining convergences of SNPs. Convergent SNPs, observed between isolates without a recent shared ancestry, are more likely to be errors, since the low substitution rate of *M. bovis* (Table 3.3) means that there is a low probability that a mutation will occur at the same position on a genome more than once. If many of these convergent events were observed, this could be an indication that the quality processing conducted was ineffective. Convergent SNPs have been considered for WGS *M. bovis* [25] and *M. tuberculosis* [117, 232] before. Therefore, the filter sensitivity analysis could be improved by incorporating a metric describing the number of convergent SNPs, which were present in a WGS *M. bovis* dataset, following quality filtering.

An important step in the incorporation of WGS into routine bTB surveillance will be the development of a standard bioinformatics pipeline to process and handle the vast quantities of data that such surveillance would produce. Chapter 2 represents a stepping stone in that

direction. Such analyses could be used to inform the quality filtering undertaken in the automated pipeline. A fully automated pipeline would have to be generalisable, able to handle data from a variety of sequencing platforms producing data with varying degrees of quality, in addition to being at least as accurate as the semi-automated approaches currently used. An important consideration, with the generation of WGS *M. bovis* data from across the world, will be how these data can be compared whilst accounting for the large amount of genetic and epidemiological history that separates the datasets. This problem was alluded to in Chapter 3, where, although over 300 WGS *M. bovis* isolates across New Zealand were available, only a small subset of these could be used, since the large between-clade genetic distances relate to epidemiology that wasn't on the same scale as the epidemiological questions of interest.

The selection of a small subset of the available isolates for Chapter 3 demonstrates an important point. If WGS is to be used effectively, it is important to recognise its limitations, and how sampling is conducted is a big part of that. A continuation of the research in Chapter 3, using the same data, could be done to address a different question. For example, were movement data regarding the sampled cattle populations available, it might be possible to use the WGS data to recognise patterns of movements that were associated with bTB spread. In the future when more WGS data for *M.bovis* isolates are available, the inter-species transmission question could be readdressed. To quantify the role of wildlife, cattle and wildlife would ideally be sampled densely within a small spatial scale (10-20km) across as broad a time scale (at least 5-10 years) as possible. Importantly, any sampling conducted would have to be done in such a way that the dataset was representative of the underlying population densities through time.

The simulation model described in Appendix A was designed to investigate the influence of the types of biases observed in the New Zealand dataset. An analysis of the simulated genetic data using BEAST, a phylogenetic analysis platform, provided an estimated substitution rate that was consistently higher than the substitution rate specified in the model parameters. This difference may have been the result of estimating the rate on a tree based structure. If this can be shown to be true it would have important implications. Substitution rates based on a phylogenetic tree could be overestimations, in situations where the true rate is low, relative to the transmission rate. Once the difference between the specified and estimated substitution rate is resolved, the analysis could be extended by increasing the complexity of the simulation model; heterogeneity in the substitution rate could be explored. A simulated pathogen with a highly variable substitution rate might be more representative of *M. bovis*. Factors such as latency have been investigated as drivers of substitution rate variability for TB [120, 114].

The sequencing quality and sampling biases (representativeness) of the WGS data from isolates sourced from infected badgers and cattle within and surrounding Woodchester Park represent the main limitations of the analyses described in Chapter 4. The badger population was sampled more densely and at a smaller spatial scale than the cattle population, which

made it difficult to determine true cattle to badger spread. This bias could be mitigated if there were additional data regarding the badger population surrounding Woodchester Park, such as population densities and disease prevalence statistics. The bias of the sampling towards more badgers was addressed by investigating the test and movement histories of unsampled cattle, however it would be better mitigated if more WGS cattle isolates were available. The poorer sequencing quality of a number of the cattle isolates used in the analyses described in Chapter 4 is also an important consideration. The genetic data available for the poorly sequenced isolates went through the same quality filtering as the other isolates did, therefore their lower sequencing quality will result in less high quality information being available for further analyses. The loss of information may have caused an underestimation of genetic distances involving these isolates. This underestimation is unlikely to influence the results discussed above, since the spatial, network and temporal patterns were investigated independently of the genetic data. In addition, where the available genetic data place a poorly sequenced isolate within a defined clade on the phylogeny, this placement is likely to be stable. In this chapter the issue of sequencing quality has been considered, but not dealt with. Given the importance of the results found, it might be worthwhile to have the cattle isolates with poor sequencing quality re-cultured and sequenced again or finding/developing phylogenetic inference methods that could explicitly account for the poorer sequencing quality.

The results presented in Chapter 4 suggest that inter-species transmission has occurred in both directions between the sampled cattle and wildlife populations. The patterns observed in the genetic, temporal, network, and spatial data act to corroborate one another and make this argument more compelling. An improvement to this research would be to examine the genetic, temporal, network, and spatial patterns that result from randomly selecting groups of isolates. The patterns of the random selections would provide a means of comparison, strengthening the conclusions if the patterns were different. In addition, ancestral state reconstruction methods could be used to quantify inter-species transmission rates in the sampled bTB system. However, as demonstrated in Chapter 3, these methods are highly sensitive to sampling biases. De Maio *et al*. [233] recently developed a method that can conduct ancestral state reconstruction methods whilst accounting both for the structured population and sampling biases. A continuation of the work described in Chapter 4 will be to evaluate whether this method can be applied to the Woodchester Park's sampled bTB system. Given that De Maio *et al*.'s [233] approach relies upon phylogenetic inference, and therefore genetic distances, an important part of this future work will be considering the impact of including the poorly sequenced cattle isolates.

Many of the sampled herds and social groups, in the Woodchester Park dataset, were linked directly via the recorded movements of the sampled or in-contact cattle and badgers. This observation suggests that cattle movements and badger dispersal are important mechanisms of bTB transmission within the sampled cattle and badger populations. Cattle movements have previously been shown to be an important means by which bTB is spread within

the UK cattle population [73, 75]. Network measures were incorporated into the investigations of the inter-isolate genetic distances for the sampled badger population. If the sequencing quality of the cattle isolates could be improved, a similar analysis could be conducted to determine whether network characteristics on a local scale relate to bTB transmission. Weber *et al*. [234, 235] recently used proximity loggers on resident badgers from Woodchester Park and found that infected badgers were socially isolated within their own social group but well connected to badgers within other social groups. The use of proximity loggers allowed badger movement to be analysed at a finer scale than was available for the research described in Chapter 4. The movements of sampled and in-contact badgers in Woodchester Park recorded via the trapping data could be further examined to determine whether this relationship between infection and spread could be observed on a larger coarser spatial scale. If infected badgers were consistently involved in more dispersal events between groups than susceptible badgers, this work could be extended to investigate whether being more mobile resulted from or in *M. bovis* infection.

A much higher proportion of the in-contact badgers were found to have tested positive for bTB in their lives than the in-contact cattle examined (Figure 4.7 , Figure 4.9, Figure 4.11, Figure 4.13, & Figure 4.15). However, the proportion of in-contact badgers testing positive for bTB at some point in their lives approximately matched that of the badger population as a whole. Therefore this observation suggests that living in the same social group as a badger who tests positive isn't an important predictor of future infection and highlights the high prevalence of bTB in this high density badger population. Understanding the importance of this finding and the difference in prevalence observed in the in-contact badgers and cattle could be an aim of continued investigations of these data.

In this thesis, the analyses conducted and lessons learnt shouldn't be restricted to the systems examined. There are wildlife reservoirs involved in the spread and persistence of bTB in livestock around the world, wild boar and deer in Spain and France, buffalo in Africa, and deer in North America [200]. If appropriate data becomes available, the methods used here could be utilised elsewhere to investigate the role of these different wildlife reservoirs. Outside of bTB systems, defining the role of species in multi-host communities is a considerable challenge and quantifying the extent and direction of inter-species transmission is an important part of this.

Conclusions

This thesis describes research that aimed to use WGS data of *M. bovis* to investigate the transmission of bTB within and between cattle and wildlife populations. Statistical, mathematical and simulation based approaches were used to address three main themes:

1. The utility of WGS in epidemiological investigations of *M. bovis* transmission.
2. The application of phylogenetic inference methods to examine the dynamics of bovine tuberculosis.
3. The role of wildlife reservoirs in the transmission and persistence of *M. bovis* infection in domestic cattle populations.

The vast quantities of data resulting from Next Generation Sequencing platforms can be handled and processed efficiently. The error rates of these platforms can be mitigated, to some extent, by conducting quality filtering, a process that can be informed using both the published literature and any available epidemiological data. Following the handling and processing of the WGS available it was possible to investigate sampled bTB systems in the UK and New Zealand. The WGS data were demonstrated to be consistent with, and provided considerably more discriminatory power than previous molecular typing methods. For the sampled bTB system in New Zealand it was possible to demonstrate a high degree of exchange between the sampled cattle and wildlife populations. In addition, the substitution rate estimate, although broad, will inform future epidemiological investigations. Using the WGS *M. bovis* isolates from a bTB system in the southwest of England it was possible to reveal potential inter-species transmission events, providing the first evidence for transmission occurring in both directions. Importantly, these analyses of inter-species transmission were both affected by sampling biases and emphasise that WGS is most powerful in combination with appropriate sampling design.

The WGS data analysed, combined with detailed epidemiological information, were able to provide additional evidence that wildlife populations, the brushtail possum in New Zealand and the European badger in the UK, are important in the spread and persistence of bTB in cattle populations. Therefore, these analyses suggest that surveillance, and possibly control, of *M. bovis* infection in wildlife reservoirs is necessary and should be accompanied by WGS of the any sampled *M. bovis*. In New Zealand, vector control operations currently target possum populations and effectively limit the risk of infection in livestock. In the UK, promising vaccination projects, like the Test-Vaccinate-Release trial in Northern Ireland, aim to limit the maintenance ability of badger populations, whilst protecting this iconic species. Cattle movements were also highlighted as an important means of spread, which emphasises the continued need for cattle testing and movement surveillance.

Chapter A

Appendix A: Description of simulation model designed to investigate the sampling biases evident in Chapter 3

A simulation modelling framework was designed to investigate the influence of the sampling patterns evident in Chapter 3. For this chapter, the Whole Genome Sequenced (WGS) *M. bovis* isolates were sampled over 30 years. This sampling window was relatively narrow, considering that the Most Recent Common Ancestor (MRCA) was estimated to have been circulating in 1859 (2.5% Lower: 1525, 97.5% Upper: 1936). In addition, the isolates sourced from wildlife were, in general, sampled earlier than the cattle isolates (Figure 3.2).

A.1 Description of the Model

A Susceptible-Infected-Removed (SIR) individual based compartmental infection model was developed, under the assumption of density dependent transmission (Equation A.1) [236]. The infection probability, p, of a single susceptible individual was a function of the number of infected individuals, I, and the per contact transmission probability T. The susceptible population size was kept constant by replacing any individual that was removed.

$$p = 1 - (1 - T)^I \tag{A.1}$$

The pathogen in the simulation model was incorporated as a mutating sequence. Each infected individual, during a simulation, carried a sequence of mutation events inherited from its source. Upon the completion of a simulation each mutation event was mapped onto a randomly generated reference genome using a Jukes Cantor substitution model [237].

A.2 Model Parameters

The simulation model was parameterized to resemble a population infected with *Mycobacterium bovis* (Table A.1).

Table A.1: The parameter settings for the simulation model.

Parameter	Symbol	Value	Units
Population size	N	1000	individuals
Per contact transmission probability	T	0.1	individual/time-step
Removal probability	r	0.2	individual/time-step
Genome size	g	4,500,000	base pairs
Substitution probability	m	0.5	per genome per time-step
Simulation length	l	120	time-steps (years)

A.3 Sampled Transmission Tree

During a simulation, the path of transmission was stored by recording who infected whom in an adjacency matrix, M. $M_{ij} = 1$ when individual i infected individual j and $M_{ij} = 0$ when there was no transmission.

A sampled transmission tree was constructed to resemble an ideal phylogenetic tree. This tree was built, based on the sampled population, using two steps:

1. Any un-sampled individuals with an out degree equal to zero were recursively removed.
2. Any un-sampled individuals with an out degree equal to one were removed.

A.4 Changing the Sampling Window

A set of simulations were completed using sampling windows of different sizes (Table A.2). Each sampling window was defined by the start time-step, after which sampling could occur, and the end time-step, when no more sampling could be conducted. Randomly sampled infectious individuals were removed and their specific sequences of mutation events were recorded. The nucleotide sequences, created for each sampled individual, were analysed in the program BEAST (v1.8.4, [238]). BEAST analyses were ran using a Jukes Cantor substitution model, a constant population size model and a Strict Clock model. A Uniform prior distribution was used to estimate the substitution rate. A chain length of 10,000,000 steps was used, sampling every 1,000 steps (burn-in = 10%). The phylogenetic tree estimated in BEAST should partially resemble the sampled transmission tree. The BEAST substitution rate estimate was compared to the rate estimated on the sampled transmission tree, so that the estimates being compared were both based on a tree like structure. The substitution rate was

estimated on the sampled transmission tree by using the observed genetic distance between the sources and sinks on the branches of the tree and length of time (in time-steps) between the infection events for the source and sink (branch lengths).

Table A.2: A table showing how the four different sampling windows were defined. Each sampling window was applied to 10 independent simulations.

Number Of Time-steps	Time-steps Sampled	Proportion Sampled Per Time-step	Number Sampled
100	20-120	0.001	100
50	70-120	0.002	100
25	95-120	0.004	100
10	110-120	0.01	100

A.5 Estimating State Transitions

The individuals in the simulation model's homogeneous population, described above, were assigned a state (A or B). Simulations were conducted to evaluate how applying different biased sampling strategies could influence the estimation of inter-state transmission rates.

