# The genetic and spatial epidemiology of bovine tuberculosis in the UK:

## from molecular typing to bacterial whole genome sequencing

Hannah Trewby


Thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy (PhD)

---

Boyd Orr Centre for Population and Ecosystem Health
Institute for Biodiversity, Animal Health and Comparative Medicine
College of Medical Veterinary and Life Sciences
University of Glasgow

---

September 2015

# Abstract

Bovine tuberculosis (bTB) is a disease of cattle and other animals caused by the bacterium *Mycobacterium bovis*. In the UK, control of the disease presents significant difficulties, with bTB currently one of the most important diseases affecting the livestock industry in this region. The involvement of infection in a wildlife reservoir, the Eurasian badger *Meles meles*, is a key challenge to the eradication of bTB in the UK and causes much public controversy. In such a situation it is essential to use all available tools to control the spread of the disease. For many years, molecular typing of the *M. bovis* population has been routinely employed to understand the epidemiology of bTB in Britain and Northern Ireland (NI), providing broad scale information on the *M. bovis* population. More recently, high-resolution bacterial whole genome sequencing (WGS) of *M. bovis* has become feasible, although its use had yet to be explored in depth for the epidemiology of bTB. In this thesis, I describe various approaches to the use of pathogen genetic and spatial information to explore the epidemiology of bTB in the UK. I start at the broad scale, analysing the molecular types of *M. bovis* across the whole of NI, and go on to evaluate the use of WGS for *M. bovis* in targeted subpopulations of cattle and badgers. Following a general overview (Chapter 1), I explore the processes underlying the pattern of relative abundances of *M. bovis* molecular types in NI, showing that simple neutral processes are not capable of generating the distribution observed, and using simulation models to demonstrate that historical increases in bTB prevalence and/or transmission heterogeneity may be responsible (Chapter 2). In Chapter 3, I examine the spatial structure of the NI *M. bovis* population, demonstrating

significant spatial clustering of *M. bovis* molecular types and correlations between the types present in cattle and badgers. I go on to develop a landscape genetics approach, which provides preliminary indications that transmission in badgers might be responsible for the spatial structure observed in *M. bovis* infections in cattle. In Chapter 4, I evaluate the use of high density bacterial WGS in *M. bovis* isolates belonging to a single molecular type. I use these data to demonstrate some of the potential of WGS, while also highlighting limitations inherent in using these approaches in such a slowly evolving pathogen. In Chapter 5, I take methods developed in the preceding chapters and apply them to the study of bTB in a well studied badger population, characterising the genetic diversity of *M. bovis* present in these badgers and their links to local cattle infections. As a whole, this body of work shows that there is much to be gained in our understanding of the epidemiology of bTB from genetic and genomic data, both from examining the large historical datasets that already exist on molecular types of *M. bovis* in the UK, as well as from the application of high resolution bacterial WGS in this system.

To my parents, and in memory of Kathleen Turnbull

# Acknowledgements

The work within this thesis was funded by a Doctoral Training Grant from the Biotechnology and Biosciences Research Council and a Case Studentship from Novartis Animal Health.

This thesis would not have been possible without some superb datasets provided by some superb collaborators. My thanks therefore go to Robin Skuce, Adrian Allen, El Breadon and the team at AFBI, Northern Ireland, David Wright at Queens University Belfast, and Clare Benton and Dez Delahay at APHA Woodchester Park. I'm very grateful to my supervisors who advised on all aspects of the thesis, and to my colleagues across the Graham Kerr and elsewhere who generously gave of their advice and friendship. Finally a huge thank you to my family and friends for their love and support over the years.

## Specific contributions to thesis chapters

**Chapter 2:** VNTR-type data and information on the breakdowns associated with each VNTR-type record were provided by the team at AFBI Northern Ireland, and by David Wright (Queens University Belfast). Dan Haydon (University of Glasgow) helped with the design of the analysis and advised on the simulation models.

**Chapter 3:** Epidemiological and demographic data were provided by AFBI and DARD, and were extracted by Stewart McBride (AFBI) and David Wright (Queens University Belfast). Jason Matthiopoulos (University of

Glasgow) helped with the design of the spatial analyses of VNTR-types. Badger genotype data were generated by Adrian Allen (AFBI), and landscape genetics analyses of badger genotypes were carried out by Jimena Guerrero Flores (University of Glasgow).

**Chapter 4:** Data were provided courtesy of AFBI and DARD, and David Wright (Queens University Belfast), and *M. bovis* culture and DNA extraction were also carried out by AFBI. Sequence library preparation and whole genome sequencing were conducted by Julie Galbraith and Glasgow Polyomics. Anthony O'Hare (University of Stirling) gave help with coding, and Richard Orton and Graham Hamilton (both University of Glasgow) gave advice on bioinformatics analysis. Samantha Lycett (University of Edinburgh) provided the code for the phylogeographic simulations as well as other phylogenetic advice, and Paul Johnson (University of Glasgow) advised on the statistics used.

**Chapter 5:** Badger data were provided by Clare Benton, Dez Delahay, and the staff at APHA Woodchester Park, and the data on British cattle and *M. bovis* spoligotypes present in cattle were provided by APHA and DEFRA. *M. bovis* isolates were cultured and DNA was extracted both at FERA York and AFBI. Library preparation and sequencing were again carried out by Julie Galbraith and Glasgow Polyomics.

I am also very grateful to Matt Denwood (University of Copenhagen), who developed the thesis template used here

# Declaration

I declare that this thesis and the research contained within it is my own work unless otherwise stated, and no part of it has been submitted as part of any other degree or qualification. The work was carried out between September 2011 and September 2015 at the University of Glasgow under the supervision of Dr Roman Biek and Professor Rowland Kao.

Hannah Trewby

March 2016

# Contents

# List of Figures

# List of Tables

# Abbreviations

**AFBI** Agri-Food and Biosciences Institute, Northern Ireland

**APHA** Animal and Plant Health Agency, UK

**BCG** Bacille Calmettte Guerin: passaged strain of *M. bovis* used for vaccination

**BEAST** Bayesian Evolutionary Analysis Sampling Trees: computer programme for Bayesian phylogenetic analysis, see Drummond *et al.* (2012)

**bTB** Bovine Tuberculosis

**CMI** Cell-Mediated Immunity

**CPHH** Country Parish Herd Holding number: reference used in GB to identify individual farm holdings

**DARD** Department for Agriculture and Rural Development, Northern Ireland

**Defra** Department for Environment, Food and Rural Affairs, UK

**DIVA** Diagnostic test that Differentiates Infected and Vaccinated Animals

**ESS** Effective Sample Size

**EU** European Union

**FAP** Four Areas Project

**GAM** Generalized Additive Model

**GB** Great Britain, comprising England, Scotland and Wales

**GLM** Generalized Linear Model

**HPD** Highest Posterior Density

**IFN-gamma** Interferon-gamma

**KL** Kullback Leibler divergence

**MCC** Maximum Clade Credibility

**ML** Maximum Likelihood

**MLE** Marginal Likelihood Estimate

| | |
|---|---|
| **MTBC** | Mycobacterium Tuberculosis Complex |
| **mtDNA** | Mitochondrial DNA |
| **NI** | Northern Ireland |
| **NND** | Nearest Neighbour Distance |
| **NTB** | Unified Neutral Theory of Biodiversity and Biogeography, see Hubbell (2001) |
| **RAD** | Relative Abundance Distribution |
| **RBCT** | Randomised Badger Culling Trial |
| **REA** | Restriction Endonuclease Analysis |
| **RFLP** | Restriction Fragment Length Polymorphism |
| **RoI** | Republic of Ireland |
| **SICCT** | Single Intradermal Comparative Cervical Tuberculin test |
| **SNP** | Single Nucleotide Polymorphism |
| **tMRCA** | Time to most recent common ancestor |
| **UK** | United Kingdom of Britain and Ireland, comprising England, Scotland, Wales and Northern Irela |
| **VNTR** | Variable Number Tandem Repeat |
| **WGS** | Whole Genome Sequencing |

# CHAPTER 1

# Background to bovine tuberculosis in Britain and Ireland

# Background to bovine tuberculosis in Britain and Ireland

## 1.1 Introduction

Tuberculosis in cattle (bovine tuberculosis, bTB), caused by the bacteria *Mycobacterium bovis*, was recognized as an important zoonosis and a cause of human tuberculosis as long ago as the 1800s. In the 1930s the prevalence of *M. bovis* infection in British dairy cattle was as high as 40% and an estimated 2500 human deaths were caused by bTB each year (Pritchard, 1988). Voluntary measures were introduced at this time to reduce the level of bTB in cattle and so reduce the risk to human health. These measures consisted of testing cattle for infection and slaughtering infected animals, a system that remains the mainstay of bTB control to this day throughout the world. In addition to cattle controls, pasteurisation of milk has significantly lowered the risk posed to human health by bTB. The test and slaughter control policy was made compulsory for cattle in Great Britain (GB, made up of England, Scotland and Wales) in the 1950s, and the success of the strategy was such that bTB appeared to be on the verge of eradication in the early 1980s (Reynolds, 2006).

Similar initial successes of the test and slaughter program were seen in Northern Ireland (NI) and the Republic of Ireland (RoI). Despite these early achievements, bTB is still present in the British Isles (which consist of GB, NI and RoI) and, with the exception of Scotland, has proved impossible to eradicate to date. However, many other countries have managed to use the test and slaughter policy to successfully eliminate bTB in cattle. Where this has not been possible, the presence of infection in wildlife reservoirs

is often suggested to be the reason for this failure (Nugent, 2011; Okafor *et al.*, 2011; Reviriego Gordejo & Vermeersch, 2006). In the British Isles the Eurasian Badger *Meles meles* is implicated as a reservoir for cattle infection, and bTB in this species appears closely linked to cattle infection in the British Isles (Woodroffe *et al.*, 2005), although the issue is controversial, with strongly held views on both sides.

This chapter aims primarily to give an overview of *M. bovis* as the causal agent of bovine tuberculosis (bTB) in the British Isles, and in so doing to explain the rationale for the PhD project put forward here. Tuberculosis in other species, primarily *M. bovis* infection of badgers and *Mycobacterium tuberculosis* infection of humans, is also discussed where the comparison is helpful to understanding, or where the information is relevant to the situation in cattle.

## 1.2 Pathogenesis

The Mycobacteria are a genus of bacteria characterised by their waxy cell wall, many of which live a saprophytic lifestyle. *M. bovis*, the causal agent of bTB, is one of several slow-growing pathogenic mycobacteria that make up the Mycobacterium tuberculosis complex (MTBC) (Brosch *et al.*, 2002). Members of the MTBC have different host preferences but cause a similar disease (tuberculosis) in their various hosts. As its name suggests, *M. bovis* is primarily a pathogen of cattle, although it has the potential to infect a wide range of host species including man, having one of the widest host ranges of any zoonotic pathogen (Reilly & Daborn, 1995). The MTBC group also includes *M. tuberculosis*, the primary cause of human tuberculosis, which give rise to 1.45 million deaths worldwide in 2010 (WHO, 2011).

The hallmark of tuberculosis is the granuloma, a lesion which typically consists of a core of infected macrophages and giant cells surrounded by layers of other immune cells, which is then (in some species) surrounded by a final layer of fibrous tissue, walling off the infection (Lin & Flynn, 2010). Depending on the stage of disease, an infected individual may have a variable number of granulomas of different sizes in various parts

of the body (Cassidy, 2006). The progression or containment of each granuloma is dependent on the balance between the immune system and bacterial replication at the level of the individual lesion, and the lesions appear to develop independently of each other (Russell *et al.*, 2009).

The majority of our understanding of tuberculosis pathology and pathogenesis comes from studies relating to *M. tuberculosis* infection of humans, and the following discussion of pathogenesis starts with an overview of the general understanding of tuberculosis before going on to discuss *M. bovis* in cattle and badgers.

Broadly speaking, in an individual lesion the balance between bacterium and immune response may proceed in three ways:

- Following infection, the bacteria may prevail and lesions will advance, seeding more bacteria through the body and causing generalised disease.
- The host's immune system and the bacteria system may reach an equilibrium, in which the bacteria survive but do not spread. This balance can continue for many years and sometimes for the entire lifespan of the host, however impairment of the immune system can shift the balance in favour of the bacterium leading to recrudescence of the disease.
- The host's immune system may win through, resulting in removal of the bacteria and resolution of the lesion.

At the host level, infection may result in several possible outcomes: the host may clear infection completely; infection may establish but be contained, generally in lesions in the lungs or associated lymph nodes (latent tuberculosis: in humans 10% of latent infections recrudesce to active tuberculosis over the host's lifetime, increasing to 10% per year with concurrent HIV infection); or the bacteria may overwhelm the host immune response, leading to primary active tuberculosis with lesions disseminated in different parts of the body (Frieden *et al.*, 2003).

In the classic understanding of latent tuberculosis it was thought that bacteria were contained in a state of non-replicative dormancy (Gomez & McKinney, 2004). However this theory has recently been challenged, and studies have suggested that bacterial replication does occur in latent infections but that this is balanced by host killing of bacteria, keeping the number of viable bacteria at a stable level (Ford *et al.*, 2011; Gill *et al.*, 2009). It has to be remembered though, that diagnosis of tuberculosis is categorical: in humans diagnostic tests are used to classify an individual as either infected or not, and the presence or absence of clinical signs of disease decides whether an infection is latent or active. In this situation, our understanding of the events occurring during latency is likely to be additionally confused by this imposition of simple categories on what is most likely to be a complex and continuous spectrum of disease states (Barry *et al.*, 2009).

Mycobacteria are intracellular pathogens and survive within vesicles inside macrophages (phagocytic cells of the immune system) (Russell, 2011). Infected macrophages stimulate cell-mediated immunity (CMI) resulting in the classic tuberculosis pathology, the granuloma, described above. CMI is key both to the immunopathology seen in tuberculosis, and also to the containment of the disease. CD4+ T-cells and other cells of the immune system produce cytokines such as interferon- (IFN-) gamma and tumour necrosis factor alpha, which orchestrate the responses of other immune cells with the aim of ultimately killing or containing infected macrophages (Lin & Flynn, 2010). Only in late stage disease does CMI wane (so-called anergy), giving way to humoral immunity and antibody responses (Pollock *et al.*, 2003), see Figure 1.1.

Excretion of bacteria in infected individuals varies depending on the stage of the disease. In latent disease, excretion of bacteria is low-level and intermittent (Behr *et al.*, 1999; Cassidy, 2006), and with pulmonary disease bacteria are generally excreted into the respiratory secretions. As the disease progresses to the point at which clinical signs are apparent the numbers of mycobacteria excreted increases hugely, with the route of excretion dependent on the site of the lesions (Neill *et al.*, 2001).

In *M. bovis* infection of cattle, advanced disease was common historically (Medlar,

1940; Stamp, 1948), with lesions spreading from the lungs to other sites in the body. Udder lesions led to excretion of bacteria into the milk, the most common route for human infection. After 50 years of control measures, the distribution of lesions in cattle in the British Isles is now very different. Cattle are generally culled in the early stages of disease, and lesions seen at post mortem are generally few and small. In the British Isles today, bTB granulomas are generally detected in the lymph nodes associated with the lungs and upper respiratory tract and cervical lymph nodes, and to a lesser degree in the lungs themselves (Liebana *et al.*, 2008). This distribution suggests that the primary route of infection is respiratory in cattle, although it is possible that cervical lesions may be associated with oral exposure.

In badgers, *M. bovis* infection can remain latent for years without any apparent detrimental effects before progressing to clinical disease and death, with intermittent shedding of bacteria in the earlier stages of pulmonary disease, and an apparent relationship between the severity of the disease and the amount of bacteria shed (Gallagher & Clifton-Hadley, 2000). Respiratory transmission is again suggested to be the primary route of spread, resulting in lesions in the lungs and associated lymph nodes, and excretion through the respiratory secretions (Corner *et al.*, 2011). However, infection via bite wounds is also seen in badgers, and progression of disease is thought to be quicker via this route of infection (Clifton-Hadley *et al.*, 1993; Jenkins *et al.*, 2007). BTB-induced mortality is more rapid in male badgers than in females (McDonald *et al.*, 2014), and the disease also appears to progress faster in badgers infected as cubs (Delahay *et al.*, 2013).

## 1.3 Diagnosis

As described above, tuberculosis is often characterised by a lengthy latent period where an infected individual may intermittently excrete bacteria at a low level, before the disease becomes clinically apparent. To prevent the spread of tuberculosis it is therefore important to diagnose and control infection at this pre-clinical stage. However, the

nature of the disease and of the immune response make this challenging: walled off lesions and intermittent shedding mean detection of the bacteria themselves is difficult in pre-clinical infection, and CMI responses are more challenging to measure than antibody responses.

In both humans and cattle the two main methods for diagnosing pre-clinical infection, the skin test and the IFN-gamma test, measure the CMI response. Responses to these tests over the course of infection are shown schematically in Figure 1.1. It is interesting to note that in cattle, if the responsiveness to the bTB tests correlates with protective immunity, we may be selectively culling out resistant animals (Amos *et al.*, 2013).

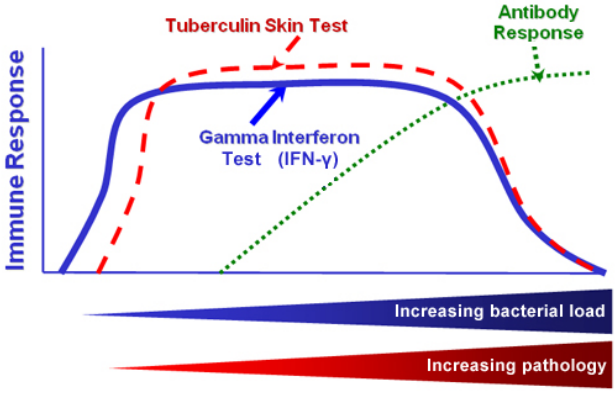


**Figure 1.1: Schematic representation of the spectrum of responses of the bovine immune system to the various tests for bTB** - reproduced with permission from de la Rua-Domenech *et al.* (2006).

### 1.3.1 Skin test

The skin test has for many decades been the mainstay of the control of tuberculosis in both cattle and humans. It is based on the measurement of delayed-type hypersensitivity in response to the intradermal injection of tuberculin, a glycerol extract from the tuberculosis bacterium (Francis, 1947). The tuberculin is recognised by the immune system of animals that have been exposed to or infected with the same species of bacterium, resulting in inflammation and influx of immune cells to the site of the injection and measured as an increase in skin thickness 48-72 hours post-injection (Pollock *et al.*, 2003). Following an injection of tuberculin, the response to further testing

can be affected for over a month afterwards (Monaghan *et al.*, 1994).

The fundamental test has remained little changed since it was conceived in the late 1800s (Monaghan *et al.*, 1994). In many countries, a single injection of tuberculin is used for diagnosis of tuberculosis in cattle. By contrast, in the British Isles cattle are injected with tuberculin derived from a strain of *M. bovis*, and this is then compared with the response to injection with a *Mycobacterium avium* tuberculin. This improves the specificity of the test (Monaghan *et al.*, 1994), an important consideration in areas such as the British Isles, with a low incidence of bTB and where a false positive test has a significant impact and cost. This is known as the single intradermal comparative cervical tuberculin (SICCT) test.

The skin test is accepted to provide a very high level of specificity (i.e. gives very few of false positive results), however the sensitivity of the test is considerably lower, meaning that the test may miss a significant number of diseased animals (de la Rua-Domenech *et al.*, 2006; Defra, 2011). The test sensitivity and specificity and the testing regime for cattle in the British Isles is discussed in more depth in following sections.

### 1.3.2 IFN-gamma test

The fact that, despite its imperfections, the skin test has for many decades been the primary means to diagnose latent tuberculosis is indicative of the difficulties involved in developing accurate diagnostic tests. In more recent years though, another test, the IFN-gamma test has shown promise.

This test involves mixing a separated blood sample from the animal with *M. bovis* purified protein derivative. In animals that have been exposed to *M. bovis*, sensitised immune cells within the blood release IFN-gamma, and this is measured after 16-24 hours using an ELISA (Ryan *et al.*, 2000). Again, in the British Isles this test is conducted comparatively, using *M. bovis* and *M. avium* purified protein derivative. The IFN-gamma test sensitivity appears to be similar to the skin test, although it may

identify a different population of infected animals (Figure 1.1) (de la Rua-Domenech *et al.*, 2006; Defra, 2011). However it does have a somewhat lower specificity Defra (2011), and therefore it may be inappropriate for stand-alone use in a low prevalence situation such as the British Isles. Additionally, this test avoids the need to revisit the farm, and it is thought to identify animals at an earlier stage of infection, and therefore may have a role alongside the skin test, especially in herds where infection is proving difficult to control.

### 1.3.3 DIVA tests

At present, a major barrier to vaccination against bTB in cattle in the British Isles (see below) is the inability of the tests described here to distinguish between animals that are truly infected and animals that have been vaccinated (European Food Safety Authority, 2013). The commonly used Bacille Calmettte Guerin (BCG) vaccine lacks several of the antigens produced by wild-type *M. bovis*, and this has facilitated the development of both IFN-gamma and tuberculin skin tests that specifically identify the immune response to these wild-type antigens (Jones *et al.*, 2012; Vordermeier *et al.*, 2011). These are termed DIVA tests due to their ability to "Differentiate between Infected and Vaccinated Animals". The next step towards vaccination will be the validation of these DIVA tests in the field, with EU authorisation likely to depend on demonstration of sensitivity and specificity comparable to that of the currently-used bTB tests (Conlan *et al.*, 2015).

### 1.3.4 Testing in badgers

Given the various issues described below regarding indiscriminate culling of badgers, targeted removal of infected badgers could be an attractive solution to the issue of bTB in this species. Although the tuberculin skin test used in cattle is not an option in badgers - badgers do not develop the hypersensitivity reaction that the test is based on (Gallagher & Clifton-Hadley, 2000) - an IFN-gamma assay, similar to the cattle

IFN-gamma test, has been developed for badgers, and shows relatively high sensitivity and specificity (Dalley *et al.*, 2008). However, in general terms it is impractical (and unethical) to keep badgers restrained for the time necessary to obtain the test results, and therefore this is not ideal as a trap-side test in badgers.

Further diagnostic tests, based on measurement of the levels of *M. bovis*-specific antibodies in the blood, have also been validated in badgers (Clifton-Hadley *et al.*, 1995; Greenwald *et al.*, 2003). These generate results in a much shorter space of time than the IFN-gamma test, making them feasible as trap-side tests, but at the cost of a reduced sensitivity in the general population. However, this sensitivity is reported to be higher (up to 78%; Chambers *et al.*, 2008) in the later stages of the disease as antibody levels increase (Figure 1.1), therefore these tests may preferentially identify the badgers most at risk of onward transmission of bTB infection.

Non-invasive quantitive PCR tests have also been developed, allowing identification and quantification of *M. bovis* bacteria in soil or faecal samples, although the sensitivity of these is very low (Courtenay *et al.*, 2006; King *et al.*, 2015), probably as a result of the low levels of bacteria shed in the faeces during infection (Corner *et al.*, 2011). Due to the territorial nature of badgers and the use of latrines as boundary markers, this test may enable detection and targeted removal of infected badger social groups.

## 1.4 Vaccination

Vaccination against tuberculosis is as challenging as diagnosis. In human medicine, despite much research, the only current vaccine is the BCG, which consists of a strain of *M. bovis* attenuated through repeated laboratory culture. This method of vaccination is now almost a century old and, although it provides a degree of protection against severe childhood forms of tuberculosis and has shown to be relatively safe (with the exception of immunocompromised individuals), it is less effective against adult pulmonary tuberculosis (Roche *et al.*, 1995).

Finding a suitable vaccination for cattle is even more difficult. The BCG vaccine is currently the only available vaccine, and although it does reduce severity of infection and shedding, it does not provide complete immunity (although challenge studies of vaccine efficacy use higher infection doses than would be encountered in the field, and therefore field efficacy may be higher than reported by laboratory studies; Chambers *et al.*, 2014). However, it is not possible to differentiate BCG-vaccinated animals from infected animals using the currently approved tests (but see the discussion of DIVA tests above), and therefore widespread vaccination would be illegal under EU legislation, compromising the test and slaughter program and causing major issues for cattle exports (European Food Safety Authority, 2013).

Vaccination of badgers has also been proposed as an alternative means of reducing the levels of bTB in the cattle population. An injectable formulation of the BCG vaccine has been licenced for use in badgers: again, this provides imperfect immunity, however the reduction in disease progression and shedding has been shown to provide herd immunity in the field (Carter *et al.*, 2012; Chambers *et al.*, 2011). Delivery of the injectable vaccine requires individual trapping of badgers, and as such is highly labour intensive and is estimated to cost between £2000-£4000 per $km^2$ (Chambers *et al.*, 2014). Despite this, vaccination of badgers is ongoing in Wales and is also being carried out under licence by groups in various areas of England. In New Zealand, an oral formulation of BCG has been shown to be effective against bTB infection in possums (Tompkins *et al.*, 2009). This route of administration would clearly be a much easier means of delivering the vaccine to wild badgers, and work validating the use of oral BCG to vaccinate badgers is ongoing (Chambers *et al.*, 2014).

## 1.5 Badgers and bTB

The Eurasian badger is thought to be the primary wildlife reservoir of bTB in the British Isles. As its name suggests it is found through Europe and into Asia, and in the British Isles badgers are present in high numbers, reaching population densities of

30 badgers/km$^2$ and higher in some areas of southwest GB (Roper, 2010). On average though, densities are estimated at 1.39 badgers/km$^2$ in GB, 2.75 badgers/km$^2$ in NI and 1.19 badgers/km$^2$ in RoI. In GB, badger numbers are increasing, possibly as a result of their protected status under UK law, and in NI numbers appear to be static (Reid *et al.*, 2011) while in RoI they may be decreasing (Sleeman *et al.*, 2009). At high population densities badgers live in reasonably stable, territorial, mixed-sex social groups, with each group occupying one or more setts (underground chambers connected by tunnels), although in lower density areas badger social groups become less stable (Kruuk & Parish, 1987), and badgers also appear to range further (reviewed in Byrne *et al.*, 2014). Badger movements tend to be restricted to their own and neighbouring social groups, with the majority of badger movements occurring over distances of under 5km, although longer distance dispersal events do occur more rarely (Byrne *et al.*, 2014; Pope *et al.*, 2006). Badgers are omnivorous and can live off a variety of foods, but in the British Isles earthworms form a large part of their diet. This is relevant as earthworms tend to be abundant on grazed cattle pastures and so this dietary preference can put badgers in close proximity to cattle areas (Krebs *et al.*, 1997; Roper, 2010).

*M. bovis* infection has been identified in a large number of wildlife species in the UK and elsewhere (Delahay *et al.*, 2001). In badgers, infection with *M. bovis* was first identified in the 1970s, and was quickly linked to bTB in cattle (Anon, 1975; Muirhead *et al.*, 1974). In order to assess the risk a species poses to cattle infection, several factors need to be evaluated: the prevalence of infection in the wildlife host, the abundance of the host species, whether the species comes into contact with cattle, and how likely bTB is to be transmitted if this occurs.

Badger populations in the British Isles have a high prevalence of bTB infection (estimated at 16.6% in GB (Bourne *et al.*, 2007) and 19.5% in RoI (Griffin *et al.*, 2005a)), though this may reach up to 40% in some populations (Menzies & Neill, 2000). Badgers are found at high densities, especially in the southwest of GB, a notable hotspot of cattle bTB infection. Badgers' preference for earthworms means they frequent cattle pastures, and they are also known to enter cattle housing to access feed. Badgers in

the final stages of bTB may be more likely to leave their social group and take up residence somewhere with easy access to food (such as cattle housing), and cattle are thought likely to investigate moribund or dead badgers if they encounter them (Corner, 2006; Roper, 2010). Experimental studies have demonstrated that badgers are able to infect cattle with bTB (Little *et al.*, 1982) , and culling trials in GB and the RoI have shown that reducing badger numbers significantly alters the incidence of bTB in cattle (Bourne *et al.*, 2007; Gilbert *et al.*, 2005). In light of all the above, badgers are seen as the most likely wildlife reservoir of bTB infection in the British Isles. Interestingly though, bTB prevalence in badgers does not appear to increase with badger population density as might be expected given the nature of the disease. In fact, the opposite occurs and prevalence is actually higher at reduced densities and smaller group sizes (although absolute number of infected animals may still be lower), thought to be a reflection of social behaviours at different densities (Woodroffe *et al.*, 2009).

It has been suggested that *M. bovis* infection in deer may also prove a risk to cattle, as *M. bovis* has been found in wild deer at significant prevalences, and deer species also appear to develop the advanced lesions which facilitate onward transmission of the disease (Delahay *et al.*, 2007). However, bTB in deer is thought to be more likely to be the result of spillover infection from other species (Nugent, 2011), except where deer are found at artificially high densities such as farmed red deer in New Zealand and white tailed deer in Michigan (Griffin & Mackintosh, 2000; Nugent, 2011; Okafor *et al.*, 2011).