For a two-state system, Equation A.1 was changed to incorporate state transition probabilities (Equation A.2).

$$p_A = 1 - [(1 - T * Q_{AA})^{I_A} * (1 - T * Q_{BA})^{I_B}] \tag{A.2}$$

The probability, p_A, of a single susceptible individual, of state A, was a function of the number of infected individuals in each state, I_A and I_B, and the per contact transmission probability T. Each infected individual's infectiousness was scaled by the state transition probability defined in the matrix Q:

$$Q = \begin{pmatrix} & A & B \\ A & 0.8 & 0.2 \\ B & 0.2 & 0.8 \end{pmatrix}$$

For the estimation of the state transition rates, involving the off diagonal elements of Q, Pagel *et al.*'s method [214] was modified to approximate the rates based upon a transmission tree. The approximation was simplified, to avoid the necessity for Markov Chain Monte Carlo estimations, by retaining the states and timings of the ancestral nodes on the transmission tree. For a given set of state transition rates Q, the likelihood of a given branch, of length t, beginning in state i and ending in state j was given by $P(t)_{ij}$. The matrix P, for a given branch length t, was calculated as the matrix exponential of the matrix Q, multiplied by time, t (Equation A.3). The likelihood of the transmission tree was the product of the branch-specific likelihoods.

$$P(t) = e^{Q*t} \tag{A.3}$$

Three different sampling strategies were evaluated, each sampling 100 individuals. These strategies were designed to replicate the sampling patterns evident in Chapter 3:

1. Individuals of either state had an equal probability of being sampled.
2. Individuals of state A were preferentially sampled.
3. Individuals of state A were sampled early whilst individuals of state B were sampled later.

For each scenario the state transition rates were estimated on the complete and sampled transmission trees using the likelihood function described above. Each simulated scenario was replicated 100 times.

A.6 Modelling output

A.6.1 Changing the Sampling Window

As the size of the sampling window was reduced, the estimations of the substitution rate remained accurate (Figure A.1). The rate estimates in Figure A.1 were higher than the mutation rate specified in Table A.1 and closer to the rate estimated using the sampled transmission tree.

A.6.2 Estimating State Transitions

When sampling biases were introduced, the state transition rates estimated on the sampled transmission tree were inaccurate (Figure A.2). Sampling only 100 individuals reduced the precision of the rate estimates. Sampling individuals of state A more than those of state B meant the state transition rate from B to A was overestimated. Sampling individuals of state A earlier meant the state transition rate from A to B was slightly overestimated and highly variable.

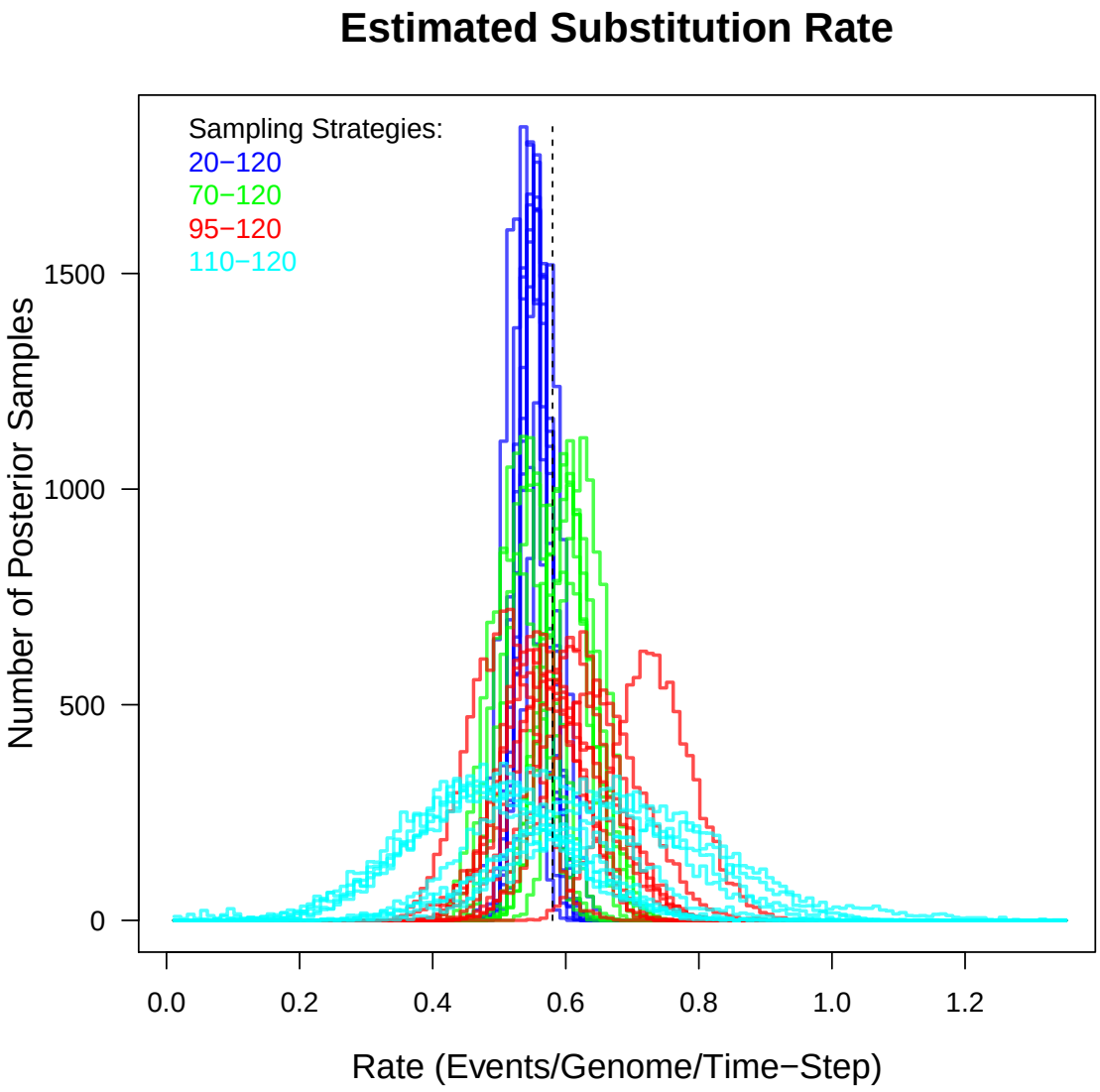

Figure A.1: The influence of the sampling window size on the estimation of the substitution rate. Histograms of the sampled (n=9000, 10% burn-in removed) posterior distributions for the substitution rate estimates produced by BEAST are plotted. Using sampling windows of varying sizes - 100 time-steps (blue), 50 time-steps (green), 25 time-steps (red) and 10 time-steps (cyan) - 100 individuals were sampled from a simulated outbreak and their sequences were analysed in BEAST. 10 replicates were completed for each of the four window sizes. The substitution rate estimated on the sampled transmission tree was shown as a vertical dashed black line.

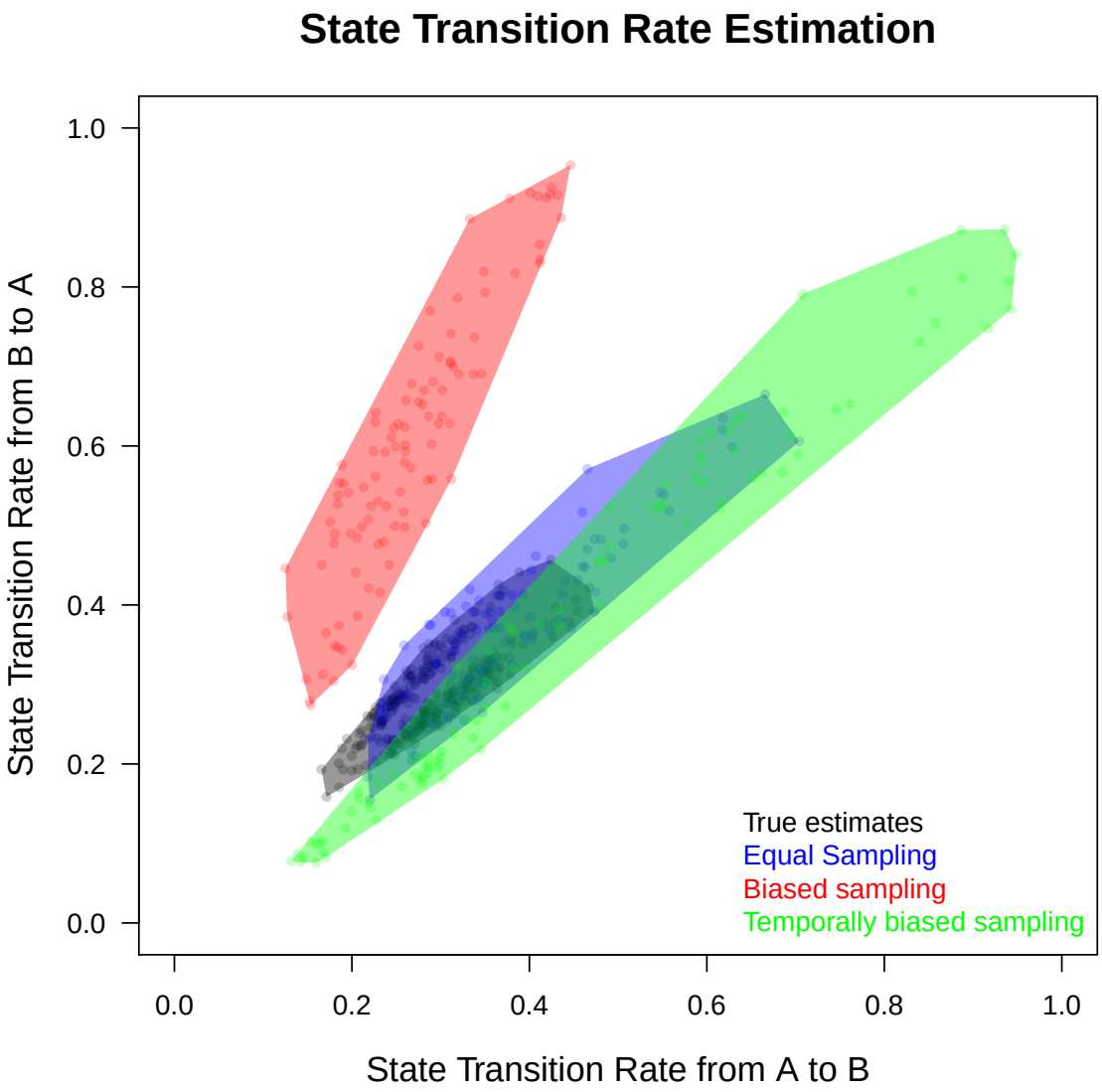

Figure A.2: The influence of different sampling regimes on the estimated state transition rates. Three different sampling scenarios were investigated and the state transition rates were estimated using the sampled transmission tree. The state transition rate estimates based on the sampled transmission tree were compared to those estimated with the full transmission tree (black). A minimum convex polygon was drawn around the estimates, taken from 100 simulations, for each sampling regime. 100 individuals were sampled over 100 time-steps in each of the regimes, equally (blue - scenario 1), with a bias towards state A (red - scenario 2), with individuals of state A sampled in the first 50 time-steps and individuals in state B sampled thereafter (green - scenario 3).

Chapter B

Appendix B: Investigating the mislabelling of the Whole Genome Sequenced isolates used in Chapter 4

Woodchester park, in the southwest of England, is home to a high density badger population. This population is naturally infected with *Mycobacterium bovis*, the causative agent of bovine tuberculosis (bTB). Badgers living in this population are routinely captured and tested for bTB [185]. Two hundred and thirty *M. bovis* isolates from samples, taken during trapping operations, were Whole Genome Sequenced (WGS) and preliminary analyses of these data revealed some mislabelled isolates. These isolates were spoligotyped on two separate occasions and the types didn't match for five isolates.

The aim of this investigation was to use the epidemiological data to determine the extent of the mislabelling of the WGS badger isolates from Woodchester Park (Chapter 4). This investigation assumes that mislabelling will result in the isolate being decoupled from its sampled badger, and the epidemiological data associated with that badger. As a result the genetic relationships between the mislabelled isolates and all other isolates shouldn't agree well with the epidemiological data describing how the sampled badgers were related.

B.1 Comparing the genetic and epidemiological data

Of the 230 WGS *M. bovis* isolates available, 176 had sufficient sequence quality ($\geq$90% of the *M. bovis* genome with read depth $\geq$20) to be used in further analyses. The mislabelling investigation was only conducted on these 176 isolates. Each of the 176 remaining WGS *M. bovis* isolates were associated with an individual badger as identified by it's tattoo. Using the tattoo it was possible to examine each of the sampled badger's capture histories. In order to determine whether any further isolates were mislabelled, the available WGS data were compared to the epidemiological data. To compare each isolate's genetic data, inter-isolate genetic distances were calculated by counting the number of sites that differed between isolates. Epidemiological metrics, described in Table 2.4 were used to make inter-isolate epidemiological comparisons.

Inter-isolate comparisons generated epidemiological metrics that were fitted to the respective pairwise genetic distances between the isolates. A Random Forest regression model [197] was used to fit the epidemiological metrics to the genetic distances in the statistical programming environment R (v3.2.3 [194]). Once the Random Forest model was fitted, it was used to predict the inter-isolate genetic distances. Similarly, a Boosted Regression model [227] was used to fit the epidemiological metrics to the inter-isolate genetic distances and predict those same distances. These different machine learning approaches, the Random Forest and Boosted Regression models, were used independently to increase the robustness of the analyses.

As Figure B.1 demonstrates, the isolates originate from a number of highly distinct clades producing a skewed inter-isolate genetic distance distribution as a result of having multiple, highly distinct clades. To avoid the Random Forest or Boosted Regression models fitting to very large genetic distances between clades that very distantly epidemiologically related and unlikely to be captured by the available data, the epidemiological metrics were only fitted to genetic distances that were <15 SNPs - i.e. within-clade distances (Figure B.1).

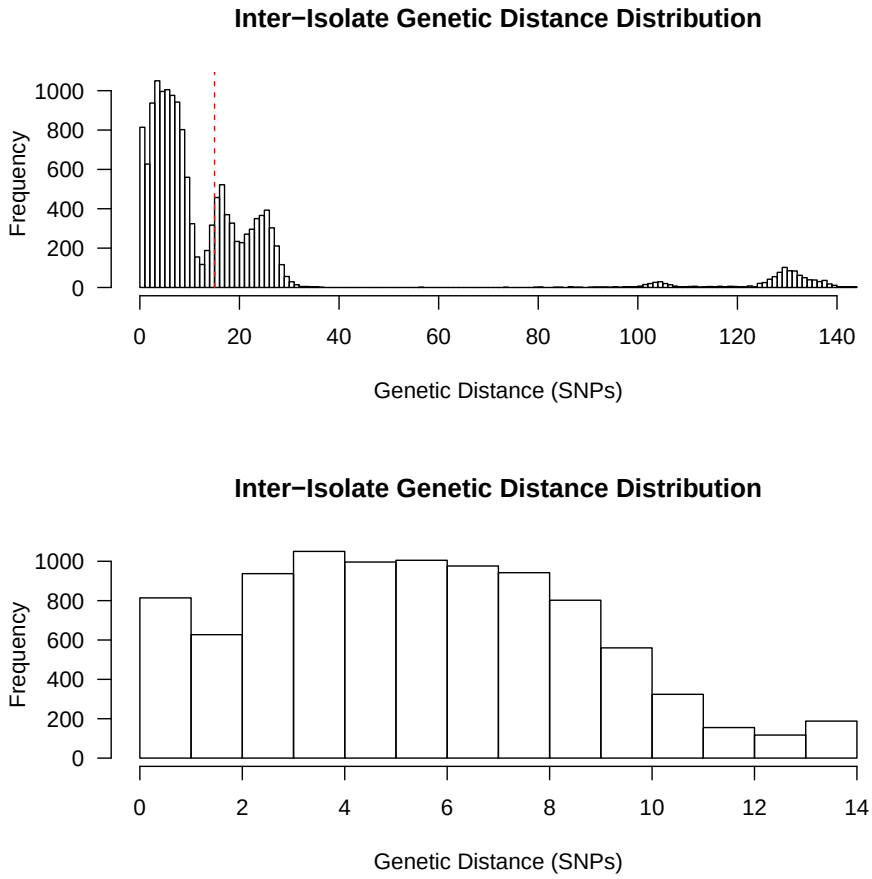

Figure B.1: Histograms of the inter-isolate genetic distance distribution. The genetic distance was calculated as the number of differences (Single Nucleotide Polymorphisms - SNPs) between the isolates. TOP: The full inter-isolate genetic distance distribution, the vertical red dashed line marks a genetic distance of 15. BOTTOM: The inter-isolate genetic distance distribution for distances <15 SNPs.

B.2 Shuffling the isolates

To examine the effect of further isolate shuffling, on the fitting of the Random Forest model, an increasing proportion of the isolates were shuffled, and the impact of shuffling on the fit of epidemiological metrics to the inter-isolate genetic distances was investigated. A range of shuffling proportions were examined, from 0 to 1 in steps of 0.05. For each shuffling proportion, 10 replicates were completed.

B.3 Output from comparing the genetic and epidemiological data

The predicted inter-isolate genetic distances from both the Random Forest and Boosted Regression fitted models showed a good agreement to the actual genetic distances. A Pearsons correlation between the predicted and actual distances, produced a value of 0.83 for the Random Forest model and 0.89 for the Boosted Regression model.

The good agreement between the predicted and actual inter-isolate genetic distances resulted in a low median difference between them (Figure B.2 & Figure B.3). There were some clear outliers - isolates for which both the fitted Random Forest and Boosted Regression models weren't able to accurately predict the genetic distance between them and other isolates. These outliers appear to have been sampled around 2003 and included three of the five isolates that were identified as mislabelled (Table B.1). In addition, seven of the outliers of the fitted Random Forest and Boosted Regression models were the same.

Table B.1: The isolates that were poorly fitted by the Random Forest and/or the Boosted Regression models and those that had spoligotype mismatches.

Isolate ID	Outlier Random Forest	Outlier Boosted Regression	Spoligotype Mismatch
WB65	YES	YES	NO
WB71	NO	YES	YES
WB72	NO	NO	YES
WB74	YES	YES	NO
WB75	NO	YES	NO
WB91	YES	NO	NO
WB96	YES	NO	NO
WB98	YES	YES	NO
WB99	YES	YES	NO
WB100	YES	YES	YES
WB105	YES	YES	YES
WB106	YES	YES	NO
WB107	NO	NO	YES

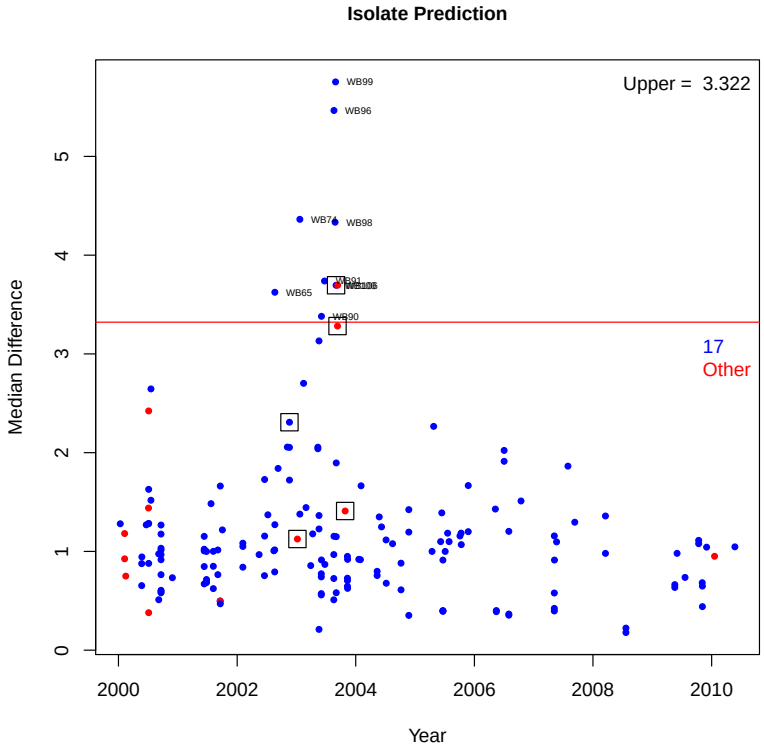

Figure B.2: Each isolate's sampling date against the median difference between the actual and predicted inter-isolate genetic distances produced by the fitted Random Forest model. A horizontal red line is plotted at the 95% percentile. The isolate labels are shown for those isolates whose median difference falls in the highest 5% of the median difference distribution. Those points highlighted with a square are the median values for the isolates that are known to be mislabelled. Isolates of spoligotype 17 and other types are represented by blue and red circles, respectively.

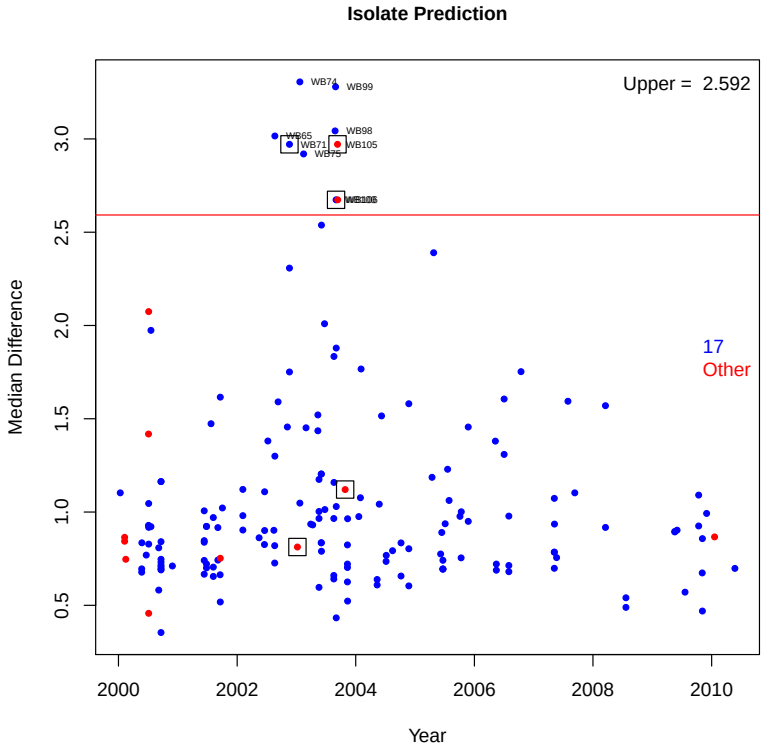

Figure B.3: Each isolate's sampling date against the median difference between the actual and predicted inter-isolate genetic distances produced by the fitted Boosted Regression model. A horizontal red line is plotted at the 95% percentile. The isolate labels are shown for those isolates whose median difference falls in the highest 5% of the median difference distribution. Those points highlighted with a square are the median values for the isolates that are known to be mislabelled. Isolates of spoligotype 17 and other types are represented by blue and red circles, respectively.

B.4 Output from further shuffling

With no shuffling the fitted Random Forest model reports that approximately 70% of the variation in the inter-isolate genetic distances was explained by the epidemiological metrics (Figure B.4). As the shuffling proportion of the isolates increases towards one, the variation explained by the Random Forest model declined towards zero.

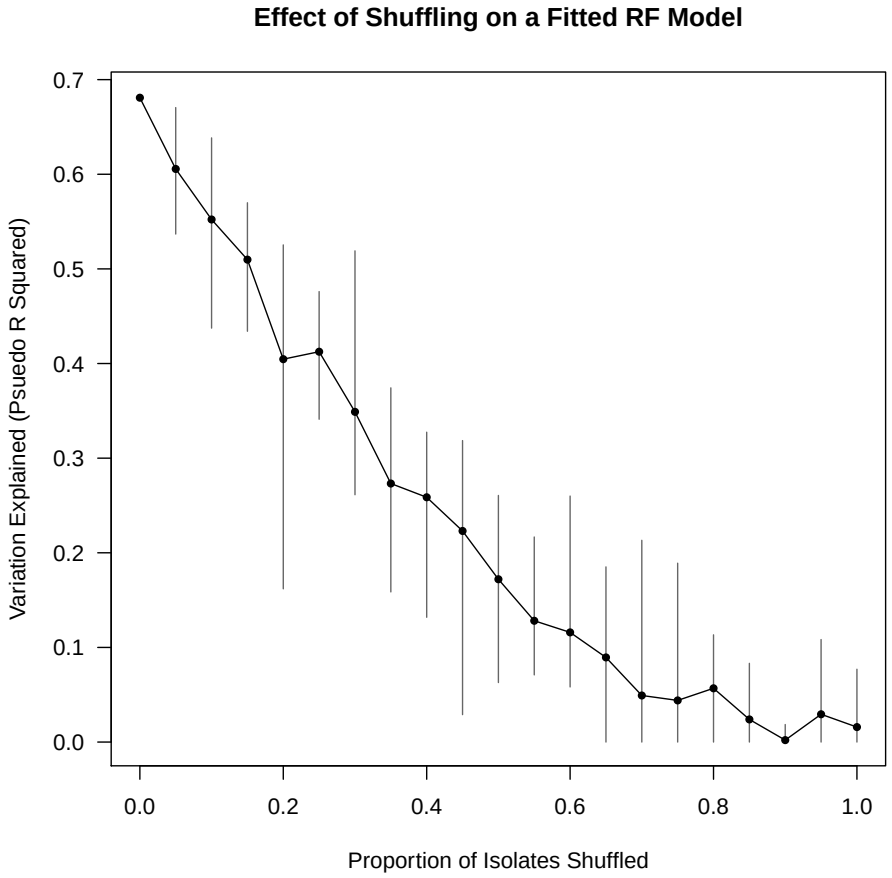

Figure B.4: The effect of shuffling a varying proportion of the isolate sequences on the variation explained by a fitted Random Forest model. The mean of the 10 replicates completed for each shuffling proportion was shown as a black point. The vertical grey lines span the range of values between the minimum and maximum variation explained by the fitted Random Forest model.

Bibliography

[1] W. Y. Ayele, S. D. Neill, J. Zinsstag, M. G. Weiss, and I. Pavlik, "Bovine tuberculosis: An old disease but a new threat to Africa," *The International Journal of Tuberculosis and Lung Disease*, vol. 8, no. 8, pp. 924–37, 2004.

[2] I. N. de Kantor and V. Ritacco, "An update on bovine tuberculosis programmes in Latin American and Caribbean countries," *Veterinary Microbiology*, vol. 112, pp. 111–118, 2006.

[3] C. Thoen, P. Lobue, and I. de Kantor, "The importance of *Mycobacterium bovis* as a zoonosis," *Veterinary Microbiology*, vol. 112, pp. 339–45, 2006.

[4] F. J. Reviriego Gordejo and J. P. Vermeersch, "Towards eradication of bovine tuberculosis in the European Union," *Veterinary Microbiology*, vol. 112, no. 2-4, pp. 101–9, 2006.