Despite much research into badgers and their role in bTB, many uncertainties remain, and the exact route by which badgers may transmit bTB to cattle is still unknown. Infection by aerosol is the most efficient method of transmission, and respiratory spread would fit with the pattern of lesions in infected cattle (Liebana *et al.*, 2008). However, aerosol transmission would require close contact between cattle and badgers as infectious aerosols are subject to rapid desiccation (Nardell, 2004), but close contact appears unusual (Bentham & Broom, 1989; Drewe *et al.*, 2013), though it may occur in the situations mentioned above. The other route by which badgers may infect cattle

is by contamination of pasture, feed, or the environment. Transmission of *M. bovis* by ingestion requires a much (1000x) higher infectious dose than by aerosol, probably due to the barrier presented by stomach acid (Corner *et al.*, 2011). However, on investigation of an object cattle will give an initial strong exhalation, and this might be enough to aerosolise any *M. bovis* contaminating the area (Krebs *et al.*, 1997). It is also worth noting that transmission of bTB between cattle and badgers does not occur solely in one direction: suspension of bTB testing in cattle during the foot and mouth epidemic of 2001 was followed by a rise in bTB prevalence in badgers (Woodroffe *et al.*, 2006), implying that a rise in the levels of bTB in cattle results in increased spillover from cattle into the badger population. In fact, badgers may have originally become infected with bTB as a result of transmission from infected cattle (Anon, 1975).

Several culling strategies and trials of badger culling have been implemented in different areas of the British Isles, and these are discussed in more detail in the following section.

## 1.6 Controlling bTB in the British Isles

Despite the initial success of the test and slaughter program in the British Isles, eradication of the disease has proved impossible to date, except in Scotland. BTB is currently found throughout NI and RoI at relatively stable herd level incidences (Abernethy *et al.*, 2013). In GB, by contrast, the level of bTB in cattle is not stable, and in the southwest in particular has increased markedly since the mid-1980s (Godfray *et al.*, 2013). The majority of bTB in GB is found in a core area in the southwest and Wales, which is expanding year on year (Brunton *et al.*, 2015), while short-lived outbreaks outside of this area appear to be seeded by cattle moved out of the high-risk areas (Gilbert *et al.*, 2005). By contrast, the EU has recently classified Scotland as officially bTB free (Hall, 2010).

### 1.6.1 Cattle controls

As cattle vaccination is currently prohibited under EU legislation, control of bTB in cattle in Britain and Ireland is centred on test and slaughter protocols in combination with routine abattoir surveillance. Additionally, research into the genetic basis of susceptibility to bTB in cattle raises the possibility that in the future, selective breeding for bTB-resistant cattle may also play a role in managing the disease (Allen *et al.*, 2010; Bermingham *et al.*, 2014; Tsairidou *et al.*, 2014; Wu *et al.*, 2015).

Cattle control programs are similar throughout the British Isles and are governed by EU legislation. This stipulates testing of cattle herds every one to four years according to local prevalence using the SICCT test described above, with the IFN-gamma test also permitted for ancillary testing (de la Rua-Domenech *et al.*, 2006). In addition to this, carcasses are examined for bTB lesions at slaughter and any suspicious visible lesions are sent for culture and histopathology.

Testing is conducted on an annual basis in NI (DARD NI, 2015a) and RoI (DAFM, 2015b). In England, the country is divided up into separate areas based on bTB risk (Defra, 2014). In high risk areas with endemic bTB in the south and west of England , and in edge areas acting as a buffer around the high risk areas, bTB testing is carried out annually, with pre-movement testing of cattle also carried out in the high risk areas. In low risk areas where bTB has yet to become established, cattle are tested every four years. In Wales, all cattle are tested annually, with specific focus on an intensive action area in the west of the country (Welsh Government, 2015). In Scotland, no routine tests are necessary although post-movement testing is conducted on animals imported from endemic areas (Reviriego Gordejo & Vermeersch, 2006).

For routine testing, the standard interpretation of the SICCT test is used to identify infected cattle (so-called reactors). A positive reaction to the standard test is defined as a difference of >4mm in skin thickness between the *M. avium* and *M. bovis* injection sites 72 hours post injection. On discovery of reactor animal(s) within a herd (termed a herd breakdown), movement restrictions are imposed and reactor animals

are slaughtered. If bTB lesions are seen at post mortem or if *M. bovis* is cultured from tissue samples then the animal is known as a confirmed reactor. Some positive tests are not confirmed at post-mortem: in GB this amounts to 6% (Bourne *et al.*, 2007), however this is not necessarily a sign of a false positive - small lesions (under 2mm) are difficult to detect on post mortem examination, and mycobacteria can be difficult to culture. Testing of the breakdown herd is repeated at a 60-day interval using the severe interpretation of the SICCT test, whereby a >2mm difference in skin thickness is a classified as a reaction (so increasing test sensitivity, with a corresponding reduction in specificity). The movement restrictions, testing at 60-day intervals using the severe interpretation, and the herd breakdown itself, continue until the herd has had two consecutive clear tests.

The above testing protocol is similar across the British Isles, with the IFN-gamma test also being used in problem herds. In RoI, herd depopulation followed by cleansing and disinfection of the premises and a period of at least four months without cattle is employed for problem herds in which repeated testing has failed to resolve bTB infection (Good *et al.*, 2011).

These cattle control measures come at a cost, most of it borne by the government (and therefore the tax-payer): £500 million has been spent on bTB in England in the past 10 years (Defra, 2014). With pasteurisation of milk, the risk to public health has been minimised, and the majority of human *M. bovis* infections in the British Isles are a result of latent infection reactivated after decades, or from infection contracted abroad. In view of this, some have suggested that the taxpayer should not be responsible for what is effectively a trade issue (Torgerson & Torgerson, 2009; White *et al.*, 2008). The counter arguments note the increasing prevalence of HIV-AIDS, which facilitates infection (Amanfu, 2006), the recent cluster of *M. bovis* infection in six people in Birmingham (Evans *et al.*, 2007), and the fact that if controls were stopped, progressive lesions in cattle would increase the level of zoonotic risk (Amanfu, 2006). The EU still considers bTB of importance to human health (Reviriego Gordejo & Vermeersch, 2006).

Although the true sensitivity and specificity of bTB tests are difficult to measure in the

absence of a perfect gold standard test to diagnose *M. bovis* infection (Karolemeas *et al.*, 2012), the imperfect sensitivity of the available tests pose a considerable hindrance to control of the disease as it makes it likely that infected cattle are left undetected in herds. Despite this, the summary report on the Randomised Badger Culling Trial in GB (see below) concluded that the rising incidence of disease can be reversed by the rigid application of cattle-based control measures alone (Bourne *et al.*, 2007). A continued focus on improving the control of the bTB in cattle is likely to prove important in the journey towards eradication of the disease (Brooks-Pollock *et al.*, 2014).

### 1.6.2 Badger controls

Despite the imperfections of the bTB test, the test and slaughter protocol has been successful in eradicating the disease in many countries around the world (Amanfu, 2006). The failure to eradicate the disease in Britain and Ireland after many decades of cattle controls points to the likelihood of bTB infection in a wildlife reservoir (Reviriego Gordejo & Vermeersch, 2006).

Badgers are a charismatic and much-loved native wildlife species in the British Isles, and there are significant sociological barriers to the principle of badger culling, in GB especially (see for example www.badger.org.uk, http://badger-killers.co.uk/), although the vocal and well-publicised nature of the opposition to badger culling may make it seem larger than it actually is (White & Whiting, 2000). However, it does mean that any controls that seek to reduce bTB in cattle by interfering with the badger population must be underpinned by rigorous scientific evidence if they are to avoid impassioned criticism. While cattle controls are similar across different regions of the British Isles, the situation with badgers is much more varied, as discussed below.

In GB various culling strategies have been implemented since the 1970s. These were initially based on gassing of setts, and then used snaring after welfare issues were identified with gassing. The most recent strategy was the interim strategy which during the 1980s and 1990s, involving reactive small-scale removal of badgers in response to

cattle outbreaks (Krebs *et al.*, 1997).

In GB between 1998 and 2006 a large scale, controlled, randomised, and replicated, culling trial was conducted, the Randomised Badger Culling Trial (RBCT), overseen by the Independent Scientific Group and summarised in Bourne *et al.* (2007). It was carried out to investigate the effects of reactive badger culling (similar to the interim strategy described above), proactive culling (continued culling of badgers across a large area), compared to control areas where no culling was conducted. The trial was carried out between 1998-2007, and its results were unexpected. Reactive culling was suspended in 2003 as it was found to increase bTB incidence in cattle by 27% in comparison to no cull areas (but with wide confidence intervals). BTB incidence did fall in proactive areas by 23%, although this was offset by increases in the incidence of bTB in cattle of 25% in the 2km border surrounding the proactive areas. These results were explained by the perturbation hypothesis: culling is thought to disrupt stable badger social groups, leading to increased contact between different groups and immigration of badgers from surrounding areas, and therefore increasing the chances for bTB to spread between badgers (Donnelly *et al.*, 2006). This is supported by the recorded increase in bTB prevalence in badgers after proactive culling, and also by disruption of the clustered structure of *M. bovis* strains in the culled areas (Woodroffe *et al.*, 2005). The Independent Scientific Group concluded that badger culling had "no role in the cost-effective management of bTB in cattle in GB" (Bourne *et al.*, 2007).

More recently, England and Wales have diverged in their measures to control bTB in badgers. No badger culling is currently being conducted in Wales, while in England, pilot badger culls were licenced in 2013 in Somerset and Gloucestershire to evaluate the use of free shooting to remove badgers (Defra, 2014). Despite controversy over the extent that the English culls met their stated aims (Abraham *et al.*, 2014; Bateson *et al.*, 2015), they have continued and are about to be extended to include Dorset (Anon, 2015a). In both England and Wales, injectable delivery of the BCG badger vaccination is carried out under licence: in Wales this is specifically targeted to the Intensive Action Area in the west of the country (Anon, 2015b).

**Table 1.1: Possible reasons for the disparity between the findings of the Randomised Controlled Badger Culling Trial and the Four Areas Project** - reproduced with permission from O'Connor *et al.* (2011).

| | Randomised Badger Culling Trial (RBCT) | Four Areas Project (FAP) |
|---|---|---|
| Cattle demography | Larger herd sizes in GB provide opportunities for persistence of infection within herds, and through cattle movements, transmission of infection between herds, potentially affecting the incidence of cattle within trial sites | The smaller herd sizes in RoI and the associated differences in management in comparison to GB may have affected herd incidence and persistence of infection within the herd which would in turn have influenced the outcomes of the trials. |
| Badger demography | The higher badger density in GB may have affected the degree of perturbation observed in the badger populations during the RBCT. | The lower badger density in RoI may have meant that with consistent culling, disturbance to the social structure of badgers would have been minimised. |
| Epidemiology of disease in cattle | bTB epidemic with rapid spatial spread of high bTB risk areas that has partially impacted trial sites. | Endemic bTB situation, with the effective reproduction number likely close to one; evidence of transient hotspots. |
| Epidemiology of disease in badgers | Analysis of badgers from proactively culled sites during the RBCT revealed an increase in prevalence on successive culls. | Prevalence, as assessed through analysis of continued badger culling operations, appears to be stable at a national level. |
| Previous culling | The higher density of badgers removed in trial sites may have resulted in a socially unstable badger population prior to the start of the RBCT. | The lower density of badgers removed prior to the trial could mean that the badger groups within the FAP trial sites were more socially stable than those in the RBCT. |
| Consent to cull | Un-culled land may have provided a residual population of perturbed badgers who could have promptly re-populated trial sites following cessation of culling. | Vast majority of land within culling sites allowed for badger culling, therefore reducing residual population of badgers within trial sites. |
| Trappability as affected by trapping method | Trappability may have been affected by animals missed by cage-trapping or cage shy badgers, leading to residual populations. | The use of stopped restraints in FAP may have resulted in more badgers being trapped and thus less residual badgers remaining with disturbed social behaviours. |
| Application of barriers/ buffer zones to trial sites | Immigration of badgers from social groups outside of trial sites was not impeded, allowing free movement of badgers with potential augmentation of behaviour to re-populate sites. | Barriers surrounding trial sites may have hindered the immigration of badgers into trial sites, which coupled with consistent culling, would have reduced residual populations and immigrant badgers in removal sites. |
| Badger culling consistency | Inconsistent culling of proactive sites may have resulted in residual populations of badgers and in turn behavioural change in remaining badgers, potentially leading to increased incidence of bTB in cattle. | Five years of consistent culling would have removed resident badgers and continued to remove immigrating badgers over culling period, preventing observation of behavioural changes in badgers in trial sites. |

In NI a Test and Vaccinate or Remove (TVR) project is currently underway in a localised area of the country (DARD NI, 2015b). The strategy consists of trapping and testing of badgers using an antibody test, vaccinating and releasing those badgers that test negative, and culling badgers that test positive. Vaccinated badgers are identified using microchipping.

In the RoI, another large-scale badger culling trial, the Four Areas Project (FAP) conducted between 1997 and 2002, produced very different results from the RBCT, identifying a 59% reduction in bTB incidence (data from Griffin *et al.*, 2005a presented in Bourne *et al.*, 2007). The FAP was conducted in four areas matched to different types of farmland present in RoI. Two notable differences between the FAP and the RBCT were the inclusion of geographical barriers to badger movement in the FAP (or where this was impossible, buffer zones), and the comparison of proactive areas with reactively culled areas, rather than the use of control no-culling areas as controls. Based on the results of the FAP trial, the present strategy in RoI involves the culling of badgers where they are epidemiologically implicated in cattle breakdowns, and work is ongoing to evaluate the use of oral badger vaccination (DAFM, 2015a)

The conflicting findings from the GB and RoI trials, far from clarifying the situation, have added to the uncertainty over bTB in cattle and badgers. The differences between the results of the two trials are likely to reflect underlying differences in the complex system surrounding the disease in cattle between the two countries, as well as differences in trial design and implementation, and an overview of these is given in Table 1.1 (O'Connor *et al.*, 2011). Independent reviews of both trials have been critical of each of them. King *et al.* (2007) suggested that the ISGs conclusion was unnecessarily negative and that culling *can* be cost-effective in GB , whereas Godfray *et al.* (2005), while agreeing with the qualitative findings of the FAP, suggest that the actual size of the reduction in cattle incidence that was reported may have been inflated.