[5] M. De Garine-Wichatitsky, A. Caron, R. Kock, R. Tschop, M. Munyeme, M. Hofmeyr, and A. Michel, "A review of bovine tuberculosis at the wildlife–livestock–human interface in sub-Saharan Africa," *Epidemiology and Infection*, vol. 141, no. 07, pp. 1342–1356, 2013.

[6] H. C. J. Godfray, C. A. Donnelly, R. R. Kao, D. W. Macdonald, R. McDonald, G. Petrokofsky, J. L. N. Wood, R. Woodroffe, D. B. Young, and A. R. McLean, "A restatement of the natural science evidence base relevant to the control of bovine tuberculosis in Great Britain," *Proceedings of the Royal Society B: Biological Sciences*, vol. 280, no. 1768, pp. 1–18, 2013.

[7] P. G. Livingstone, N. Hancox, G. Nugent, and G. W. de Lisle, "Toward eradication: The effect of *Mycobacterium bovis* infection in wildlife on the evolution and future direction of bovine tuberculosis management in New Zealand," *New Zealand Veterinary Journal*, vol. 63 Suppl 1, no. December 2014, pp. 4–18, 2014.

[8] "Council Directive of 26 June 1964 on animal health problems affecting intra-community trade in bovine animals and swine," Tech. Rep. 61, Official Journal of the European Comunities, 1964.

[9] D. V. Cousins and J. L. Roberts, "Australia's campaign to eradicate bovine tuberculosis: The battle for freedom and beyond," *Tuberculosis*, vol. 81, no. 1-2, pp. 5–15, 2001.

[10] S. Hall, "Official bovine tuberculosis-free status in Scotland," *The Veterinary Record*, vol. 166, no. 8, pp. 245–6, 2010.

[11] R. J. Delahay, A. N. S. De Leeuw, A. M. Barlow, R. S. Clifton-Hadley, and C. L. Cheeseman, "The status of *Mycobacterium bovis* infection in UK wild mammals: A review," *The Veterinary Journal*, vol. 164, no. 2, pp. 90–105, 2002.

[12] S. Schmitt, D. J. O'Brien, C. S. Bruning-Fann, and S. D. Fitzgerald, "Bovine tuberculosis in Michigan wildlife and livestock," *New York Academy of Sciences*, vol. 969, pp. 262–268, 2002.

[13] L. A. L. Corner, "The role of wild animal populations in the epidemiology of tuberculosis in domestic animals: How to assess the risk," *Veterinary Microbiology*, vol. 112, pp. 303–312, 2006.

[14] R. S. Miller and S. J. Sweeney, "*Mycobacterium bovis* (bovine tuberculosis) infection in North American wildlife: Current status and opportunities for mitigation of risks of further infection in wildlife populations," *Epidemiology and Infection*, vol. 141, no. 7, pp. 1357–70, 2013.

[15] R. Delahay, "Control of bovine tuberculosis in New Zealand in the face of a wildlife host: A compiled review of 50 years of programme policy, design and research," *New Zealand Veterinary Journal*, vol. 63, pp. 2–3, 2015.

[16] G. Nugent, B. Buddle, and G. Knowles, "Epidemiology and control of *Mycobacterium bovis* infection in brushtail possums (*Trichosurus vulpecula*), the primary wildlife host of bovine tuberculosis in New Zealand," *New Zealand Veterinary Journal*, vol. 63, no. sup1, pp. 28–41, 2015.

[17] Parliment UK, "Protection of Badgers Act 1992," Tech. Rep. September, 1992.

[18] C. Gortazar, R. J. Delahay, R. A. Mcdonald, M. Boadella, G. J. Wilson, D. Gavier-Widen, and P. Acevedo, "The status of tuberculosis in European wild mammals," *Mammal Review*, vol. 42, no. 3, pp. 193–206, 2012.

[19] R. De La Rua-Domenech, "Human *Mycobacterium bovis* infection in the United Kingdom: Incidence, risks, control measures and review of the zoonotic aspects of bovine tuberculosis," *Tuberculosis*, vol. 86, no. 2, pp. 77–109, 2006.

[20] J. A. Drewe, A. J. Tomlinson, N. J. Walker, and R. J. Delahay, "Diagnostic accuracy and optimal use of three tests for tuberculosis in live badgers," *PLoS ONE*, vol. 5, no. 6, p. e11196, 2010.

[21] G. W. de Lisle, G. F. Yates, D. M. Collins, R. W. MacKenzie, K. B. Crews, and R. Walker, "A study of bovine tuberculosis in domestic animals and wildlife in the MacKenzie Basin and surrounding areas using DNA fingerprinting," *New Zealand Veterinary Journal*, vol. 43, no. 7, pp. 266–71, 1995.

[22] C. Furphy, E. Costello, D. Murphy, L. A. L. Corner, and E. Gormley, "DNA typing of *Mycobacterium bovis* isolates from badgers (*Meles meles*) culled from areas in Ireland with different levels of tuberculosis prevalence.," *Veterinary Medicine International*, pp. 1–6, 2012.

[23] A. Balseiro, P. González-Quirós, Ó. Rodríguez, M. Francisca C, I. Merediz, L. de Juan, M. A. Chambers, R. J. Delahay, N. Marreros, L. J. Royo, J. Bezos, J. M. Prieto, and C. Gortázar, "Spatial relationships between Eurasian badgers (*Meles meles*) and cattle infected with *Mycobacterium bovis* in Northern Spain," *The Veterinary Journal*, vol. 197, no. 3, pp. 739–745, 2013.

[24] R. Biek, A. O'Hare, D. Wright, T. Mallon, C. McCormick, R. J. Orton, S. McDowell, H. Trewby, R. A. Skuce, and R. R. Kao, "Whole genome sequencing reveals local transmission patterns of *Mycobacterium bovis* in sympatric cattle and badger populations," *PLoS Pathogens*, vol. 8, no. 11, 2012.

[25] H. Trewby, D. Wright, E. L. Breadon, S. J. Lycett, T. R. Mallon, C. McCormick, P. Johnson, R. J. Orton, A. R. Allen, J. Galbraith, P. Herzyk, R. A. Skuce, R. Biek, and R. R. Kao, "Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis," *Epidemics*, vol. 14, pp. 26–35, 2016.

[26] L. Glaser, M. Carstensen, S. Shaw, S. Robbe-Austerman, A. Wunschmann, D. Grear, T. Stuber, and B. Thomsen, "Descriptive epidemiology and whole genome sequencing analysis for an outbreak of bovine tuberculosis in beef cattle and white-tailed deer in northwestern Minnesota," *PloS ONE*, vol. 11, no. 1, pp. 1–21, 2016.

[27] J. M. Pollock and S. D. Neill, "*Mycobacterium bovis* infection and tuberculosis in cattle," *Veterinary Journal*, vol. 163, no. 2, pp. 115–27, 2002.

[28] G. S. Dean, S. G. Rhodes, M. Coad, O. Whelan, P. J. Cockle, D. J. Clifford, R. G. Hewinson, and H. M. Vordermeier, "Minimum effective dose of *Mycobacterium bovis* in cattle," *Infection and Immunity*, vol. 73, no. 10, pp. 6467–6471, 2005.

[29] V. Goodchild and R. S. Clifton-Hadley, "Cattle-to-cattle transmission of *Mycobacterium bovis*," *Tuberculosis (Edinburgh, Scotland)*, vol. 81, no. 1-2, pp. 23–41, 2001.

[30] R. de la Rua-Domenech, A. T. Goodchild, H. M. Vordermeier, R. G. Hewinson, K. H. Christiansen, and R. S. Clifton-Hadley, "Ante mortem diagnosis of tuberculosis in cattle: A review of the tuberculin tests, γ-interferon assay and other ancillary diagnostic techniques," *Research in Veterinary Science*, vol. 81, no. 2, pp. 190–210, 2006.

[31] J. P. Cassidy, "The pathogenesis and pathology of bovine tuberculosis with insights from studies of tuberculosis in humans and laboratory animal models," *Veterinary Microbiology*, vol. 112, no. 2-4, pp. 151–161, 2006.

[32] A. J. K. Conlan, T. J. McKinley, K. Karolemeas, E. Brooks-Pollock, A. V. Goodchild, A. P. Mitchell, C. P. D. Birch, R. S. Clifton-Hadley, and J. L. N. Wood, "Estimating the Hidden Burden of Bovine Tuberculosis in Great Britain," *PLoS Computational Biology*, vol. 8, no. 10, pp. 1–14, 2012.

[33] J. J. O'Keeffe, "Description of a medium term national strategy toward eradication of Tuberculosis in cattle in Ireland," *Proceedings of Veterinary Epidemiology and Economics*, vol. 11, pp. 1–7, 2006.

[34] DEFRA, "Strategy for consultation on a strategy for achieving "Officially Bovine Tuberculosis-Free" status for England," tech. rep., Department for Environment Food & Rural Affairs, 2013.

[35] E. Costello, J. W. A. Egan, F. C. Quigley, and P. F. O'Reilly, "Performance of the single intradermal comparative tuberculin test in identifying cattle with tuberculous lesions in Irish herds," *Veterinary Record*, vol. 141, no. 9, pp. 222–224, 1997.

[36] B. Buddle, G. de Lisle, J. Griffin, and S. Hutchings, "Epidemiology, diagnostics, and management of tuberculosis in domestic cattle and deer in New Zealand in the face of a wildlife reservoir," *New Zealand Veterinary Journal*, vol. 63, pp. 19–27, 2015.

[37] T. A. Clegg, A. Duignan, C. Whelan, E. Gormley, M. Good, J. Clarke, N. Toft, and S. J. More, "Using latent class analysis to estimate the test characteristics of the γ-interferon test, the single intradermal comparative tuberculin test and a multiplex immunoassay under Irish conditions," *Veterinary Microbiology*, vol. 151, no. 1-2, pp. 68–76, 2011.

[38] S. A. J. Strain, J. MacNair, and S. W. J. Mcdowell, "A review of the international application of the Interferon-gamma test," Tech. Rep. March, Agri-Food and Biosciences Institute, 2012.

[39] J. Álvarez, A. Perez, J. Bezos, S. Marqués, A. Grau, J. L. Saez, O. Mínguez, L. de Juan, and L. Domínguez, "Evaluation of the sensitivity and specificity of bovine tuberculosis diagnostic tests in naturally infected cattle herds using a Bayesian approach," *Veterinary Microbiology*, vol. 155, no. 1, pp. 38–43, 2012.

[40] J. Claridge, P. Diggle, C. M. McCann, G. Mulcahy, R. Flynn, J. McNair, S. Strain, M. Welsh, M. Baylis, and D. J. L. Williams, "*Fasciola hepatica* is associated with the failure to detect bovine tuberculosis in dairy cattle," *Nature Communications*, vol. 3, no. 853, pp. 1–8, 2012.

[41] L. Garza-Cuartero, J. O'Sullivan, A. Blanco, J. McNair, M. Welsh, R. J. Flynn, D. Williams, P. Diggle, J. Cassidy, and G. Mulcahy, "*Fasciola hepatica* infection

reduces *Mycobacterium bovis* burden and mycobacterial uptake and suppresses the pro-inflammatory response," *Parasite Immunology*, vol. 38, no. 7, pp. 387–402, 2016.

[42] E. Gormley, M. B. Doyle, T. Fitzsimons, K. McGill, and J. D. Collins, "Diagnosis of *Mycobacterium bovis* infection in cattle by use of the gamma-interferon (Bovigam) assay," *Veterinary Microbiology*, vol. 112, no. 2-4, pp. 171–9, 2006.

[43] E. Liebana, L. Johnson, J. Gough, P. Durr, K. Jahans, R. Clifton-Hadley, Y. Spencer, R. G. Hewinson, and S. H. Downs, "Pathology of naturally occurring bovine tuberculosis in England and Wales," *Veterinary Journal*, vol. 176, no. 3, pp. 354–60, 2008.

[44] W. I. Morrison, F. J. Bourne, D. R. Cox, C. A. Donnelly, G. Gettinby, J. P. McInerney, and R. Woodroffe, "Pathogenesis and diagnosis of infecions with *Mycobacterium bovis* in cattle," 2000.

[45] B. J. Wards, D. M. Collins, and G. W. de Lisle, "Detection of *Mycobacterium bovis* in tissues by polymerase chain reaction," *Veterinary Microbiology*, vol. 43, no. 2-3, pp. 227–240, 1995.

[46] A. Mishra, V. M. Katoch, K. Srivastava, S. S. Thakral, S. S. Bharadwaj, V. Sreenivas, and H. K. Prasad, "Direct detection and identification of *Mycobacterium tuberculosis* and *Mycobacterium bovis* in bovine samples by a novel nested PCR assay: Correlation with conventional techniques," *Journal of Clinical Microbiology*, vol. 43, no. 6, pp. 5670–5678, 2005.

[47] B. M. Buddle, F. E. Aldwell, M. A. Skinner, G. W. Lisle, M. Denis, H. M. Vordermeier, R. G. Hewinson, and D. N. Wedlock, "Effect of oral vaccination of cattle with lipid-formulated BCG on immune responses and protection against bovine tuberculosis," *Vaccine*, vol. 23, no. 27, pp. 3581–3589, 2005.

[48] D. N. Wedlock, M. Denis, H. M. Vordermeier, R. G. Hewinson, and B. M. Buddle, "Vaccination of cattle with Danish and Pasteur strains of *Mycobacterium bovis* BCG induce different levels of IFNγ post-vaccination, but induce similar levels of protection against bovine tuberculosis," *Veterinary Immunology and Immunopathology*, vol. 118, pp. 50–58, 2007.

[49] H. M. Vordermeier, B. Villarreal-Ramos, P. J. Cockle, M. McAulay, S. G. Rhodes, T. Thacker, S. C. Gilbert, H. McShane, A. V. S. Hill, Z. Xing, and R. G. Hewinson, "Viral booster vaccines improve *Mycobacterium bovis* BCG-induced protection against bovine tuberculosis," *Infection and Immunity*, vol. 77, no. 8, pp. 3364–73, 2009.

[50] B. M. Buddle, D. N. Wedlock, M. Denis, H. M. Vordermeier, and R. G. Hewinson, "Update on vaccination of cattle and wildlife populations against tuberculosis," *Veterinary Microbiology*, vol. 151, no. 1-2, pp. 14–22, 2011.

[51] M. L. Thom, M. McAulay, H. M. Vordermeier, D. Clifford, R. G. Hewinson, B. Villarreal-Ramos, and J. C. Hope, "Duration of immunity against *Mycobacterium bovis* following neonatal vaccination with bacillus Calmette-Guérin Danish: Significant protection against infection at 12, but not 24, months," *Clinical and Vaccine Immunology*, vol. 19, no. 8, pp. 1254–60, 2012.

[52] J. M. Pollock, J. McNair, H. Bassett, J. P. Cassidy, E. Costello, H. Aggerbeck, I. Rosenkrands, and P. Andersen, "Specific delayed-type hypersensitivity responses to ESAT-6 identify tuberculosis-infected cattle," *Journal of Clinical Microbiology*, vol. 41, no. 5, pp. 1856–60, 2003.

[53] A. O. Whelan, D. Clifford, B. Upadhyay, E. L. Breadon, J. McNair, G. R. Hewinson, and M. H. Vordermeier, "Development of a skin test for bovine tuberculosis for differentiating infected from vaccinated animals," *Journal of Clinical Microbiology*, vol. 48, no. 9, pp. 3176–3181, 2010.

[54] H. M. Vordermeier, G. J. Jones, B. M. Buddle, and R. G. Hewinson, "Development of immune-diagnostic reagents to diagnose bovine tuberculosis in cattle," *Veterinary Immunology and Immunopathology*, pp. 3–7, 2016.

[55] C. Whelan, E. Shuralev, G. O'Keeffe, P. Hyland, H. F. Kwok, P. Snoddy, A. O'Brien, M. Connolly, P. Quinn, M. Groll, T. Watterson, S. Call, K. Kenny, A. Duignan, M. J. Hamilton, B. M. Buddle, J. Johnston, W. C. Davis, S. Olwill, and J. Clarke, "Multiplex immunoassay for serological diagnosis of *Mycobacterium bovis* infection in cattle," *Clinical and Vaccine Immunology*, vol. 15, no. 12, pp. 1834–1838, 2008.

[56] C. Whelan, E. Shuralev, H. F. Kwok, K. Kenny, A. Duignan, M. Good, W. C. Davis, and J. Clarke, "Use of a multiplex enzyme-linked immunosorbent assay to detect a subpopulation of *Mycobacterium bovis*-infected animals deemed negative or inconclusive by the single intradermal comparative tuberculin skin test," *Journal of Veterinary Diagnostic Investigation*, vol. 23, no. 3, pp. 499–503, 2011.

[57] D. Abernethy, P. Upton, I. M. Higgins, G. McGrath, V. Goodchild, S. J. Rolfe, J. M. Broughan, S. H. Downs, R. Clifton-Hadley, F. D. Menzies, R. de la Rua-Domenech, M. J. Blissitt, A. Duignan, and S. J. More, "Bovine tuberculosis trends in the UK and the Republic of Ireland, 1995–2010," *Veterinary Record*, vol. 172, no. 12, pp. 312–312, 2013.

[58] P. G. Livingstone, N. Hancox, G. Nugent, G. Mackereth, and S. A. Hutchings, "Development of the New Zealand strategy for local eradication of tuberculosis from wildlife and livestock," *New Zealand Veterinary Journal*, pp. 98–107, 2015.

[59] OSPRI, "Our first year | 2013 / 2014," tech. rep., OSPRI, 2014.

[60] OSPRI, "Annual Report 2014-2015," tech. rep., OSPRI, 2015.

[61] S. A. Hutchings, N. Hancox, and P. G. Livingstone, "A strategic approach to eradication of bovine TB from wildlife in New Zealand," *Transboundary and Emerging Diseases*, vol. 60 Suppl 1, no. SUPPL1, pp. 85–91, 2013.

[62] N. H. Smith, S. V. Gordon, R. de la Rua-Domenech, R. S. Clifton-Hadley, and R. G. Hewinson, "Bottlenecks and broomsticks: The molecular evolution of *Mycobacterium bovis*," *Nature Reviews. Microbiology*, vol. 4, no. 9, pp. 670–681, 2006.

[63] J. J. Carrique-Mas, G. F. Medley, and L. E. Green, "Risks for bovine tuberculosis in British cattle farms restocked after the foot and mouth disease epidemic of 2001," *Preventive Veterinary Medicine*, vol. 84, no. 1-2, pp. 85–93, 2008.

[64] B. Warburton and P. Livingstone, "Managing and eradicating wildlife tuberculosis in New Zealand," *New Zealand Veterinary Journal*, pp. 1–37, 2015.

[65] T. Roper, *Badger*. HarperCollins UK, 2010.

[66] Parliment UK, "Badgers Act 1973," tech. rep., 1973.

[67] E. Costello, M. L. Doherty, M. L. Monaghan, F. C. Quigley, and P. F. O'Reilly, "A study of cattle-to-cattle transmission of *Mycobacterium bovis* infection," *The Veterinary Journal*, vol. 155, no. 3, pp. 245–250, 1998.

[68] J. P. Cassidy, D. G. Bryson, J. M. Pollock, R. T. Evans, F. Forster, and S. D. Neill, "Lesions in cattle exposed to *Mycobacterium bovis* - inoculated Calves," *Journal of Comparative Pathology*, vol. 121, no. 4, pp. 321–337, 1999.

[69] L. E. Green and S. J. Cornell, "Investigations of cattle herd breakdowns with bovine tuberculosis in four counties of England and Wales using VETNET data," *Preventive Veterinary Medicine*, vol. 70, no. 3-4, pp. 293–311, 2005.

[70] M. A. Schoenbaum, B. H. Espe, and B. Behring, "Epidemic of bovine tuberculosis cases originating from an infected beef herd in Oklahoma, USA," *Preventive Veterinary Medicine*, vol. 13, no. 2, pp. 113–120, 1992.

[71] J. W. Chalmers, F. Jamieson, and P. Rafferty, "An outbreak of bovine tuberculosis in two herds in south west Scotland: Veterinary and human public health response," *Journal of Public Health Medicine*, vol. 18, no. 1, pp. 54–8, 1996.

[72] N. D. Barlow, J. M. Kean, G. Hickling, P. G. Livingstone, and A. B. Robson, "A simulation model for the spread of bovine tuberculosis within New Zealand cattle herds," *Preventive Veterinary Medicine*, vol. 32, no. 1-2, pp. 57–75, 1997.

[73] M. Gilbert, A. Mitchell, D. Bourn, J. Mawdsley, R. Clifton-Hadley, and W. Wint, "Cattle movements and bovine tuberculosis in Great Britain," *Nature*, vol. 435, no. 7041, pp. 491–6, 2005.

[74] A. M. Ramírez-Villaescusa, G. F. Medley, S. Mason, and L. E. Green, "Risk factors for herd breakdown with bovine tuberculosis in 148 cattle herds in the south west of England," *Preventive Veterinary Medicine*, vol. 95, no. 3-4, pp. 224–30, 2010.

[75] D. M. Green, I. Z. Kiss, A. P. Mitchell, and R. R. Kao, "Estimates for local and movement-based transmission of bovine tuberculosis in British cattle," *Proceedings of the Royal Society B: Biological Sciences*, vol. 275, no. 1638, pp. 1001–5, 2008.

[76] R. Jackson, G. W. de Lisle, and R. S. Morris, "A study of the environmental survival of *Mycobacterium bovis* on a farm in New Zealand," *New Zealand Veterinary Journal*, vol. 43, no. 7, pp. 346–52, 1995.

[77] B. J. Duffield and D. A. Young, "Survival of *Mycobacterium bovis* in defined environmental conditions," *Veterinary Microbiology*, vol. 10, pp. 193–197, 1984.

[78] J. S. Young, E. Gormley, and E. M. H. Wellington, "Molecular detection of *Mycobacterium bovis* and *Mycobacterium bovis* BCG (Pasteur) in soil," *Applied and Environmental Microbiology*, vol. 71, no. 4, pp. 1946–1952, 2005.

[79] A. E. Fine, C. Bolin, J. C. Gardiner, and J. B. Kaneene, "A study of the persistence of *Mycobacterium bovis* in the environment under natural weather conditions in Michigan, USA," *Veterinary Medicine International*, vol. 2011, p. 765430, 2011.

[80] D. M. Wolfe, O. Berke, D. F. Kelton, P. W. White, S. J. More, J. O'Keeffe, and S. W. Martin, "From explanation to prediction: A model for recurrent bovine tuberculosis in Irish cattle herds," *Preventive Veterinary Medicine*, vol. 94, no. 3-4, pp. 170–177, 2010.

[81] K. Karolemeas, T. J. McKinley, R. S. Clifton-Hadley, A. V. Goodchild, A. Mitchell, W. T. Johnston, A. J. K. Conlan, C. A. Donnelly, and J. L. N. Wood, "Recurrence of bovine tuberculosis breakdowns in Great Britain: Risk factors and prediction," *Preventive Veterinary Medicine*, vol. 102, no. 1, pp. 22–29, 2011.

[82] M. J. Gallagher, I. M. Higgins, T. A. Clegg, D. H. Williams, and S. J. More, "Comparison of bovine tuberculosis recurrence in Irish herds between 1998 and 2008," *Preventive Veterinary Medicine*, vol. 111, no. 3-4, pp. 237–44, 2013.

[83] K. L. Dawson, M. A. Stevenson, J. A. Sinclair, and M. A. Bosson, "Recurrent bovine tuberculosis in New Zealand cattle and deer herds, 2006–2010," *Epidemiology and Infection*, vol. 142, no. 10, pp. 2065–2074, 2014.

[84] M. Good, T. A. Clegg, A. Duignan, and S. J. More, "Impact of the national full herd depopulation policy on the recurrence of bovine tuberculosis in Irish herds, 2003 to 2005," *Veterinary Record*, vol. 169, p. 581, 2011.

[85] I. Schiller, B. Oesch, H. M. Vordermeier, M. V. Palmer, B. N. Harris, K. A. Orloski, B. M. Buddle, T. C. Thacker, K. P. Lyashchenko, and W. R. Waters, "Bovine tuberculosis: A review of current and emerging diagnostic techniques in view of their relevance for disease control and eradication," *Transboundary and Emerging Diseases*, vol. 57, no. 4, pp. 205–220, 2010.

[86] C. C. Okafor, D. L. Grooms, C. S. Bruning-Fann, J. J. Averill, and J. B. Kaneene, "Descriptive epidemiology of bovine tuberculosis in michigan (1975-2010): Lessons learned," *Veterinary Medicine International*, 2011.

[87] I. Schiller, W. RayWaters, H. M. Vordermeier, T. Jemmi, M. Welsh, N. Keck, A. Whelan, E. Gormley, M. L. Boschiroli, J. L. Moyen, C. Vela, M. Cagiola, B. M. Buddle, M. Palmer, T. Thacker, and B. Oesch, "Bovine tuberculosis in Europe from the perspective of an officially tuberculosis free country: Trade, surveillance and diagnostics," *Veterinary Microbiology*, vol. 151, no. 1-2, pp. 153–159, 2011.

[88] E. Brooks-Pollock, G. O. Roberts, and M. J. Keeling, "A dynamic model of bovine tuberculosis spread and control in Great Britain," *Nature*, vol. 511, no. 7508, pp. 228–231, 2014.

[89] D. Collins and G. W. de Lisle, "DNA restriction endonuclease analysis of *Mycobacterium bovis* and other members of the tuberculosis complex," *Journal of Clinical Microbiology*, vol. 21, no. 4, pp. 562–4, 1985.

[90] D. M. Collins, G. W. De Lisle, and D. M. Gabric, "Geographic distribution of restriction types of *Mycobacterium bovis* isolates from brush-tailed possums (*Trichosurus vulpecula*) in New Zealand," *The Journal of Hygiene*, vol. 96, no. 3, pp. 431–8, 1986.

[91] D. van Soolingen, P. E. W. de Haas, and K. Kremer, "Restriction fragment length polymorphism typing of Mycobacteria," *Methods in Molecular Medicine*, vol. 54, no. 13, pp. 165–203, 2001.

[92] M. Price-Carter, S. Rooker, and D. M. Collins, "Comparison of 45 variable number tandem repeat (VNTR) and two direct repeat (DR) assays to restriction endonuclease analysis for typing isolates of *Mycobacterium bovis*," *Veterinary Microbiology*, vol. 150, no. 1-2, pp. 107–114, 2011.

[93] D. M. Collins, D. M. Gabric, and G. W. de Lisle, "Typing of *Mycobacterium bovis* isolates from cattle and other animals in the same locality," *New Zealand Veterinary Journal*, vol. 36, no. 1, pp. 45–46, 1988.

[94] D. M. Collins, "DNA typing of *Mycobacterium bovis* strains from the Castlepoint area of the Wairarapa," *New Zealand Veterinary Journal*, vol. 47, no. 6, pp. 207–209, 1999.

[95] R. A. Skuce, D. Brittain, M. S. Hughes, L. A. Beck, and S. D. Neill, "Genomic fingerprinting of *Mycobacterium bovis* from cattle by restriction fragment length polymorphism analysis," *Journal of Clinincal Microbiology*, vol. 32, no. 7814471, pp. 2387–2392, 1994.