## 1.7 Molecular markers and genetic epidemiology of *M. bovis*

The genetic makeup of the MTBC group shows very little diversity, with MTBC bacteria identical across 99.95% of their genome (Garnier *et al.*, 2003). Different lineages within the MTBC have been identified by deletions at specific genomic regions (Brosch *et al.*, 2002), and animal strains including *M. bovis* are found nested within the wider diversity of lineages primarily infecting humans (Figure 1.2). Phenotypic differences have been recorded between different lineages of *M. tuberculosis* in humans, for example the Beijing lineage seems more predisposed to the acquisition of drug resistance (Coscolla & Gagneux, 2014), while by contrast only limited phenotypic variation has been identified within *M. bovis* (Wright *et al.*, 2013a,b).

It is generally accepted that little or no recombination occurs within the MTBC group (Hershberg *et al.*, 2008; Pepperell *et al.*, 2013; Wirth *et al.*, 2008, although see also Namouchi *et al.*, 2012), and therefore the genome is inherited as a single unit from parent to offspring. As a consequence of this clonality, population bottlenecks (reduction in genetic diversity through reduction in population size) and selective sweeps (reduction in genetic diversity due to selective pressures favouring a low number of genetic variant(s)) are thought to have had a strong impact on the evolution of the MTBC (Smith *et al.*, 2006).

### 1.7.1 Molecular typing for epidemiology

The lack of recombination within the MTBC group makes molecular strain typing considerably simpler than it might otherwise be. However, the lack of genetic diversity present within the group also means that sequencing of housekeeping genes, as used in Multilocus Sequence Typing schemes in other bacteria, does not provide sufficient discrimination to be used as molecular markers for the MTBC group. Instead, the typing schemes used for MTBC bacteria rely on more variable regions within the genome.

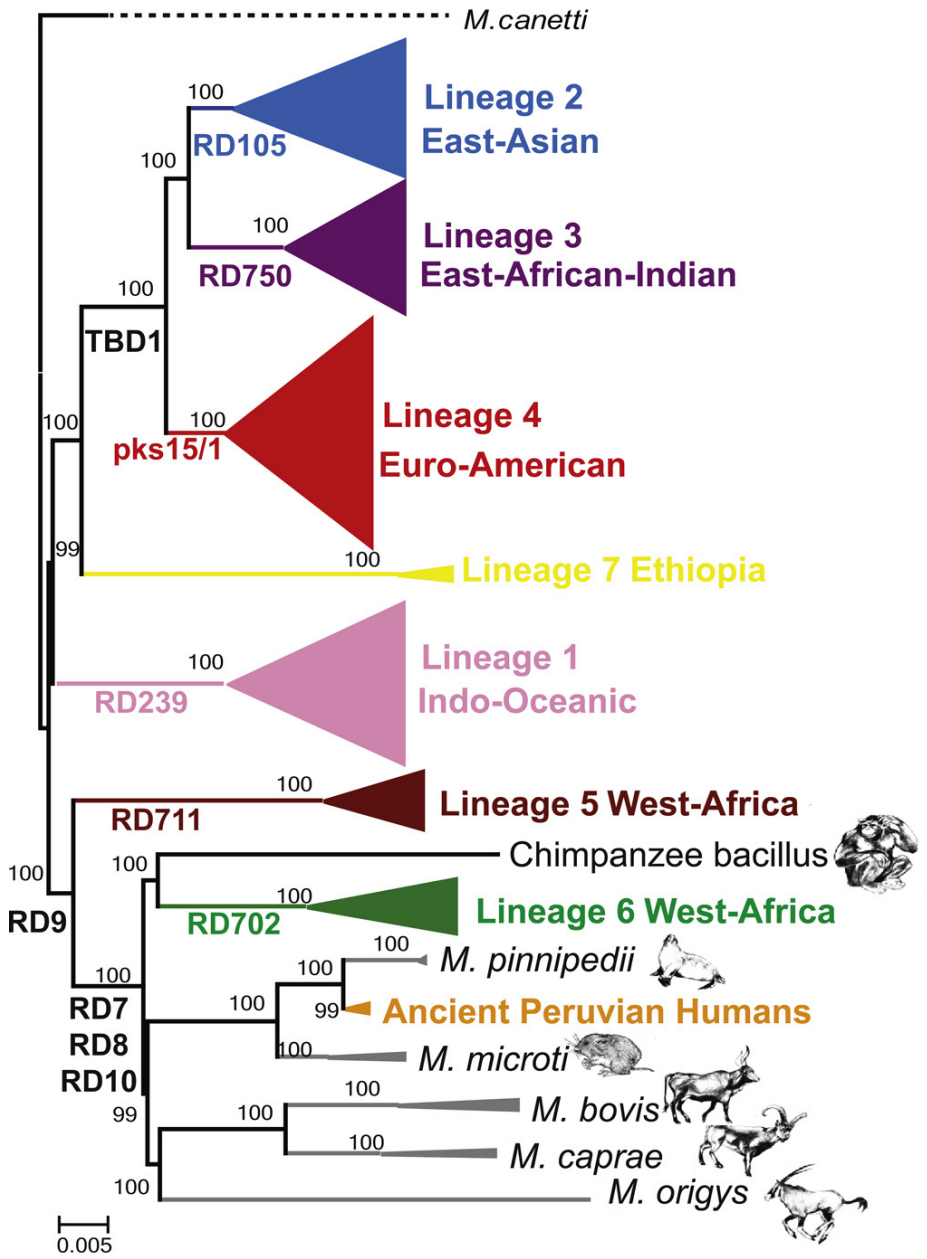


**Figure 1.2: Maximum likelihood phylogeny of the $M.\ bovis$ complex** - Tree is rooted by the outgroup *M. canettii*, scale bar gives nucleotide substitutions per site, and branch labels indicate percent bootstrap node support (bootstraps). Lineage-defining genomic deletions (Brosch *et al.*, 2002) are indicated along the relevant branches. Reproduced with permission from Coscolla & Gagneux (2014).

In *M. bovis*, initial typing methods were based on the use of restriction enzymes to digest bacterial DNA, followed by gel electrophoresis to separate the resultant fragments and staining or hybridization of specific probes to identify discriminatory patterns or polymorphisms. The first of these methods to be developed was Restriction Endonuclease Analysis (REA), used extensively in New Zealand, where the overall pattern of restriction fragments is analysed (Collins, 2011). Another method, IS*6110* Restriction Fragment Length Polymorphism (RFLP), which identifies polymorphisms associated with the IS*6110* region, is widely used to investigate human tuberculosis. However, the low diversity of IS*6110* repeats present in many *M. bovis* strains mean that this method has little discriminatory power for typing *M. bovis* isolates (Smith *et al.*, 2003). The other common restriction enzyme method to be employed is PGRS-RFLP, measuring polymorphisms in the PGRS family of genes (Cousins *et al.*, 1998).

Although the restriction enzyme typing methods such as REA typing can offer great discriminatory power in typing isolates, they are laborious to perform and can be difficult to standardise. With the advent of PCR protocols, strain-typing methods based on DNA amplification have become widespread. In the spoligotyping method, the Direct Repeat (DR) locus, made up of a variable number of repetitive elements separated by spacer regions, is amplified by PCR (Kamerbeek *et al.*, 1997), and the presence or absence of the spacer oligonucleotide repeats is then assessed by DNA hybridisation. Although simpler to perform and more reproducible than the restriction enzyme methods, spoligotyping gives considerably poorer discrimination. This is especially the case in situations such as bTB in the British Isles, where diversity is low and the majority of *M. bovis* isolates consist of a single spoligotype, SB0140 (Smith *et al.*, 2003).

In these situations, Variable Number Tandem Repeat (VNTR) typing has been used to distinguish different strains of *M. bovis* (Frothingham & Meeker-O'Connell, 1998; Groenen *et al.*, 1993; Mazars *et al.*, 2001). In this technique, a series of polymorphic genetic regions known collectively as VNTR loci are used, having been identified from the genome sequences of *M. bovis* and *M. tuberculosis*. Amplification of these regions by PCR and measurement of the size of the resulting products gives a string of numbers,

with each number corresponding to the number of repeats present at the relevant VNTR locus.

VNTR-typing is now used extensively to type *M. bovis* in the British Isles. This method gives results that are significantly more discriminating than spoligotyping alone. However, it also has a higher potential to generate homoplasies, where unrelated bacteria independently converge on the same VNTR-type (Reyes *et al.*, 2012), and therefore this is generally used in combination with the more stable spoligotyping method (Skuce & Neill, 2001). VNTR-typing is currently used routinely (in conjunction with spoligotyping) in both NI and GB, although different panels of VNTR-typing loci are used in each region (Skuce *et al.*, 2005; Smith *et al.*, 2006), and therefore the results are not directly comparable between the two regions.

Molecular typing has been used for many years to refine and target contact tracing and outbreak investigations for both human and bovine tuberculosis. Its uses in epidemiology centre on the assumption that infections of the same molecular type share a common ancestry. Isolates belonging to different molecular types are unlikely to be linked epidemiologically, while in human tuberculosis in low prevalence areas, if infections belong to the same molecular type this is taken to indicate that they are part of the same outbreak (van Soolingen, 2001). However, given the relatively slow timescale over which molecular types evolve, isolates belonging to the same molecular type cannot be assumed to be directly linked epidemiologically in higher prevalence settings.

In human epidemiology, molecular typing has been involved in tracing the spread of disease in various settings including hospitals and prisons, and in identifying risk factors involved in outbreak transmission (van Soolingen, 2001). It has shown that recent transmission plays a larger role than previously thought in the spread of tuberculosis, as compared to the reactivation of historically acquired infections, and also highlighted the presence of individuals infected with multiple strains of tuberculosis, especially in high prevalence settings (Cohen *et al.*, 2012).

Molecular typing has been applied in a similar manner to aid in the investigation and control of bTB in various countries around the world, and has furthered our understanding of the historical spread of the disease worldwide (Skuce & Neill, 2001). In Britain and Ireland, the molecular types of *M. bovis* show a strongly clustered geographic pattern (AHVLA, 2012; Diggle *et al.*, 2005; Goodchild *et al.*, 2012; Olea-Popelka *et al.*, 2005; Skuce *et al.*, 2010; Woodroffe *et al.*, 2005). In GB this has allowed the identification of home range areas for the dominant molecular types: a breakdown occurring outside of its home range is assumed to be associated with translocation of infected cattle (AHVLA, 2013; Gopal *et al.*, 2006; Smith *et al.*, 2006), although even with detailed data on cattle movements it is not always possible to identify the original link back to the home range area (AHVLA, 2013). The geographic clustering of molecular types has also been used to identify associations between *M. bovis* isolates infecting badgers and cattle in the same area, giving further evidence for transmission of infection between the two hosts, although without indicating the direction of transmission between them (Goodchild *et al.*, 2012; Olea-Popelka *et al.*, 2005; Woodroffe *et al.*, 2005).

### 1.7.2 Whole genome sequencing

Within the last decade there has been a step-change in the availability of sequence data, brought about by advances in sequencing technologies (Metzker, 2010).The new technologies generate huge amounts of sequence data as short sequencing reads, but are more error prone than the traditional capillary sequencing methods. Complex computational processing is therefore necessary to convert the raw sequence data to usable output, and as yet no standardised one-size-fits-all bioinformatics solution exists. The large amounts of sequence data generated also raise challenges related to data storage.

In comparison to other bacteria, the use of whole genome sequencing (WGS) for *M. bovis* and other members of the MTBC is relatively uncomplicated. Good quality reference genomes exist for *M. bovis* (Garnier *et al.*, 2003) as well as other MTBC lineages (Bentley *et al.*, 2012; Cole *et al.*, 1998), and the lack of heterogeneity and minimal

recombination in the MTBC make mapping of the raw sequence data to these references simpler. However, the presence of repetitive elements within the mycobacterial genome (e.g. VNTR regions, and repetitive genes such as the PE/PPE family) present a considerable problem for the WGS approaches, and until sequencing technologies are able to generate good quality sequences long enough to bridge these repetitive regions they will remain a challenge.

In addition to the potential for improving our understanding of the biology of the MTBC, and for drug and vaccine discovery, the possibility of sequencing the entire genome of many bacterial isolates offers considerable advances in the application of genetics to the epidemiology of tuberculosis (Kao *et al.*, 2014). At its most basic, WGS offers all the advantages of molecular typing but with considerable increase in discriminatory power. WGS has been shown to give significant increase in resolution over molecular typing methods for tuberculosis, while also removing the risk of homoplasy associated with e.g. VNTR-typing (e.g. Walker *et al.*, 2012). In addition to this, the use of sequence data allows the potential to tap into the powerful phylogenetic and phylodynamic approaches originally developed for rapidly evolving viruses (Grenfell *et al.*, 2004). Phylogenetics broadly concerns the evolutionary relationships between organisms, while phylodynamics is focussed more specifically towards pathogens, using pathogen genetic variation and phylogenetics to understand the epidemiological dynamics of the pathogen population. However, the application of phylodynamic approaches to study the epidemiology of tuberculosis needs more investigation: these approaches rely on measurably evolving populations, where genetic mutations are accumulating on an epidemiologically relevant timescale. MTBC bacteria have been shown to evolve very slowly, even when mutations across the whole genome are recorded, which may make phylodynamic inference challenging (Biek *et al.*, 2012; Bryant *et al.*, 2013b; Roetzer *et al.*, 2013; Walker *et al.*, 2012). Additionally, considerable variation in the evolutionary rate has also been noted for *M. tuberculosis* (Bryant *et al.*, 2013b), possibly due to differences in the rate of accumulation of mutations between active and latent infection states (Colangeli *et al.*, 2014, although see also Ford *et al.*, 2011).

There are many reports of the use of WGS to investigate the epidemiology of human tuberculosis (see for example Bryant *et al.*, 2013a; Casali *et al.*, 2014; Eldholm *et al.*, 2015; Gardy *et al.*, 2011; Roetzer *et al.*, 2013; Walker *et al.*, 2012) and substantial interest exists in rolling WGS out for routine clinical use in hospitals (Gardy *et al.*, 2015; Köser *et al.*, 2012a). Where the genetic basis for antimicrobial resistance is known, WGS can further be used to identify drug resistant strains, which are becoming increasingly important in human disease (due to the cost-limited treatment options for tuberculosis in non-human animals, drug resistance is not a problem that has been encountered in *M. bovis* strains). Promising work has also been reported in generating WGS data directly from clinical samples (Brown *et al.*, 2015), removing the need for the costly and time consuming culturing of isolates.

Although not as advanced as in the human field, the use of WGS to understand the spread of bTB has been reported for NI (Biek *et al.*, 2012) and GB (Roberts *et al.*, 2014), and its use is also under investigation for bTB surveillance in New Zealand and the United States.

## 1.8 Thesis overview

This thesis begins by further exploring the use of molecular types to understand the underlying processes involved in the spread and maintenance of bTB. Molecular typing of *M. bovis* isolates, primarily from cattle infections, has been employed for many years for routine surveillance of bTB in Britain and NI, building up a large dataset on the different *M. bovis* types present in this region.