[96] J. Kamerbeek, L. Schouls, A. Kolk, M. van Agterveld, D. van Soolingen, S. Kuijper, A. Bunschoten, H. Molhuizen, R. Shaw, M. Goyal, and J. van Embden, "Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology," *Journal of Clinical Microbiology*, vol. 35, no. 4, pp. 907–14, 1997.

[97] A. Aranaz, E. Liébana, A. Mateos, L. Domínguez, and D. Cousins, "Restriction fragment length polymorphism and spacer oligonucleotide typing: A comparative analysis of fingerprinting strategies for *Mycobacterium bovis*," *Veterinary Microbiology*, vol. 61, no. 4, pp. 311–324, 1998.

[98] D. Cousins, S. Williams, E. Liébana, A. Aranaz, A. Bunschoten, J. Van Embden, and T. Ellis, "Evaluation of four DNA typing techniques in epidemiological investigations of bovine tuberculosis," *Journal of Clinical Microbiology*, vol. 36, no. 1, pp. 168–78, 1998.

[99] S. Roring, D. Brittain, A. E. Bunschoten, M. S. Hughes, R. A. Skuce, J. D. A. Van Embden, and S. D. Neill, "Spacer oligotyping of *Mycobacterium bovis* isolates compared to typing by restriction fragment length polymorphism using PGRS, DR and IS6110 probes," *Veterinary Microbiology*, vol. 61, no. 1-2, pp. 111–120, 1998.

[100] N. H. Smith, J. Dale, J. Inwald, S. Palmer, S. V. Gordon, R. G. Hewinson, and J. Smith, "The population structure of *Mycobacterium bovis* in Great Britain: Clonal expansion," *Proceedings of the National Academy of Sciences: PNAS*, vol. 100, no. 25, pp. 15271–15275, 2003.

[101] A. Aranaz, L. D. Juan, N. Montero, M. Galka, C. Delso, J. Álvarez, B. Romero, J. Bezos, A. I. Vela, V. Briones, A. Mateos, L. Domínguez, C. Sa, and A. Julio, "Bovine tuberculosis (*Mycobacterium bovis*) in wildlife in Spain," *Journal of Clinical Microbiology*, vol. 42, no. 6, pp. 2602–2608, 2004.

[102] T. Garnier, K. Eiglmeier, J.-C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P. R. Wheeler, J. Parkhill, B. G. Barrell, S. T. Cole, S. V. Gordon, and R. G. Hewinson, "The complete genome sequence of *Mycobacterium bovis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 13, pp. 7877–82, 2003.

[103] S. Roring, A. Scott, D. Brittain, I. Walker, G. Hewinson, S. Neill, and R. Skuce, "Development of variable-number tandem repeat typing of *Mycobacterium bovis*: Com-

parison of results with those obtained by using existing exact tandem repeats and spoligotyping," *Journal of Clinical Microbiology*, vol. 40, no. 6, pp. 2126–33, 2002.

[104] R. A. Skuce, S. W. McDowell, T. R. Mallon, B. Luke, E. L. Breadon, P. L. Lagan, C. M. McCormick, S. H. McBride, and J. M. Pollock, "Discrimination of isolates of *Mycobacterium bovis* in Northern Ireland on the basis of variable numbers of tandem repeats (VNTRs)," *The Veterinary Record*, vol. 157, no. 17, pp. 501–4, 2005.

[105] S. Roring, A. N. Scott, R. Glyn Hewinson, S. D. Neill, and R. A. Skuce, "Evaluation of variable number tandem repeat (VNTR) loci in molecular typing of *Mycobacterium bovis* isolates from Ireland," *Veterinary Microbiology*, vol. 101, no. 1, pp. 65–73, 2004.

[106] M. Hilty, C. Diguimbaye, E. Schelling, F. Baggi, M. Tanner, and J. Zinsstag, "Evaluation of the discriminatory power of variable number tandem repeat (VNTR) typing of *Mycobacterium bovis* strains," *Veterinary Microbiology*, vol. 109, no. 3-4, pp. 217–22, 2005.

[107] X. Didelot, R. Bowden, D. J. Wilson, T. E. Peto, and D. W. Crook, "Transforming clinical microbiology with bacterial genome sequencing," *Nature Reviews Genetics*, vol. 13, no. 9, pp. 601–612, 2012.

[108] M. L. Metzker, "Sequencing technologies - the next generation," *Nature Reviews. Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[109] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu, "A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.," *BMC Genomics*, vol. 13, no. 1, p. 341, 2012.

[110] M. Ruffalo, M. Koyutürk, S. Ray, and T. LaFramboise, "Accurate estimation of short read mapping quality for next-generation genome sequencing," *Bioinformatics (Oxford, England)*, vol. 28, no. 18, pp. 349–355, 2012.

[111] R. R. Kao, D. T. Haydon, S. J. Lycett, and P. R. Murcia, "Supersize me: How whole-genome sequencing and big data are transforming epidemiology," *Trends in Microbiology*, vol. 22, no. 5, pp. 282–91, 2014.

[112] N. J. Croucher and X. Didelot, "The application of genomics to tracing bacterial pathogen transmission," *Current Opinion in Microbiology*, vol. 23, pp. 62–67, 2015.

[113] R. R. Kao, M. Price-Carter, and S. Robbe-Austerman, "Use of genomics to track bovine tuberculosis transmission," *Revue Scientifique et Technique (International Office of Epizootics)*, vol. 35, no. 1, pp. 241–58, 2016.

[114] C. B. Ford, P. L. Lin, M. R. Chase, R. R. Shah, O. Iartchouk, J. Galagan, N. Mohaideen, T. R. Ioerger, J. C. Sacchettini, M. Lipsitch, J. L. Flynn, and S. M. Fortune, "Use of

whole genome sequencing to estimate the mutation rate of Mycobacterium tuberclosis during latent infection," *Nature Genetics*, vol. 43, no. 5, pp. 482–6, 2011.

[115] J. L. Gardy, J. Johnston, S. J. H. Sui, V. J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, R. Varhol, I. Birol, M. Lem, M. K. Sharma, K. Elwood, S. J. M. Jones, F. S. L. Brinkman, R. C. Brunham, and P. Tang, "Whole genome sequencing and social-network analysis of a tuberculosis outbreak," *New England Journal of Medicine*, vol. 364, no. 8, pp. 730–739, 2011.

[116] T. M. Walker, C. L. C. Ip, R. H. Harrell, J. T. Evans, G. Kapatai, M. J. Dedicoat, D. W. Eyre, D. Wilson, P. M. Hawkey, D. W. Crook, J. Parkhill, D. Harris, S. A. Walker, R. Bowden, P. Monk, G. E. Smith, and T. E. A. Peto, "Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: A retrospective observational study," *The Lancet Infectious Diseases*, 2012.

[117] J. M. Bryant, A. C. Schürch, H. van Deutekom, S. R. Harris, J. L. de Beer, V. de Jager, K. Kremer, S. A. F. T. van Hijum, R. J. Siezen, M. Borgdorff, S. D. Bentley, J. Parkhill, and D. van Soolingen, "Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data," *BMC Infectious Diseases*, vol. 13, no. 110, pp. 1–12, 2013.

[118] A. Roetzer, R. Diel, T. A. Kohl, C. Rückert, U. Nübel, J. Blom, T. Wirth, S. Jaenicke, S. Schuback, S. Rüsch-Gerdes, P. Supply, J. Kalinowski, and S. Niemann, "Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: A longitudinal molecular epidemiological study," *PLoS Medicine*, vol. 10, no. 2, pp. 1–12, 2013.

[119] L. Perez-Lago, I. Comas, Y. Navarro, F. Gonzalez-Candelas, M. Herranz, E. Bouza, and D. Garcia de Viedma, "Whole Genome Sequencing of intrapatient microevolution in *Mycobacterium tuberculosis*: Potential impact on the inference of tuberculosis transmission," *Journal of Infectious Diseases*, 2013.

[120] R. Colangeli, V. L. Arcus, R. T. Cursons, A. Ruthe, N. Karalus, K. Coley, S. D. Manning, S. Kim, E. Marchiano, and D. Alland, "Whole genome sequencing of *Mycobacterium tuberculosis* reveals slow growth and low mutation rates during latent infections in humans," *PLoS ONE*, vol. 9, no. 3, pp. 1–9, 2014.

[121] R. Biek, O. G. Pybus, J. O. Lloyd-Smith, and X. Didelot, "Measurably evolving pathogens in the genomic era," *Trends in Ecology & Evolution*, vol. 30, no. 6, pp. 306–313, 2015.

[122] G. Nugent, C. Gortazar, and G. Knowles, "The epidemiology of *Mycobacterium bovis* in wild deer and feral pigs and their roles in the establishment and spread of bovine tuberculosis in New Zealand wildlife," *New Zealand Veterinary Journal*, vol. 63 Suppl 1, no. sup1, pp. 54–67, 2015.

[123] D. T. Haydon, S. Cleaveland, L. H. Taylor, and M. K. Laurenson, "Identifying reservoirs of infection: A conceptual and practical challenge," *Emerging Infectious Diseases*, vol. 8, no. 12, pp. 1468–73, 2002.

[124] B. Warburton, P. Cowan, and J. Shepherd, "How many possums are now in New Zealand following control and how many would there be without it?," tech. rep., TBfree, New Zealand, 2009.

[125] M. Girot, J. M. Ferro, P. Canhão, J. Stam, M.-G. Bousser, F. Barinagarrementeria, and D. Leys, "Predictors of outcome in patients with cerebral venous thrombosis and intracerebral hemorrhage," *Stroke; A Journal of Cerebral Circulation*, vol. 38, no. 2, pp. 337–42, 2007.

[126] G. Nugent, "Maintenance, spillover and spillback transmission of bovine tuberculosis in multi-host wildlife complexes: A New Zealand case study," *Veterinary Microbiology*, vol. 151, no. 1-2, pp. 34–42, 2011.

[127] B. M. Buddle, F. E. Aldwell, A. Pfeffer, and G. W. de Lisle, "Experimental *Mycobacterium bovis* infection in the brushtail possum (*Trichosurus vulpecula*): Pathology, haematology and lymphocyte stimulation responses," *Veterinary Microbiology*, vol. 38, no. 3, pp. 241–254, 1994.

[128] D. Ramsey and P. Cowan, "Mortality rate and movements of brushtail possums with clinical tuberculosis (*Mycobacterium bovis*) infection," *New Zealand Veterinary Journal*, vol. 51, no. 4, pp. 179–85, 2003.

[129] L. A. L. Corner, D. U. Pfeiffer, G. W. de Lisle, R. S. Morris, and B. M. Buddle, "Natural transmission of *Mycobacterium bovis* infection in captive brushtail possums (*Trichosurus vulpecula*)," *New Zealand Veterinary Journal*, vol. 50, no. 4, pp. 154–162, 2002.

[130] J. Whitford, C. Rouco, D. Tompkins, and G. Nugent, "First direct estimate of the detection probability of bovine tuberculosis in possums by possum transmission," *European Journal of Wildlife Research*, vol. 60, no. 5, pp. 827–830, 2014.

[131] B. M. Paterson and R. S. Morris, "Interactions between beef cattle and simulated tuberculous possums on pasture," *New Zealand Veterinary Journal*, vol. 43, no. 7, pp. 289–93, 1995.

[132] I. J. Yockney, G. Nugent, M. C. Latham, M. Perry, M. L. Cross, and A. E. Byrom, "Comparison of ranging behaviour in a multi-species complex of free-ranging hosts of bovine tuberculosis in relation to their use as disease sentinels," *Epidemiology and Infection*, vol. 141, no. 7, pp. 1407–16, 2013.

[133] G. Nugent, *The role of wild deer in the epidmeiology and management of bovine tuberculosis in New Zealand*. PhD thesis, Lincoln University, 2005.

[134] J. D. Coleman and M. M. Cooke, "*Mycobacterium bovis* infection in wildlife in New Zealand," *Tuberculosis*, vol. 81, no. 3, pp. 191–202, 2001.

[135] D. P. Anderson, D. S. L. Ramsey, G. W. de Lisle, M. Bosson, M. L. Cross, and G. Nugent, "Development of integrated surveillance systems for the management of tuberculosis in New Zealand wildlife," *New Zealand Veterinary Journal*, vol. 63, no. sup1, pp. 89–97, 2015.

[136] P. Caley and J. Hone, "Assessing the host disease status of wildlife and the implications for disease control: *Mycobacterium bovis* infection in feral ferrets," *Journal of Applied Ecology*, vol. 42, no. 4, pp. 708–719, 2005.

[137] R. J. Delahay, G. C. Smith, A. M. Barlow, N. Walker, A. Harris, R. S. Clifton-Hadley, and C. L. Cheeseman, "Bovine tuberculosis infection in wild mammals in the South-West region of England: A survey of prevalence and a semi-quantitative assessment of the relative risks to cattle," *Veterinary Journal*, vol. 173, no. 2, pp. 287–301, 2007.

[138] F. Vial, W. T. Johnston, and C. A. Donnelly, "Local cattle and badger populations affect the risk of confirmed tuberculosis in British cattle herds," *PloS ONE*, vol. 6, no. 3, pp. 1–9, 2011.

[139] A. W. Byrne, D. Paddy S, J. O'Keeffe, and J. Davenport, "The ecology of the European badger (*Meles meles*) in Ireland: A review," *Biology & Environment: Proceedings of the Royal Irish Academy*, vol. 112, no. 1, pp. 105–132, 2012.

[140] C. H. Benton, R. J. Delahay, A. Robertson, A. Mcdonald, A. J. Wilson, T. A. Burke, D. Hodgson, C. H. Benton, and D. Hodgson, "Blood thicker than water: Kinship, disease prevalence and group size drive divergent patterns of infection risk in a social mammal," *Proceedings of the Royal Society B: Biological Sciences*, vol. 283, pp. 1–8, 2016.

[141] R. S. Clifton-Hadley, J. W. Wilesmith, and F. Stuart, "*Mycobacterium bovis* in the European badger (*Meles meles*): Epidemiological findings in tuberculous badgers from a naturally infected population," *Epidemiology and Infection*, vol. 111, no. 1, pp. 9–19, 1993.

[142] J. Gallagher and R. S. Clifton-Hadley, "Tuberculosis in badgers: A review of the disease and its significance for other animals," *Research in Veterinary Science*, vol. 69, no. 3, pp. 203–17, 2000.

[143] L. A. L. Corner, E. Costello, S. Lesellier, D. O'Meara, and E. Gormley, "Experimental tuberculosis in the European badger (*Meles meles*) after endobronchial inoculation with *Mycobacterium bovis*: II. Progression of infection," *Research in Veterinary Science*, vol. 85, no. 3, pp. 481–490, 2008.

[144] L. A. L. Corner, D. Murphy, and E. Gormley, "*Mycobacterium bovis* infection in the Eurasian badger (*Meles meles*): The disease, pathogenesis, epidemiology and control," *Journal of Comparative Pathology*, vol. 144, no. 1, pp. 1–24, 2011.

[145] J. Gallagher, R. Monies, M. Gavier-Widen, and B. Rule, "Role of infected, non-diseased badgers in the pathogenesis of tuberculosis in the badger," *The Veterinary Record*, vol. 142, no. 26, pp. 710–4, 1998.

[146] D. Wilkinson, G. C. Smith, R. J. Delahay, L. M. Rogers, C. L. Cheeseman, and R. S. Clifton-Hadley, "The effects of bovine tuberculosis (*Mycobacterium bovis*) on mortality in a badger (*Meles meles*) population in England," *Journal of Zoology*, vol. 250, no. 3, pp. 389–395, 2000.

[147] A. I. Ward, B. A. Tolhurst, N. J. Walker, T. J. Roper, and R. J. Delahay, "Survey of badger access to farm buildings and facilities in relation to contact with cattle," *The Veterinary Record*, vol. 163, no. 4, pp. 107–11, 2008.

[148] B. T. Garnett, R. J. Delahay, and T. J. Roper, "Use of cattle farm resources by badgers (*Meles meles*) and risk of bovine tuberculosis (*Mycobacterium bovis*) transmission to cattle," *Proceedings of the Royal Society B: Biological Sciences*, vol. 269, no. 1499, pp. 1487–91, 2002.

[149] M. Böhm, M. R. Hutchings, and P. C. L. White, "Contact networks in a wildlife-livestock host community: Identifying high-risk individuals in the transmission of bovine TB among badgers and cattle," *PLoS ONE*, vol. 4, no. 4, 2009.

[150] J. A. Drewe, N. Weber, S. P. Carter, S. Bearhop, X. A. Harrison, S. R. X. Dall, R. A. McDonald, and R. J. Delahay, "Performance of proximity loggers in recording intra- and inter-species interactions: A laboratory and field-based validation study," *PLoS ONE*, vol. 7, no. 6, pp. 1–9, 2012.

[151] J. A. Drewe, H. M. O'Connor, N. Weber, R. A. McDonald, and R. J. Delahay, "Patterns of direct and indirect contact between cattle and badgers naturally infected with tuberculosis," *Epidemiology and Infection*, vol. 141, no. 07, pp. 1467–1475, 2013.

[152] E. M. Mullen, T. MacWhite, P. K. Maher, D. J. Kelly, N. M. Marples, and M. Good, "Foraging Eurasian badgers *Meles meles* and the presence of cattle in pastures. Do badgers avoid cattle?," *Applied Animal Behaviour Science*, vol. 144, no. 3-4, pp. 130–137, 2013.

[153] C. L. Cheeseman, J. W. Wilesmith, F. A. Stuart, and P. J. Mallinson, "Dynamics of tuberculosis in a naturally infected badger population," *Mammal Review*, vol. 18, no. 1, pp. 61–72, 1988.

[154] T. W. A. Little, C. Swan, H. V. Thompson, and J. W. Wilesmith, "Bovine tuberculosis in domestic and wild mammals in an area of Dorset. II. The badger population, its ecology and tuberculosis status," *Journal of Hygiene*, vol. 89, pp. 211–224, 1982.

[155] R. M. Davidson, "Role of the opossum in spreading tuberculosis," *New Zealand Journal of Agriculture*, 1976.

[156] C. Eason, A. Miller, S. Ogilvie, and A. Fairweather, "An updated review of the toxicology and ecotoxicology of sodium fluoroacetate (1080) in relation to its use as a pest control tool in New Zealand," *New Zealand Journal of Ecology*, vol. 35, no. 1, pp. 1–20, 2011.

[157] R. S. Clifton-Hadley, J. W. Wilesmith, M. S. Richards, P. Upton, and S. Johnston, "The occurrence of *Mycobacterium bovis* infection in cattle in and around an area subject to extensive badger (*Meles meles*) control," *Epidemiology and Infection*, vol. 114, no. 1, pp. 179–93, 1995.

[158] D. O. Máirtín, D. H. Williams, J. M. Griffin, L. A. Dolan, and J. A. Eves, "The effect of a badger removal programme on the incidence of tuberculosis in an Irish cattle population," *Preventive Veterinary Medicine*, vol. 34, no. 1, pp. 47–56, 1998.

[159] J. M. Griffin, D. H. Williams, G. E. Kelly, T. A. Clegg, I. O'Boyle, J. D. Collins, and S. J. More, "The impact of badger removal on the control of tuberculosis in cattle herds in Ireland," *Preventive Veterinary Medicine*, vol. 67, no. 4, pp. 237–66, 2005.

[160] F. J. Olea-Popelka, P. Fitzgerald, P. White, G. McGrath, J. D. Collins, J. O'Keeffe, D. F. Kelton, O. Berke, S. More, and S. W. Martin, "Targeted badger removal and the subsequent risk of bovine tuberculosis in cattle herds in county Laois, Ireland," *Preventive Veterinary Medicine*, vol. 88, no. 3, pp. 178–84, 2009.

[161] F. J. Bourne, "Bovine TB: The scientific evidence," *Department for Environment, Food and Rural Affairs (DEFRA) Publications*, no. June, pp. 3–289, 2007.

[162] C. A. Donnelly, R. Woodroffe, D. R. Cox, F. J. Bourne, C. L. Cheeseman, R. S. Clifton-Hadley, G. Wei, G. Gettinby, P. Gilks, H. Jenkins, W. T. Johnston, A. M. Le Fevre, J. P. McInerney, and W. I. Morrison, "Positive and negative effects of widespread badger culling on tuberculosis in cattle," *Nature*, vol. 439, no. 7078, pp. 843–846, 2006.

[163] R. Woodroffe, C. A. Donnelly, D. R. Cox, F. J. Bourne, C. L. Cheeseman, R. J. Delahay, G. Gettinby, M. J. P, and W. I. Morrison, "Effects of culling on badger *Meles meles* spatial organization: Implications for the control of bovine tuberculosis," *Journal of Applied Ecology*, vol. 43, no. 1, pp. 1–10, 2006.

[164] G. J. Wilson, S. P. Carter, and R. J. Delahay, "Advances and prospects for management of TB transmission between badgers and cattle," *Veterinary Microbiology*, vol. 151, no. 1-2, pp. 43–50, 2011.

[165] C. M. O'Connor, D. T. Haydon, and R. R. Kao, "An ecological and comparative perspective on the control of bovine tuberculosis in Great Britain and the Republic of Ireland," *Preventive Veterinary Medicine*, vol. 104, no. 3-4, pp. 185–97, 2012.

[166] S. Lesellier, L. Corner, E. Costello, K. Lyashchenko, R. Greenwald, J. Esfandiari, M. Singh, R. G. Hewinson, M. Chambers, and E. Gormley, "Immunological responses and protective immunity in BCG vaccinated badgers following endobronchial infection with *Mycobacterium bovis*," *Vaccine*, vol. 27, no. 3, pp. 402–9, 2009.

[167] M. A. Chambers, F. Rogers, R. J. Delahay, S. Lesellier, R. Ashford, D. Dalley, S. Gowtage, D. Dave, S. Palmer, J. Brewer, T. Crawshaw, R. Clifton-Hadley, S. Carter, C. Cheeseman, C. Hanks, A. Murray, K. Palphramand, S. Pietravalle, G. C. Smith, A. Tomlinson, N. J. Walker, G. J. Wilson, L. L. Corner, S. P. Rushton, M. D. F. Shirley, G. Gettinby, R. A. McDonald, and R. G. Hewinson, "Bacillus Calmette-Guerin vaccination reduces the severity and progression of tuberculosis in badgers," *Proceedings of the Royal Society B: Biological Sciences*, vol. 278, no. 1713, pp. 1913–1920, 2011.

[168] L. A. L. Corner, E. Costello, D. O'Meara, S. Lesellier, F. E. Aldwell, M. Singh, R. G. Hewinson, M. A. Chambers, and E. Gormley, "Oral vaccination of badgers (*Meles meles*) with BCG and protective immunity against endobronchial challenge with *Mycobacterium bovis*," *Vaccine*, vol. 28, no. 38, pp. 6265–6272, 2010.

[169] S. P. Carter, M. A. Chambers, S. P. Rushton, M. F. Shirley, P. Schuchert, S. Pietravalle, A. Murray, F. Rogers, G. Gettinby, G. C. Smith, R. J. Delahay, R. G. Hewinson, and R. A. McDonald, "BCG vaccination reduces risk of tuberculosis infection in vaccinated badgers and unvaccinated badger cubs," *PLoS ONE*, vol. 7, no. 12, pp. 1–8, 2012.

[170] C. Benton and G. Wilson, "Badger Vaccine Deployment Project Final Lessons Learned Report," Tech. Rep. March, Animal and Plant Health Agency, 2015.

[171] K. P. Lyashchenko, R. Greenwald, J. Esfandiari, D. J. O'Brien, S. M. Schmitt, M. V. Palmer, and W. R. Waters, "Rapid detection of serum antibody by dual-path platform VetTB assay in white-tailed deer infected with *Mycobacterium bovis*," *Clinical and Vaccine Immunology: CVI*, vol. 20, no. 6, pp. 907–911, 2013.

[172] DAERA-NI, "The test and vaccinate or remove (TVR) wildlife intervention research project - year 2 report," tech. rep., Department of Agriculture, Environment and Rural Affairs, 2015.

[173] J. Hone, "Diminishing returns in bovine tuberculosis control," *Epidemiology and Infection*, vol. 141, no. 7, pp. 1382–9, 2013.

[174] P. R. Torgerson and D. Torgerson, "Public health and bovine tuberculosis: What's all the fuss about?," *Trends in Microbiology*, vol. 18, no. 2, pp. 67–72, 2010.

[175] R. Skuce, T. R. Mallon, C. M. McCormick, S. H. McBride, G. Clarke, A. Thompson, C. Couzens, W. Gordon, and S. W. J. McDowell, "*Mycobacterium bovis* genotypes in Northern Ireland: Herd-level surveillance (2003 to 2008)," *The Veterinary Record*, vol. 167, no. 18, pp. 684–9, 2010.

[176] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics (Oxford, England)*, vol. 25, no. 16, pp. 2078–9, 2009.

[177] O. Bishop, *Bioinformatics and data analysis in microbiology*. Norfolk, UK: Caister Academic Press, 2014.

[178] A. Van Belkum, J. Kluytmans, W. Van Leeuwen, R. Bax, W. Quint, A. Fluit, A. Vandenbroucke-grauls, A. Van Den Brule, H. Koeleman, W. Melchers, J. Meis, A. Elaichouni, M. Vaneechoutte, F. Moonens, N. Maes, M. Struelens, F. Tenover, and H. Verbrugh, "Multicenter evaluation of arbitrarily primed PCR for typing of *Staphylococcus aureus* strains. Multicenter evaluation ofaArbitrarily primed PCR for typing of *Staphylococcus aureus* strains," *Journal of Clinical Microbiology*, vol. 33, no. 6, pp. 1537–1547, 1995.

[179] M. J. Struelens, "Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems," *Clinical Microbiology and Infection*, pp. 1–10, 1996.

[180] N. K. Fry, J. M. Bangsborg, S. Bernander, J. Etienne, B. Forsblom, V. Gaia, P. Hasenberger, D. Lindsay, A. Papoutsi, C. Pelaz, M. Struelens, S. Uldum, P. Visca, and T. G. Harrison, "Assessment of intercentre reproducibility and epidemiological concordance of *Legionella pneumophila* serogroup 1 genotyping by amplified fragment length polymorphism analysis," *European Journal of Clinical Microbiology & Infectious Diseases*, vol. 19, no. 10, pp. 773–780, 2000.

[181] V. Gaia, N. K. Fry, B. Afshar, C. P. Luck, H. Meugnier, J. Etienne, R. Peduzzi, and T. G. Harrison, "Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila* consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*," *Journal of Clinical Microbiology*, vol. 43, no. 5, pp. 2047–2052, 2005.

[182] C. Pourcel, P. Visca, B. Afshar, S. D'Arezzo, G. Vergnaud, and N. K. Fry, "Identification of Variable-Number Tandem-Repeat (VNTR) sequences in *Legionella pneumophila* and development of an optimized multiple-locus VNTR Analysis typing scheme," *Journal of Clinical Microbiology*, vol. 45, no. 4, pp. 1190–1199, 2007.

[183] T. Parish and N. G. Stoker, *Mycobacterium tuberculosis Protocols*. Humana Press, 2001.