In Chapter 2, I examine the pattern of relative abundances of different VNTR-types isolated from cattle in NI. This builds on Hubbell's neutral theory of biodiversity and biogeography (Hubbell, 2001), as well as on other work examining the pattern in molecular types of *M. tuberculosis* in various settings and *M. bovis* in GB. I go on to use simulation modelling to explore four possible mechanisms underlying the observed relative abundance pattern.

Chapter 3 also focuses on the data on *M. bovis* VNTR-types in NI. Here, VNTR-type data is combined with extensive spatial information on VNTR-types and cattle herds in NI, data on the VNTR-types identified in a survey of badgers found dead on the roads in NI, and on the genotypes of the badgers themselves. These data were used to first assess the correlation between *M. bovis* VNTR-types in cattle and badgers in these data, and secondly to take a landscape genetics approach to evaluating the hypothesis that spread of *M. bovis* within the badger population is driving the spatial clustering seen in VNTR-types infecting cattle.

The second half of the thesis takes a higher resolution viewpoint, using pathogen WGS to focus in on the epidemiology of bTB in defined subpopulations of cattle (Chapter 4) and badgers (Chapter 5).

Chapter 4 builds on the study reported in Biek *et al.* (2012), applying high-density whole genome sequencing to all available *M. bovis* isolates for a single VNTR-type in NI. In this chapter I address the use of WGS in this slowly evolving pathogen to understand the fine-scale spread and persistence of *M. bovis* in the cattle population, focusing specifically on undersampled bacterial populations, the use of genetic divergence to identify mechanisms driving transmission, and the application of phylogeographic tools.

In Chapter 5 I use both WGS and molecular typing data for a well studied, naturally infected badger population, examining the dynamics of *M. bovis* infection within this group as well as their links to infection in cattle. I assess the diversity present in the badgers at the spoligotype and WGS levels, evaluate the temporal and within-badger evolution of the pathogen, and examine the spatial correlations present between spoligotypes present in these badgers and in the surrounding cattle.

# CHAPTER 2

# Multiple processes account for the relative abundances of *M. bovis* VNTR-types in Northern Ireland

# Multiple processes account for the relative abundances of *M. bovis* VNTR-types in Northern Ireland

## 2.1 Summary

The distribution of relative abundances of different species is a commonly studied pattern in ecology. In this chapter I examine the relative abundances of different molecular types of *Mycobacterium bovis* isolated from cattle in Northern Ireland (NI). These show a left-skewed distribution, with a small number of types present at very high frequency and the majority of types being very rare. I demonstrate that, for the VNTR-types, this skew is too extreme to be accounted for by simple neutral ecological processes, and use simulation models to explore four biologically plausible hypotheses as to the underlying processes driving VNTR-type relative abundances. I show that the process of VNTR-type speciation and the manner in which the VNTR typing loci were chosen in NI cannot account for the observed skew. Structuring of the pathogen population to include, for example, a reservoir of infection in a separate host appears to drive the relative abundance distribution in the opposite direction to that observed, generating more even abundances of molecular types. This corresponds to other work showing that structuring of a population through spatial dispersal limitation also evens out the relative abundances of different species. By contrast, historical increases in bTB prevalence, and transmission heterogeneity (superspreading) are both capable of generating the skewed VNTR-type distribution. I conclude that, of the multiple processes that may be affecting the distribution of different VNTR-types, increases in

pathogen population size and/or superspreading are likely to be driving the pattern observed.

## 2.2 Introduction

For many decades there has been a general interest in ecology in relating the relative abundances of different species to the underlying ecological processes at work in the ecosystem. However, care must be taken not to over-interpret these relative abundance distributions (RADs), as in isolation they do not necessarily provide sufficient information to identify or distinguish between the specific ecological events that shape them (Magurran & Magurran, 2005; McGill *et al.*, 2007; Rosindell *et al.*, 2012). In fact, the interactions between simple neutral processes, outlined in Hubbell's neutral theory of biodiversity and biogeography (Hubbell, 2001; hereafter referred to as NTB), have shown a surprising ability to predict the RADs recorded in a wide variety of ecological datasets (Rosindell *et al.*, 2011).

NTB is a theory of ecological drift, traditionally set against the non-neutral hypotheses of niche theory and selective pressures. It is based around the idea of per-capita equivalence: all individuals, irrespective of species, have an equal chance of birth, death, reproduction, and immigration and/or speciation, with the analytical solutions put forward in Hubbell (2001) also relying on an assumption of constant population size. Further developments extend the theory to consider the explicit spatial structure of the population by incorporating the spatially limited dispersal of individuals (Hubbell, 2001; Rosindell *et al.*, 2011). Although it is widely accepted that real ecosystems are not truly neutral and species are not identical, the success of NTB demonstrates that simple processes are often sufficient to drive the diversity observed in many natural systems.

Interest in RADs is not confined to classical ecology, and several studies have examined the relative abundances of different molecular types of tuberculosis-causing mycobacteria in various settings. These tend to show left-skewed distributions similar to those

recorded in ecological datasets, with a small number of molecular types present at extremely high frequency and the majority of strains very rare, however in these datasets the interest lies in elucidating the epidemiological mechanisms responsible for driving the skewed distributions. Luciani *et al.* (2008) showed that the RADs seen in several human tuberculosis datasets did not fit the predictions generated by NTB, and demonstrated that an increasing bacterial population size could account for the distribution in at least one of these datasets. Ypma *et al.* (2013) also examined the RAD of different molecular types of human tuberculosis, this time in the Netherlands, and while suggesting that an increasing bacterial population was unlikely in this system, they demonstrated that variation in the numbers of offspring per infection (superpreading) could explain the skewed distribution observed. Smith *et al.* (2003) suggested that clonal expansion of different bacterial lineages accounted for the RAD of different molecular types of *M. bovis* recorded in cattle in Great Britain (GB).

Bovine tuberculosis (bTB), caused by *M. bovis*, is one of the most important diseases facing the livestock industry in Britain and Ireland, and control is complicated by infection in a wildlife reservoir, the Eurasian badger (Donnelly & Nouvellet, 2013; Donnelly *et al.*, 2006; Griffin *et al.*, 2005b). In NI, all cattle over six months of age are tested for the disease on an annual basis, and intensive molecular typing of these *M. bovis* has been carried out since 2003. Similar to the molecular types in GB (Smith *et al.*, 2003), the molecular types of *M. bovis* in NI also show a left-skewed RAD (Skuce *et al.*, 2005), however the possible epidemiological processes underlying the skew have not been examined for this dataset. In common with *M. bovis* in other parts of the UK and Ireland (Goodchild *et al.*, 2012; Olea-Popelka *et al.*, 2005; Smith *et al.*, 2006; Woodroffe *et al.*, 2005), *M. bovis* infections in NI also appears to cluster geographically by molecular type (Skuce *et al.*, 2010).

Molecular typing of *M. bovis* in NI is conducted through a combination of spoligotyping, based on the presence of multiple spacer oligonucleotides within the Direct Repeat region of the genome (Kamerbeek *et al.*, 1997), and Variable Nucleotide Tandem Repeat (VNTR)-typing (Frothingham & Meeker-O'Connell, 1998; Supply *et al.*, 2001), which

measures the number of repeats at various VNTR loci present in the mycobacterial genome. VNTR-typing gives a higher level of discrimination than spoligotyping, but it is more prone to homoplasies, i.e. the occurrence of the same VNTR-type in unrelated lineages (Reyes *et al.*, 2012). In NI, the specific VNTR-loci used in VNTR-typing have been chosen to optimise the discriminatory power of the technique in this population (Skuce *et al.*, 2005), and VNTR-types are used in conjunction with the more stable spoligotyping.

*M. bovis* as a species shows very little genetic diversity, especially in Britain and Ireland (Smith *et al.*, 2006), and studies have identified only limited differences in certain epidemiological phenotypes (the size of herd outbreaks and the ability of the skin test to detect infection) between different molecular types (Wright *et al.*, 2013a,b). In a low diversity population such as NI, selective pressures are less likely to play a prominent role in shaping the diversity of the *M. bovis* population. However, historical increases in bTB incidence over recent decades (Abernethy *et al.*, 2006) may have affected the relative abundances of the different molecular types, and furthermore, in Great Britain at least, superspreading has been indicated to play a role in the spread of the disease (Brooks-Pollock *et al.*, 2014), which may also impact the RAD observed.

In this chapter, I investigate the processes driving the RAD of molecular types of *M. bovis* isolated from cattle in NI. I first evaluate the null hypothesis that the observed RAD of molecular types in NI can be accounted for by NTB in the absence of other processes, comparing the observed distribution of molecular types to that predicted under NTB. Additionally, I compare the RADs reported in Smith *et al.* (2003) for cattle in GB, and Ypma *et al.* (2013) for human tuberculosis in the Netherlands, to the distributions predicted for these populations under neutral theory to assess whether these conform to NTB. I then go on to use simulation models to examine four alternative hypotheses regarding the mechanisms involved in shaping the distribution of VNTR-types in NI:

- **Hypothesis 1** - The process of VNTR-type speciation and/or the manner in

which VNTR-loci were chosen in NI can account for the observed RAD.

- **Hypothesis 2** - An unsampled pool of infection in a separate but linked population such as a wildlife reservoir can account for the observed RAD.
- **Hypothesis 3** - Historical increases in the prevalence of bTB in NI can account for the observed RAD (Luciani *et al.*, 2008).
- **Hypothesis 4** - Superspreading, i.e. variability in the number of onward infections transmitted from an infected herd, can account for the observed RAD (Ypma *et al.*, 2013).

## 2.3 Materials and Methods

All analyses and simulations were conducted in R v3.3.1 (R Core Team & R Develoment Core Team, 2014).

### 2.3.1 Data on molecular types of *M. bovis* in NI

In NI, all cattle over the age of six weeks are routinely tested for bTB on an annual basis, and bTB surveillance is additionally conducted routinely at slaughter. A breakdown of bTB in a cattle herd is defined as the period of time beginning with the first detection of bTB in an animal from the herd, and ending when the entire herd has passed two consecutive bTB tests at least 60 days apart. Since 2003, all NI bTB breakdowns from which *M. bovis* was isolated have had at least one isolate typed with a combination of spoligotyping and VNTR typing using standard protocols (Skuce *et al.*, 2005). In NI, the specific loci used for VNTR-typing are MV2163B/QUB11B, MV4052/QUB26A, MV2461/ETRB, MV2165/ETRA, MV2163/QUB11A and MV323/QUB3232, although since 2006 an extra VNTR locus, MV1895/QUB1895, has been added to the NI panel in order to split one of the common types into two geographically distinct strains (Skuce *et al.*, 2010). To ensure consistency across the sampled timeframe, in this study

I considered VNTR-types as defined using the original seven-locus VNTR panel plus spoligotype.

**Table 2.1: Numbers of *M. bovis* spoligotypes, VNTR-types, and typed breakdowns recorded by year in NI**

| Year | Number unique spoligotypes | Number unique VNTR-types | Number typed breakdowns | Total number of breakdowns |
|---|---|---|---|---|
| 2003 | 12 | 57 | 1347 | 2902 |
| 2004 | 13 | 61 | 1621 | 2954 |
| 2005 | 12 | 61 | 1432 | 2556 |
| 2006 | 9 | 72 | 1285 | 2033 |
| 2007 | 10 | 72 | 1235 | 1862 |
| 2008 | 11 | 72 | 1115 | 1789 |
| 2009 | 10 | 72 | 1080 | 1771 |
| 2010 | 9 | 72 | 934 | 1617 |
| Total (unique) 2003-2010 | 16 | 185 | 10049 | 17484 |

Details of herd breakdowns were extracted from the APHIS database (Houston, 2001) by David Wright (Queens University Belfast) and restricted to those breakdowns beginning in the years 2003-2010 inclusive. Each molecular typing record was assigned to a breakdown based on date and herd identifier, and the dataset was downsampled to include only the first typing record for each breakdown, as multiple isolates per breakdown were only routinely typed after 2008. A total of 10049 out of 17484 breakdowns (57.5%) for this period were linked to at least one molecular typing record (Table 2.1), and 18699 out of 20322 molecular typing records (92%) were successfully assigned to a breakdown (2656 typed breakdowns had more than one molecular typing result recorded). Sixteen unique spoligotypes (Table 2.2) and 183 unique VNTR-types (Table 2.3) were recorded over the study period.

### 2.3.2 Null hypothesis - Neutral ecological drift

The R package untb (Hankin, 2007) was used to test whether the observed RAD of *M. bovis* molecular types conformed to predictions generated by NTB. This package

**Table 2.2: Number of breakdowns recorded for each NI spoligotype, 2003-2010**

| Spoligotype | Number of breakdowns |
|---|---|
| SB0054 | 12 |
| SB0129 | 44 |
| SB0130 | 46 |
| SB0131 | 16 |
| SB0140 | 5684 |
| SB0142 | 2654 |
| SB0145 | 297 |
| SB0146 | 1 |
| SB0263 | 923 |
| SB0273 | 348 |
| SB0290 | 3 |
| SB0484 | 3 |
| SB0666 | 7 |
| SB0668 | 3 |
| SB0978 | 7 |
| SB1076 | 1 |

applies the theory put forward in Hubbell (2001) to generate stochastic realisations of the predicted RAD for a set group of species. It does this by first calculating the fundamental biodiversity number $\theta$, which is estimated from the number of species present in the community $S$, using statistical inference based on Ewens sampling formula (Ewens, 1972; Hankin, 2007). Theta a composite parameter discussed extensively by Hubbell, and is related to the total number of individuals in present in the community ($J_M$) and the speciation rate ($\nu$) by $\theta = J_M \frac{\nu}{1-\nu}$ (Rosindell *et al.*, 2011). When $\theta$ is combined with the total number of individuals present in the community ($J_M$), it allows stochastic generation of the RADs that would be expected in the community under neutral theory, in the absence of dispersal limitation, based on the algorithm outlined in Hubbell (2001, p289). As these predicted RADs are calculated without reference to the relative abundances observed in the community under study, comparison between these predictions and the observed RADs provide a means to test whether the observed abundances deviate from that expected under NTB (Rosindell *et al.*, 2011).