[184] G. W. de Lisle, R. G. Bengis, S. M. Schmitt, and D. J. O'Brien, "Tuberculosis in free-ranging wildlife: Detection, diagnosis and management," *Revue Scientifique et Technique (International Office of Epizootics)*, vol. 21, no. 2, pp. 317–34, 2002.

[185] R. J. Delahay, N. Walker, G. S. Smith, D. Wilkinson, R. S. Clifton-Hadley, C. L. Cheeseman, A. J. Tomlinson, and M. A. Chambers, "Long-term temporal trends and estimated transmission rates for *Mycobacterium bovis* infection in an undisturbed high-density badger (*Meles meles*) population," *Epidemiology and Infection*, vol. 141, no. 07, pp. 1445–1456, 2013.

[186] R. J. Delahay, S. Langton, G. C. Smith, R. S. Clifton-Hadley, and C. L. Cheeseman, "The spatio-temporal distribution of *Mycobacterium bovis* (bovine tuberculosis) infection in a high-density badger population," *Journal of Animal Ecology*, vol. 69, no. 3, pp. 428–441, 2000.

[187] J. Goodger, A. Nolan, W. Russell, D. Dalley, C. Thorns, F. Stuart, P. Croston, and D. Newell, "Serodiagnosis of *Mycobacterium bovis* infection in badgers: development of an indirect ELISA using a 25 k1," *Veterinary Record*, vol. 135, no. 4, pp. 82–85, 1994.

[188] K. P. Lyashchenko, R. Greenwald, J. Esfandiari, M. A. Chambers, J. Vicente, C. Gortazar, N. Santos, M. Correia-Neves, B. M. Buddle, R. Jackson, D. J. O'Brien, S. Schmitt, M. V. Palmer, R. J. Delahay, and W. R. Waters, "Animal-side serologic assay for rapid detection of *Mycobacterium bovis* infection in multiple species of free-ranging wildlife," *Veterinary Microbiology*, vol. 132, no. 3-4, pp. 283–92, 2008.

[189] S. Andrews, "FastQC: A quality control tool for high throughput sequence data," 2010.

[190] F. Krueger, "Trim Galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files," 2015.

[191] R. Schmieder and R. Edwards, "Quality control and preprocessing of metagenomic datasets," *Bioinformatics (Oxford, England)*, vol. 27, no. 6, pp. 863–4, 2011.

[192] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics (Oxford, England)*, vol. 25, no. 14, pp. 1754–60, 2009.

[193] S. L. Sampson, "Mycobacterial PE/PPE proteins at the host-pathogen interface," *Clinical & Developmental Immunology*, 2011.

[194] R. C. Team, "R: A language and environment for statistical computing," 2016.

[195] "Sequence Alignment / Map format specification," tech. rep., The SAM/BAM Format Specification Working Group, 2015.

[196] B. F. Zhan, "Three fastest shortest path algorithms on real road networks: Data structures and procedures," *Journal of Geographic Information and Decision Analysis*, vol. 1, no. 1, pp. 70–82, 1997.

[197] A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News*, vol. 2, no. December, pp. 18–22, 2002.

[198] J. A. Benavides, P. C. Cross, G. Luikart, and S. Creel, "Limitations to estimating bacterial cross-species transmission using genetic and genomic markers: Inferences from simulation modeling," *Evolutionary Applications*, vol. 7, no. 7, pp. 774–787, 2014.

[199] A. L. Michel, B. Müller, and P. D. van Helden, "*Mycobacterium bovis* at the animal-human interface: A problem, or not?," *Veterinary Microbiology*, vol. 140, no. 3-4, pp. 371–81, 2010.

[200] C. Gortazar and P. Cowan, "Dealing with TB in wildlife," *Epidemiology and Infection*, vol. 141, no. 7, pp. 1339–41, 2013.

[201] R. S. Morris and D. U. Pfeiffer, "Directions and issues in bovine tuberculosis epidemiology and control in New Zealand," *New Zealand Veterinary Journal*, vol. 43, no. 7, pp. 256–65, 1995.

[202] Y. Navarro, B. Romero, M. F. Copano, E. Bouza, L. Domínguez, L. de Juan, and D. García-de Viedma, "Multiple sampling and discriminatory fingerprinting reveals clonally complex and compartmentalized infections by *M. bovis* in cattle," *Veterinary Microbiology*, vol. 175, no. 1, pp. 99–104, 2015.

[203] G. W. de Lisle, R. Pamela Kawakami, G. F. Yates, and D. M. Collins, "Isolation of *Mycobacterium bovis* and other mycobacterial species from ferrets and stoats," *Veterinary Microbiology*, vol. 132, no. 3-4, pp. 402–407, 2008.

[204] J. Felsenstein, "PHYLIP - Phylogeny inference package - v3.2," 1989.

[205] A. Rambaut, "Path-O-Gen: Temporal signal investigation tool v1.4," 2009.

[206] A. L. Vestal, "Procedures for the isolation and identification of Mycobacteria," *Dept. of Health, Education, and Welfare, Center for Disease Control, Atlanta*, 1975.

[207] A. J. Drummond and A. Rambaut, "BEAST: Bayesian Evolutionary Analysis by Sampling Trees," *BMC Evolutionary Biology*, vol. 7, no. 1, p. 214, 2007.

[208] C. Firth, A. Kitchen, B. Shapiro, M. A. Suchard, E. C. Holmes, and A. Rambaut, "Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses," *Molecular Biology and Evolution*, vol. 27, no. 9, pp. 2038–2051, 2010.

[209] P. Lemey, A. Rambaut, A. J. Drummond, and M. A. Suchard, "Bayesian phylogeography finds its roots," *PLoS Computational Biology*, vol. 5, no. 9, pp. 1–16, 2009.

[210] S. Y. W. Ho and B. Shapiro, "Skyline-plot methods for estimating demographic history from nucleotide sequences," *Molecular Ecology Resources*, vol. 11, no. 3, pp. 423–34, 2011.

[211] A. Rambaut and A. J. Drummond, "Tracer v1. 4," 2007.

[212] M. Hasegawa, H. Kishino, and T. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution*, vol. 22, no. 2, pp. 160–74, 1985.

[213] M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, and M. A. Suchard, "Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci," *Molecular Biology and Evolution*, vol. 30, no. 3, pp. 713–24, 2013.

[214] M. Pagel, A. Meade, and D. Barker, "Bayesian estimation of ancestral character states on phylogenies," *Systematic Biology*, vol. 53, no. 5, pp. 673–84, 2004.

[215] B. Binney, P. Biggs, P. Carter, B. Holland, and N. French, "Quantification of historical livestock importation into New Zealand 1860–1979," *New Zealand Veterinary Journal*, vol. 62, no. 6, pp. 309–314, 2014.

[216] M. M. Cooke, B. M. Buddle, F. E. Aldwell, D. N. McMurray, and M. R. Alley, "The pathogenesis of experimental endo-bronchial *Mycobacterium bovis* infection in brushtail possums (*Trichosurus vulpecula*)," *New Zealand Veterinary Journal*, vol. 47, no. 6, pp. 187–192, 1999.

[217] J. L. Flynn and J. Chan, "Tuberculosis: Latency and reactivation," *Infection and Immunity*, vol. 69, no. 7, pp. 4195–4201, 2001.

[218] O. Mestre, T. Luo, T. Dos Vultos, K. Kremer, A. Murray, A. Namouchi, C. Jackson, J. Rauzier, P. Bifani, R. Warren, V. Rasolofo, J. Mei, Q. Gao, and B. Gicquel, "Phylogeny of *Mycobacterium tuberculosis* Beijing strains constructed from polymorphisms in genes involved in DNA replication, recombination and repair," *PloS ONE*, vol. 6, no. 1, p. e16020, 2011.

[219] M. B. O'Neill, T. D. Mortimer, and C. S. Pepperell, "Diversity of *Mycobacterium tuberculosis* across evolutionary scales," *PLoS Pathogens*, vol. 11, no. 11, pp. 1–48, 2015.

[220] J. R. Lawes, K. A. Harris, A. Brouwer, J. M. Broughan, N. H. Smith, and P. A. Upton, "Bovine TB surveillance in Great Britain in 2014," *Veterinary Record*, vol. 178, no. 13, pp. 310–315, 2016.

[221] J. Hermoso de Mendoza, A. Parra, A. Tato, J. M. Alonso, J. M. Rey, J. Peña, A. García-Sánchez, J. Larrasa, J. Teixidó, G. Manzano, R. Cerrato, G. Pereira, P. Fernández-Llario, and M. Hermoso de Mendoza, "Bovine tuberculosis in wild boar (*Sus scrofa*),

red deer (*Cervus elaphus*) and cattle (*Bos taurus*) in a Mediterranean ecosystem (1992-2004)," *Preventive Veterinary Medicine*, vol. 74, no. 2-3, pp. 239–47, 2006.

[222] V. Naranjo, C. Gortazar, J. Vicente, and J. de la Fuente, "Evidence of the role of European wild boar as a reservoir of *Mycobacterium tuberculosis* complex," *Veterinary Microbiology*, vol. 127, pp. 1–9, 2008.

[223] P. Durr, R. S. Clifton-Hadley, and R. G. Hewinson, "Molecular epidemiology of bovine tuberculosis. II. Applications of genotyping," *Revue Scientifique et Technique (International Office of Epizootics)*, vol. 19, no. 3, pp. 689–701, 2000.

[224] R. Woodroffe, C. A. Donnelly, D. R. Cox, P. Gilks, H. E. Jenkins, W. T. Johnston, A. M. Le Fevre, F. J. Bourne, C. L. Cheeseman, R. S. Clifton-Hadley, G. Gettinby, R. G. Hewinson, J. P. McInerney, P. Mitchell, W. I. Morrison, and G. H. Watkins, "Bovine tuberculosis in cattle and badgers in localized culling areas," *Journal of Wildlife Diseases*, vol. 45, no. 1, pp. 128–43, 2009.

[225] A. Aranaz, E. Liébana, A. Mateos, L. Dominguez, D. Vidal, M. Domingo, O. Gonzolez, E. F. Rodriguez-Ferri, A. E. Bunschoten, J. D. Van Embden, and D. Cousins, "Spacer oligonucleotide typing of *Mycobacterium bovis* strains from cattle and other animals: A tool for studying epidemiology of tuberculosis," *Journal of Clinical Microbiology*, vol. 34, no. 11, pp. 2734–40, 1996.

[226] M. C. Vernon, "Demographics of cattle movements in the United Kingdom," *BMC Veterinary Research*, vol. 7, no. 1, p. 31, 2011.

[227] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[228] R. Woodroffe, C. A. Donnelly, C. Ham, S. Y. B. Jackson, K. Moyes, K. Chapman, N. G. Stratton, and S. J. Cartwright, "Badgers prefer cattle pasture but avoid cattle: Implications for bovine tuberculosis control," *Ecology Letters*, vol. 19, no. 10, pp. 1201–1208, 2016.

[229] K. Karolemeas, C. A. Donnelly, A. J. K. Conlan, A. P. Mitchell, R. S. Clifton-Hadley, P. Upton, J. L. N. Wood, and T. J. McKinley, "The effect of badger culling on breakdown prolongation and recurrence of bovine tuberculosis in cattle herds in Great Britain," *PloS ONE*, vol. 7, no. 12, p. e51342, 2012.

[230] I. Aznar, G. McGrath, D. Murphy, L. Corner, E. Gormley, K. Frankena, S. J. More, W. Martin, J. O'Keeffe, and M. C. De Jong, "Trial design to estimate the effect of vaccination on tuberculosis incidence in badgers," *Veterinary Microbiology*, vol. 151, no. 1-2, pp. 104–111, 2011.

[231] A. W. Byrne, J. L. Quinn, J. J. O'Keeffe, S. Green, D. Paddy Sleeman, S. Wayne Martin, and J. Davenport, "Large-scale movements in European badgers: Has the tail

of the movement kernel been underestimated?," *Journal of Animal Ecology*, vol. 83, no. 4, pp. 991–1001, 2014.

[232] I. Comas, M. Coscolla, T. Luo, S. Borrell, K. E. Holt, M. Kato-Maeda, J. Parkhill, B. Malla, S. Berg, G. Thwaites, D. Yeboah-Manu, G. Bothamley, J. Mei, L. Wei, S. Bentley, S. R. Harris, S. Niemann, R. Diel, A. Aseffa, Q. Gao, D. Young, and S. Gagneux, "Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans," *Nature Genetics*, vol. 45, no. 10, pp. 1176–1182, 2013.

[233] N. De Maio, C. H. Wu, K. M. O'Reilly, and D. Wilson, "New routes to phylogeography: A Bayesian structured coalescent approximation," *PLoS Genetics*, vol. 11, no. 8, pp. 1–22, 2015.

[234] N. Weber, S. Bearhop, S. R. X. Dall, R. J. Delahay, R. A. McDonald, and S. P. Carter, "Denning behaviour of the European badger (*Meles meles*) correlates with bovine tuberculosis infection status," *Behavioral Ecology and Sociobiology*, vol. 67, no. 3, pp. 471–479, 2013.

[235] N. Weber, S. P. Carter, S. R. X. Dall, R. J. Delahay, J. L. McDonald, S. Bearhop, and R. A. McDonald, "Badger social networks correlate with tuberculosis infection," *Current Biology*, vol. 23, no. 20, pp. R915–6, 2013.

[236] Vynnycky E and R. White, *An introduction to infectious disease modelling*. Oxford: Oxford University Press, 2010.

[237] T. H. Jukes and C. R. Cantor, *Evolution of protein molecules*. 1969.

[238] A. J. Drummond, S. Y. W. Ho, M. J. Phillips, and A. Rambaut, "Relaxed phylogenetics and dating with confidence," *PLoS Biology*, vol. 4, no. 5, pp. 699–710, 2006.