**Table 2.3: Number of breakdowns recorded for each NI VNTR-type, 2003-2010**

| VNTR-type | No. breakdowns | VNTR-type | No. breakdowns | VNTR-type | No. breakdowns |
|---|---|---|---|---|---|
| 1 | 1741 | 86 | 1 | 211 | 6 |
| 2 | 2281 | 87 | 26 | 215 | 1 |
| 3 | 567 | 91 | 3 | 217 | 1 |
| 4 | 628 | 92 | 7 | 218 | 1 |
| 5 | 1071 | 93 | 1 | 221 | 1 |
| 6 | 843 | 97 | 1 | 226 | 1 |
| 7 | 563 | 98 | 1 | 232 | 2 |
| 8 | 108 | 99 | 3 | 237 | 3 |
| 9 | 264 | 100 | 3 | 238 | 1 |
| 10 | 48 | 103 | 19 | 240 | 2 |
| 11 | 294 | 104 | 6 | 241 | 1 |
| 12 | 10 | 105 | 3 | 243 | 1 |
| 13 | 104 | 106 | 8 | 251 | 1 |
| 14 | 51 | 107 | 5 | 257 | 2 |
| 15 | 78 | 108 | 1 | 258 | 1 |
| 16 | 68 | 109 | 3 | 259 | 1 |
| 17 | 75 | 110 | 2 | 260 | 1 |
| 18 | 42 | 111 | 2 | 261 | 1 |
| 19 | 88 | 112 | 9 | 262 | 1 |
| 20 | 16 | 113 | 20 | 263 | 1 |
| 21 | 9 | 115 | 2 | 266 | 3 |
| 22 | 9 | 117 | 108 | 272 | 1 |
| 23 | 74 | 118 | 5 | 279 | 1 |
| 24 | 41 | 119 | 3 | 280 | 1 |
| 25 | 74 | 120 | 7 | 281 | 1 |
| 26 | 2 | 122 | 54 | 283 | 3 |
| 27 | 31 | 123 | 1 | 284 | 1 |
| 28 | 7 | 124 | 1 | 288 | 1 |
| 29 | 8 | 126 | 1 | 289 | 1 |
| 31 | 6 | 129 | 3 | 292 | 1 |
| 32 | 4 | 131 | 1 | 296 | 5 |
| 33 | 5 | 132 | 1 | 297 | 5 |
| 34 | 2 | 141 | 3 | 310 | 1 |
| 36 | 3 | 142 | 1 | 322 | 1 |
| 37 | 3 | 143 | 2 | 330 | 1 |
| 40 | 1 | 144 | 6 | 341 | 1 |
| 41 | 1 | 145 | 2 | 348 | 2 |
| 42 | 120 | 146 | 2 | 352 | 1 |
| 43 | 1 | 147 | 6 | 357 | 1 |
| 44 | 9 | 149 | 1 | 358 | 1 |
| 45 | 7 | 150 | 6 | 364 | 1 |
| 49 | 88 | 151 | 2 | 379 | 2 |
| 52 | 8 | 152 | 1 | 385 | 1 |
| 53 | 18 | 155 | 2 | 386 | 2 |
| 54 | 1 | 159 | 1 | 387 | 1 |
| 55 | 1 | 161 | 1 | 408 | 1 |
| 56 | 10 | 162 | 1 | 423 | 2 |
| 58 | 2 | 163 | 1 | 426 | 1 |
| 60 | 3 | 164 | 1 | 432 | 1 |
| 61 | 2 | 173 | 1 | 438 | 1 |
| 62 | 3 | 183 | 1 | 950 | 1 |
| 64 | 1 | 192 | 1 | 953 | 2 |
| 67 | 1 | 193 | 1 | 960 | 1 |
| 68 | 7 | 194 | 1 | 962 | 1 |
| 70 | 3 | 195 | 1 | 965 | 1 |
| 72 | 45 | 201 | 11 | 966 | 1 |
| 73 | 19 | 202 | 1 | 974 | 1 |
| 79 | 1 | 204 | 1 | 975 | 1 |
| 83 | 1 | 205 | 4 | 979 | 1 |
| 84 | 2 | 206 | 9 | 987 | 1 |
| 85 | 3 | 208 | 1 | 999 | 3 |

For this analysis, different molecular types of *M. bovis* were treated as different species, and the number of herd breakdowns of each molecular type in NI was taken to represent the number of individuals per species. Molecular types were defined using either spoligotyping (spoligotypes), or using a combination of spoligotyping and VNTR-typing (VNTR-types). Using the untb package, the fundamental biodiversity number $\theta$ was first estimated from the observed RAD, and using this value and the total number of typed herd breakdowns (n=10049), 1000 stochastic realisations of the RADs predicted under NTB were generated. The 95% interval of these distributions was then plotted to compare the observed RAD with NTB predictions for the community.

The above steps were carried out for the RADs of: all recorded NI *M. bovis* spoligotypes across the whole study period (2003-2010); all recorded NI VNTR-types across the entire study period; and NI VNTR-types subdivided into those occurring in each year of the study period to identify whether there were differences between RADs over the study period, and to what extent this subdivision affected the fit to neutral predictions. Additionally, this process (estimation of $\theta$ and generation of expected RADs) was also carried out for the RADs reported in Smith *et al.* (2003) (the abundances of molecular types of *M. bovis* in cattle in GB), and in Ypma *et al.* (2013) (the abundances of molecular types of human tuberculosis in the Netherlands).

### 2.3.3 Alternative hypotheses - Basic simulation model structure

To test the alternative hypotheses regarding the processes involved in shaping the distribution of VNTR-types in NI, four different stochastic simulation models were constructed. Specific details of each model are given in the sections below, but I describe the basic model structure here as it is comparable across all subsequent models. Within the model, a single individual represents a single (VNTR-typed) herd breakdown of bTB, and species equates to VNTR-type. A speciation event represents a mutation in VNTR-type, and birth and death events are the start and end of herd breakdowns respectively, with each new breakdown infected from a parent individual. The population size in the model is the equivalent of the number of (VNTR-typed)

bTB breakdowns on herds at any one time, and the birth rate is the equivalent of the herd-level incidence rate.

The species present at time $t$ were represented as the vector $i = 1 \ldots S_t$, and the relative abundance of different species as the vector $j = 1 \ldots n_{i,t}$, with total population size at time $t$ $J_{M,t} = \sum_{i=1}^{S_t} n_{i,t}$. Simulations were started with the individuals in the population at one of two extremes to enable checking of the model results for convergence: either all individuals started as the same species, with $S_0 = 1$ and $n_{i,0} = J_{M,0}$ ; or every individual started as a different species, with $S_0 = J_{M,0}$ and $n_{i,0} = 1$.

An overview of the steps conducted in one timestep of the basic model is given in Figure 2.1. Firstly, a vector of individuals, $\bar{n}_t^d$, representing the number of individuals in each species that die in the current timestep, is identified (box 1, Figure 2.1) and removed from the population (box 2, Figure 2.1). The number of individuals to be removed is given by the death rate $d$ multiplied by the total population size $J_{M,t}$ , and individuals in each species are chosen to be removed from the model with probability proportional to the frequency of the species in the population (box 1). A vector representing the number of new births in each species at this timestep, $\bar{B}_t$, is then identified, with the total number of individuals to be born given by the birth rate $b$ multiplied by total population size $J_{M,t}$, and the probability of a species reproducing in the basic model again proportional to the frequency of that species in the population (box 3). Each of the newly born individuals then has the option to convert to a new species with speciation probability $\nu$ (boxes 4 and 5). If this occurs, a new species containing one individual is added to the population (box 6); otherwise the original species is updated to add another individual (box 7). The process is then repeated with the updated population vectors as input. To ensure convergence, for each model the simulations were run until the two starting population conditions had converged on the same species abundance distribution (allowing for stochastic variation).

The simulation model as described above generates the same output as the algorithm implemented in the untb package (see previous section; Hankin, 2007; Hubbell, 2001,

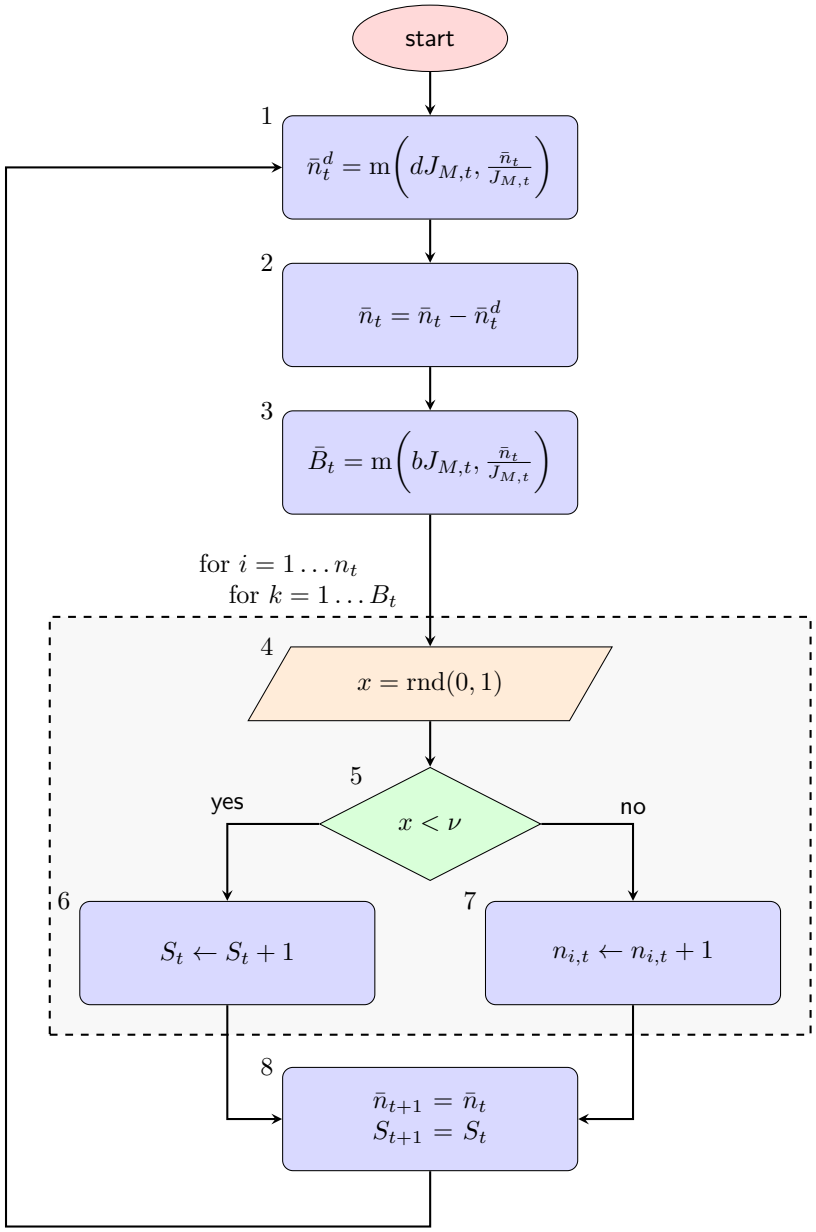


**Figure 2.1: Flowchart to show the steps carried out for the basic model, which simulates the evolution of a community of individuals of different species undergoing neutral ecological drift** - The vector of "species" in the simulated community at time $t$ is $i = 1 \ldots S_t$, and the number of individuals in each species $i$ at time $t$ is $j = 1 \ldots n_{i,t}$. $\bar{n}_t^d$ is the vector of individuals that are removed from each of the species in the model at time $t$ (boxes 1 and 2). $\bar{B}_t$ (box3) is the vector of individuals in each species chosen to reproduce at time $t$. $\nu$ is the probability of speciation occurring per birth event (box 5). Total population size at time $t$ (boxes 1 and 3) is given by: $J_{M,t} = \sum_{i=1}^{S_t} n_{i,t}$. Function $\mathrm{m}(x, y)$ (boxes 1 and 3) generates a multinomially distributed random number vector of size $x$, where $y$ is a vector giving the probability of drawing the individuals from each class . Function $\mathrm{rnd}(0, 1)$ (box 5) generates a random variable distributed evenly between 0 and 1. The dashed rectangle indicates steps conducted while looping through each birth $k$ in each species $i$.

p289). This provides a basic framework that can be altered to reproduce the mechanisms outlined in Hypotheses 1-4 (VNTR-type speciation/selection, wildlife reservoir, increasing prevalence, and superspreading), and to assess to what extent they are able to replicate the manner in which the observed distribution of NI VNTR-types diverges from NTB predictions. The individual models themselves are described in the following section. On completion of a simulation the value of the biodiversity number $\theta$ was estimated for each model run, and in order to summarise the runs for each model and to assess their divergence from NTB, these values of $\theta$ were then averaged across all runs for that specific model. Predictions of the RAD expected under NTB were then generated in the untb package as above using this average value of $\theta$, and the 95% interval of the NTB predictions was then plotted and compared to the 95% interval of the simulation results.

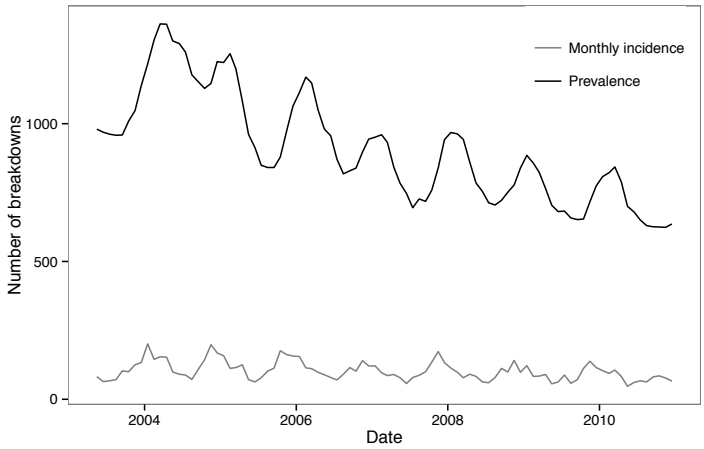


**Figure 2.2: Prevalence (black) and monthly incidence rate (grey) of VNTR-typed herd breakdowns in NI over the study period**

The parameters for the baseline model were chosen to approximate the situation in NI, with each timestep approximately representing one week. Using data on the numbers of typed breakdowns over the study timeframe (Figures 2.2 and 2.3), the number of deaths, $dJ_{M,t}$, and the number of births, $bJ_{M,t}$, per timestep were both set to 24 individuals (equivalent to the mean herd-level incidence of VNTR-typed breakdowns per week, Figure 2.2 grey line); the population size $J_{M,t}$ was set to 890 individuals (equivalent to the mean herd-level prevalence, Figure 2.2 black line); and the speciation probability $\nu$ was set as 0.015 (equivalent to the mean proportion of new VNTR-typed

breakdowns attributable to novel VNTR-types, Figure 2.3).

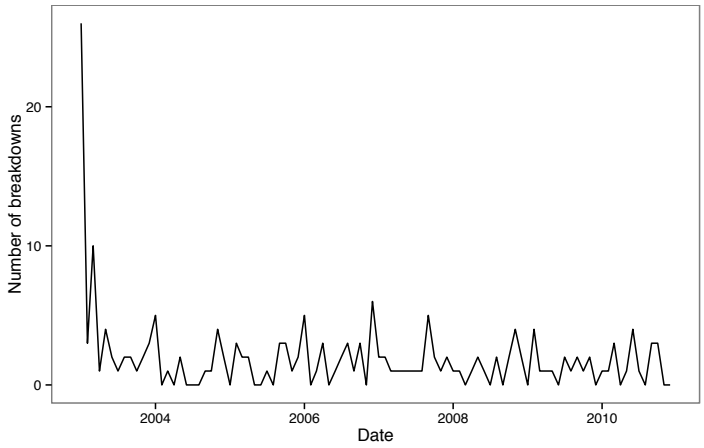


**Figure 2.3: Number of breakdowns attributed to novel VNTR-types per month in NI over the study period**

The model for Hypothesis 1 (VNTR-type speciation and/or selection) assumes a stepwise mutation process, but none of the other models described below discriminate between new species occurring through mutation or through immigration from outside the simulated community. For all models with the exception of Hypothesis 4 (superspreading), the probability of reproduction for each species is directly proportional to the frequency of the species in the population (box 4 Figure 2.1). Stochastic simulations representing each of the four hypotheses were run 500 times (250 runs starting with all individuals of the same species, and 250 runs starting with all individuals as different species), and the results were summarised to give output equivalent to the incidence of (VNTR-typed) breakdowns over the course of a year (timesteps corresponding to 52 weeks).

### 2.3.4 Hypothesis 1 - VNTR-type speciation and/or selection of VNTR-loci

Mutation in VNTR-type occurs through variation in the number of repeats present at the VNTR loci. This process can give rise to homoplasies, with unrelated lineages converging on the same VNTR-type (Reyes *et al.*, 2012), and this could act to augment the more common VNTR-types. To test whether homoplasy in VNTR-types and/or

the method by which VNTR-typing loci were selected in NI would affect the RAD of different VNTR-types, the basic model above was altered so that species was denoted by a string of twelve integers. The twelve integers represent the number of tandem repeats present at the twelve candidate VNTR-loci from which the current NI typing loci were originally selected (Skuce *et al.*, 2005). In simulations that were started with all individuals in the same species, all twelve loci were set to 5 for all individuals, while in simulations that were started with all individuals in different species, each of the twelve loci for each individual was chosen as a randomly sampled integer between 1 and 10. When a speciation event occurred in the model (box 6 Figure 2.1), this generated a stepwise mutation in VNTR-type, with one locus chosen at random from the string of twelve, and the number of repeats at this locus then increased or decreased by one.

For these simulations, the speciation rate $\nu$ was increased to 0.0257 per birth event to reflect the increased number of loci included in the simulation, as this would be expected to result in an increased rate of speciation. All other parameters were as described above.

On completion of each simulation, to mimic the manner in which the VNTR-typing loci were chosen in NI (Skuce *et al.*, 2005), the seven most diverse loci from each model run were chosen using the Hunter-Gaston index (Hunter & Gaston, 1988). The Hunter-Gaston index is a measure of the probability that two individuals sampled at random are of the same type, assuming sampling without replacement, and is a commonly used measure of diversity in microbiology, and was used in selecting the VNTR-typing loci in NI. Following identification of the most diverse loci, the species designation of each individual in the model output was then re-named based solely on these seven loci.

### 2.3.5 Hypothesis 2 - Unsampled reservoir of infection

Hypothesis 2 posited that an unsampled reservoir of *M. bovis* infection, such as infection in a wildlife population (as implicated in the epidemiology of bTB in GB), might account for the observed RAD of VNTR-types in NI. To test this, we extended

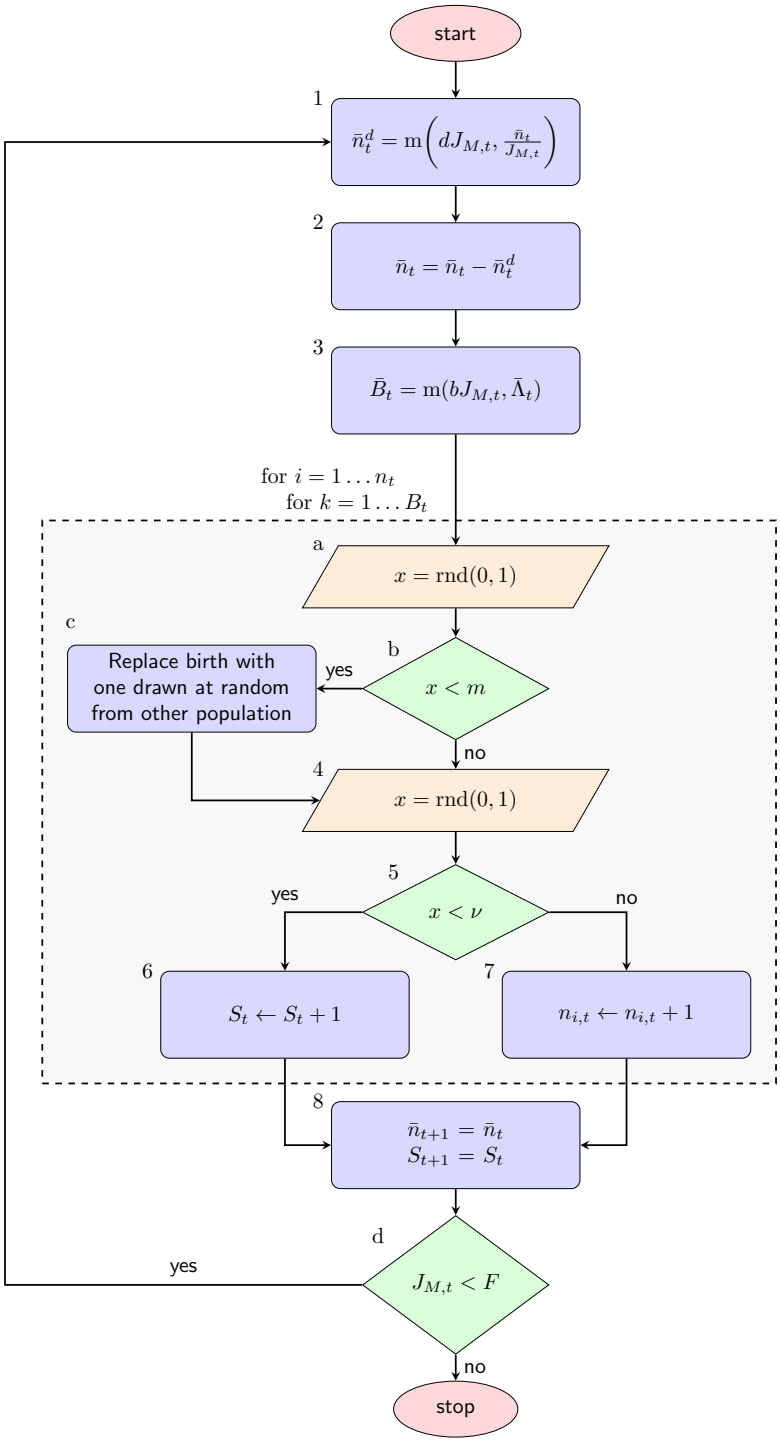


**Figure 2.4: Flowchart based on Figure 2.1 (basic model), showing the modifications made to simulate Hypotheses 2-4** - Boxes 1-8 are similar to Figure 2.1. For Hypothesis 2 (unsampled reservoir), parameter $m$ (box b) represents the likelihood of an individual migrating into the system from the separately modelled population (box c). For Hypothesis 3 (increasing population size), parameter $F$ (box d) represents the final target population size for the model. For Hypothesis 4 (superspreading), $\bar{\Lambda}_t$ (box 4) gives the vector of the probability of reproduction for each species, which is the sum of the reproduction probabilities $\lambda$, for each individual $j$, in species $i$, divided by the sum of reproduction probabilities for all individuals in the system.

the baseline model to consider two separate populations of *M. bovis*, linked through immigration of individuals between them (Figure 2.4, boxes b and c).

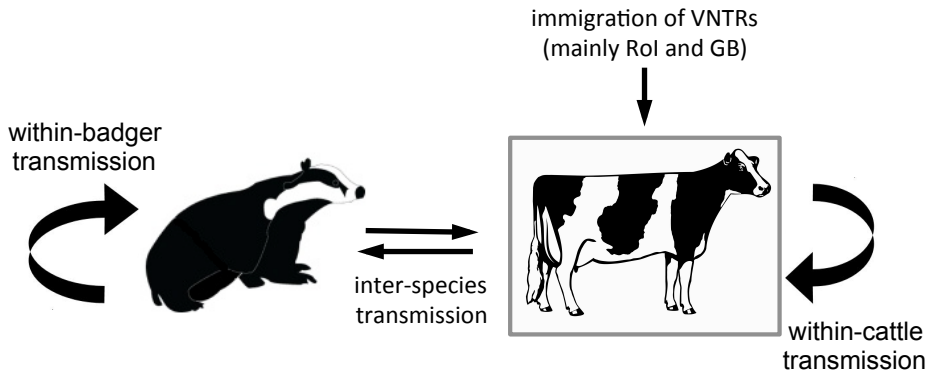


**Figure 2.5: Summary of the linked cattle-badger populations and the source of new VNTR-types in each**

The VNTR-type abundances reported here were measured solely in the observed cattle population. In infected badgers, novel VNTR-types emerge through speciation from types circulating in badgers or through import of strains from the cattle population, whereas in cattle new VNTR-types can occur through speciation from circulating cattle strains, import from the badger population, and additionally through import of VNTR-types circulating in cattle in other countries (see Figure 2.5). Therefore the rate of occurrence of new VNTR-types is likely to be higher in cattle than in badgers, and this difference in the occurrence of novel VNTR-types may affect the RAD observed. In this model we make no distinction between novel species occurring as a result of speciation or immigration, both being combined within the speciation rate parameter $\nu$. To mimic the effects of differing rates of the occurrence of novel VNTR-types between badgers and cattle, different values of speciation rates ($\nu$) were investigated for the target population (representing infections in cattle) and the linked population (representing infection in badgers), and values of migration ($m$) between the two populations were also varied. Migration rates were tested for values of 0.0208 (representing approximately one migration per timestep) and 0.0833 (approximately two migrations per timestep). Table 2.4 shows the combinations of these parameters that were investigated for these simulations.

**Table 2.4: Combinations of parameter values used in simulations for the model representing Hypothesis 2 (unobserved reservoir of infection)** - Speciation rates for the target population ($\nu_t$) and linked population ($\nu_t$) were varied, as were migration rates from target population to linked population ($m_{t,l}$), and from linked population to target population ($m_{l,t}$). Six combinations of these values were tested (labelled A-F)

| | $\nu_t$ | $\nu_l$ | $m_{t,l}$ | $m_{l,t}$ |
|---|---|---|---|---|
| A | 0.01 | 0.01 | 0.0208 | 0.0208 |
| B | 0.02 | 0.01 | 0.0208 | 0.0208 |
| C | 0.1 | 0.01 | 0.0208 | 0.0208 |
| D | 0.01 | 0.01 | 0.0208 | 0.0833 |
| E | 0.02 | 0.01 | 0.0208 | 0.0833 |
| F | 0.02 | 0.01 | 0.0833 | 0.0208 |

### 2.3.6 Hypothesis 3 - Historical increases in population size

To investigate the effects of historical bTB increases on the RAD of VNTR-types, the basic model was altered to approximate the increase in bTB prevalence that has affected NI in the decades up to 2002 (Abernethy *et al.*, 2006).

Based on data from Abernethy *et al.* (2013), peak bTB levels in NI were estimated at approximately 3200 herd breakdowns in 2002. To keep the results of this model proportional to the previous simulations I took 57.5% of this value (the proportion of NI herd breakdowns linked to at least one molecular typing record in the 2003-2010 data), giving 1840 breakdowns as the final (maximum) population size for the increasing population size model ($F$, box d Figure 2.4). Data available from www.bovinetb.info (downloaded 17th September 2014) give the historic data on numbers of cattle testing positive to bTB and are summarised in Figure 2.6. Although these figures do not include bTB infections identified at slaughter, if we take them to broadly represent historical trends this would indicate that the increase in bTB in NI started in 1986, at which point bTB levels were approximately 9.3% of their maximum.

The basic constant-size, simulation model described above was first run at the starting population size $J_{M,0} = 171$ for 100,000 timesteps to allow equilibration of the simulated RAD. A model simulating increasing population size was then run starting from this

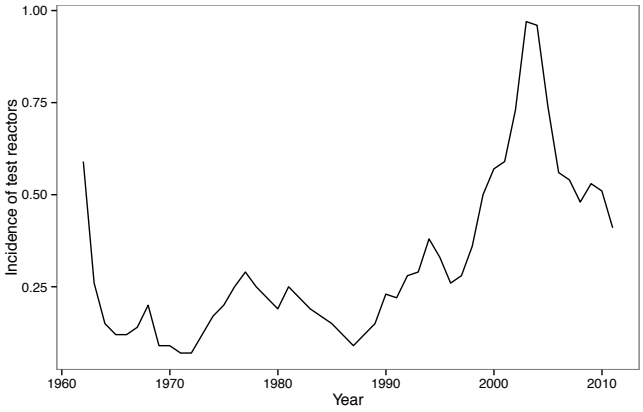


**Figure 2.6: Historical incidence of cattle testing positive for bTB in NI**

population, until the target end population size ($F$, box d Figure 2.4) was reached. To generate an increasing population size approximately equivalent to the historical situation in NI, the model was run with death rate $d = 0.02697$ (box 1 Figure 2.4) and birth rate $b = 0.02883$ (box 3, Figure 2.4).

### 2.3.7 Hypothesis 4 - Superspreading

To assess the effect of superspreading on the RAD of NI VNTR-types, the basic model was extended to incorporate transmission heterogeneity by introducing systematic variation in the number of offspring per individual.

At the point at which new births were identified in the model (box 3, Figure 2.4), each new birth was also assigned an individual reproduction number ($\lambda$), which was then retained throughout the lifespan of that individual. This individual reproduction number was drawn from a gamma distribution with a mean of 1 (equal to the effective reproduction rate in a population with a constant prevalence) and with shape parameter $k$, after Lloyd-Smith *et al.* (2005). Simulations were run with $k$ values of 0.1, 0.5, 1, and 10, to assess the effect of different degrees of variation in transmission on model outputs. As before, a set number of individuals died ($dN_t = 24$) and were born ($bN_t = 24$) at each timestep, however, in this model the chance of an individual being chosen to reproduce at each timestep was directly proportional to its pre-assigned re-

production probability $\lambda$, and therefore $\Lambda_{i,t}$, the probability of species $i$ reproducing at each timestep $t$, is given by the sum of the reproduction numbers $\lambda$ for each individual $j$ in species $i$, divided by the sum of the reproduction probabilities for all individuals in the population:

$$\Lambda_{i,t} = \frac{\sum_{j=1}^{N_{i,t}} \lambda_{i,j}}{\sum_{i=1}^{n_t} \sum_{j=1}^{N_{i,t}} \lambda_{i,j}}$$

The number of offspring generated by a single individual over the course of its lifespan is related to the reproduction number of the individual per timestep as well as the individual's lifespan (as death rate $d$ is constant, the latter will follow a geometric distribution).

## 2.4 Results and Discussion

### 2.4.1 Null hypothesis - Neutral ecological drift

The null model of neutral ecological drift described by NTB (Hubbell, 2001; Rosindell *et al.*, 2011) was unable to explain the observed relative abundances of *M. bovis* VNTR-types recorded in NI, whether considering the distribution across the whole study period (Figure 2.7A), or when divided into separate years (Figure 2.7B). There are no obvious differences in the VNTR-type RADs observed over different years of the study period (Figure 2.7B). In the observed VNTR-type distributions, the common types appear more dominant than would be expected under NTB, with a tail of VNTR-types that are rarer than expected. The RAD of *M. bovis* spoligotypes in NI does not fall outside the 95% envelope of the NTB predictions (Figure 2.7C), however it is possible there is a slight concavity in the shape of the curve, which may suggest a tendency towards the same pattern as seen in the NI VNTR-types. However, with lower numbers of spoligotypes than VNTR-types, the predicted neutral distributions show greater variability, and therefore the power to detect deviations from the neutral expectations

is reduced. Further statistical comparison between the observed spoligotype RADs and neutral predictions, for example using the methodology described in Bersier & Sugihara (1997), Mouillot *et al.* (2003) and Spatharis *et al.* (2009) would be a useful next step in analysing these results. This approach has been applied to compare results from simulation models to replicate observations of RADs found in ecological systems. The mean and variance of numbers of individuals present at each abundance rank is calculated for the observed distributions, and is compared to the mean and variance of the model outputs for that rank. This is then then summarised over all ranks to generate a summary statistic, which is compared to summary statistics generated by conducting the same procedure to compare each of the model runs in turn to the distribution of model results.

The RAD of different molecular types of human tuberculosis in the Netherlands described by Ypma *et al.* (2013) shows a similar pattern to the NI VNTR-types, and appears to differ significantly from neutral predictions (Figure 2.7D). Interestingly the RADs of *M. bovis* molecular types in GB reported in Smith *et al.* (2003) do not appear to diverge significantly from predictions (Figure 2.7E and F), suggesting that neutral ecological drift may be sufficient to drive the diversity of *M. bovis* present in this system, although this may again be a result of limited power to detect a difference given the small number of molecular types, as discussed above.

### 2.4.2 Hypothesis 1 - VNTR-type speciation and/or selection of VNTR-loci

The process through which VNTR-types are identified and new types arise is based on variation in the number of repeats present at different VNTR loci. This gives rise to the possibility of unrelated lineages converging on the same VNTR-type (Reyes *et al.*, 2012), which could act to augment the more common VNTR-types. Additionally, the VNTR-loci used for VNTR-typing in NI were originally chosen from a panel of twelve candidate loci to optimally discriminate the *M. bovis* population in NI (Skuce *et al.*, 2005). It was unclear whether either or both of these processes would affect the RAD

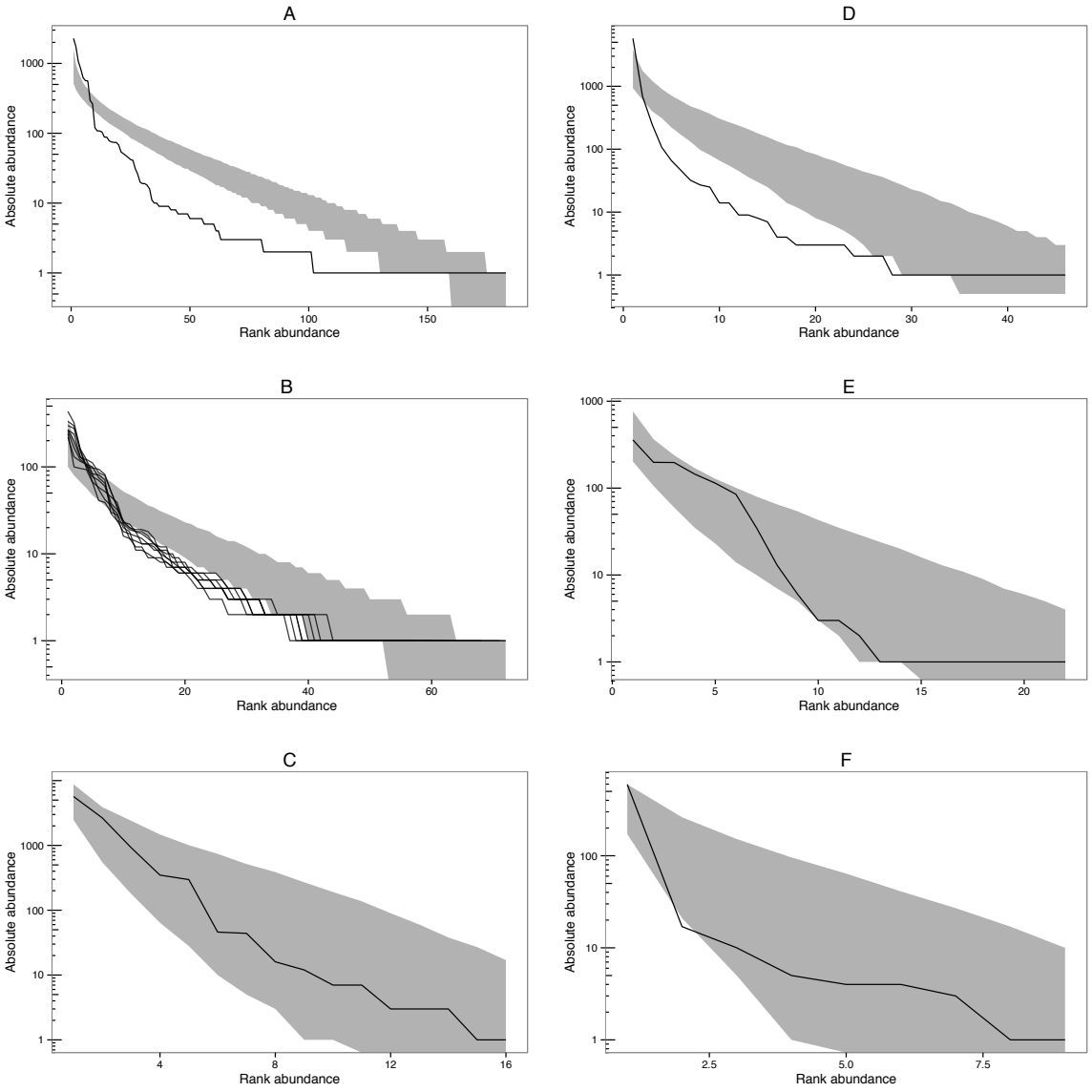


**Figure 2.7: Comparison of observed relative abundance distributions (lines) and 95% envelopes for neutral predictions (shading) for observed distributions of molecular types** - for NI *M. bovis* VNTR-types aggregated across the study period (A); NI *M. bovis* VNTR-types separated by year (B); NI *M. bovis* spoligotypes aggregated across the study period (C); spoligotypes and RFLP types of human TB in the Netherlands (D) as described in Ypma *et al.* (2013); and VNTR-types within GB *M. bovis* spoligotypes SB0140 (E) and SB0263 (F), as described in Smith *et al.* (2003). Graphs show log-scaled absolute abundance of each molecular type on the y-axis, compared to the ranked abundance of each type on the x-axis (where the most common types have a rank of 1, and increasing ranks indicate less abundant types).

of VNTR-types and their fit to neutral predictions. The results of the simulations replicating these processes show minimal differences from the distribution predicted under NTB (Figure 2.8). This indicates that the mode of VNTR-type speciation and the manner in which the panel of VNTR-typing loci were chosen in NI cannot account for the shape of the observed RAD seen in NI VNTR-types.

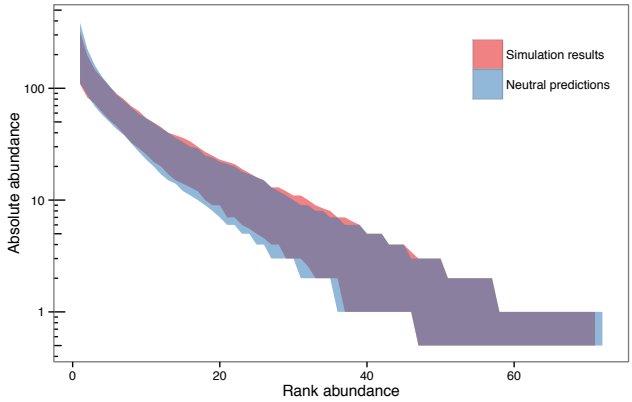


**Figure 2.8: Comparison between Hypothesis 1 simulations (VNTR-type speciation and selection) and neutral predictions.** - The 95% envelope for the relative abundance distributions generated by 500 simulations is shown in red, and the 95% envelope for predictions under neutral theory in blue. Log-scaled absolute abundance is shown on the y-axis and ranked abundance on the x-axis (species ranked in order of decreasing abundance).

### 2.4.3 Hypothesis 2 - Unsampled reservoir of infection

All molecular typing results described in this study originated from cattle infections in NI, and only more limited information is available on possible undersampled reservoirs such as the *M. bovis* types infecting the badger population in NI. The results of the model simulating an unsampled reservoir, which simulated two separate pools of infection linked through transmission of infection between the two populations, suggested that even if the rate of occurrence of novel VNTR-types differs between the two host populations, this cannot account for the observed distribution of *M. bovis* VNTR-types in NI (Figure 2.9). In fact, this simple metacommunity structure actually generates RADs that appear slightly more even than that predicted by NTB for all combinations of migration and speciation probabilities explored here.

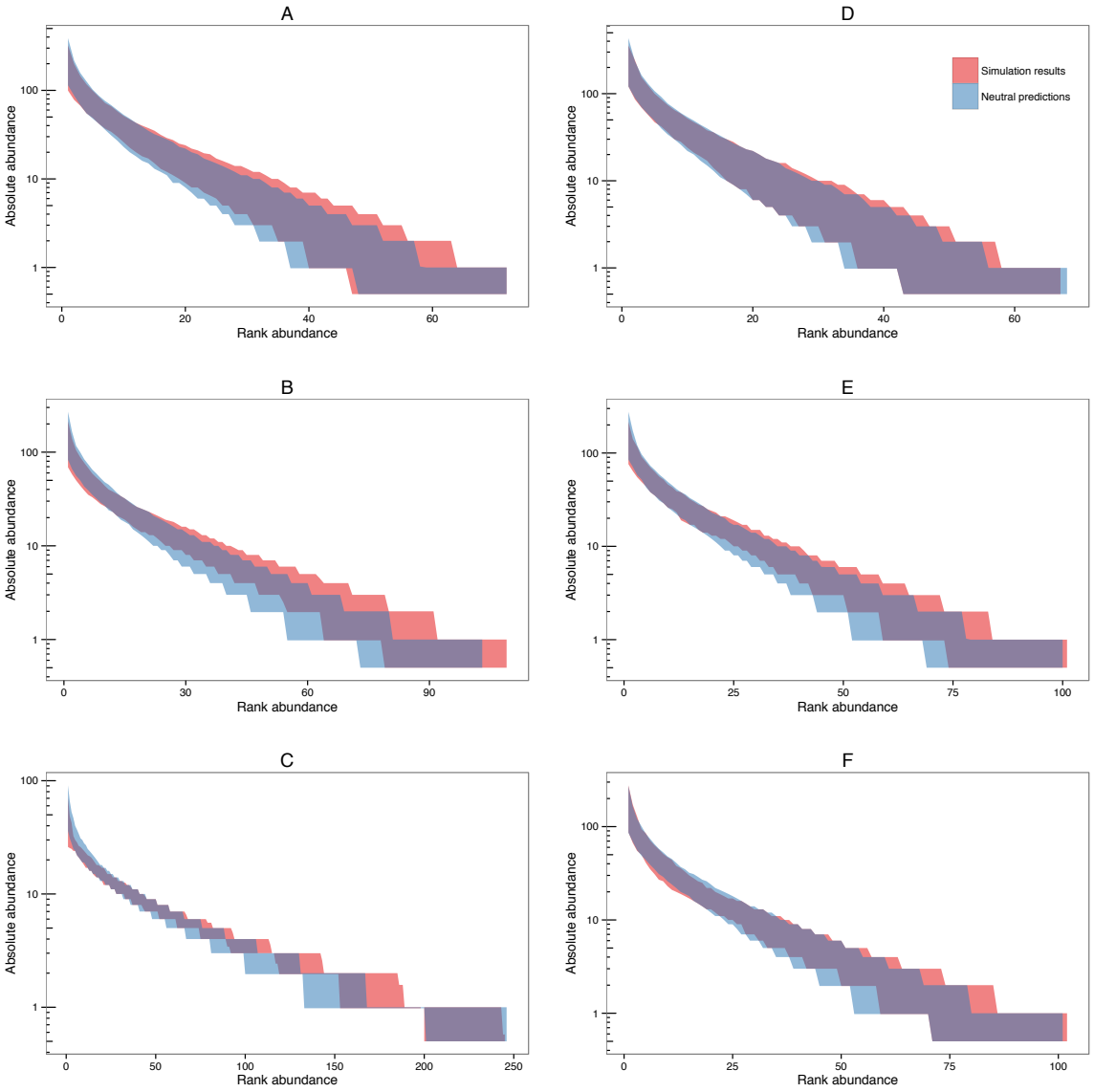


**Figure 2.9: Comparison between Hypothesis 2 simulations (unobserved reservoir) and neutral predictions, plotted for the target (cattle) population** - Results are shown for: a model with speciation probability per birth in the target (cattle) population, $\nu_{target}$, of 0.01 and speciation probability per birth in the linked (badger) population, $\nu_{linked}$, of 0.01. Migration probabilities ($m_{l,t}$ and $m_{t,l}$) were 0.0208 (approx. one migration every two timesteps) in both directions (A); a model with $\nu_{target} = 0.02$; $\nu_{linked} = 0.01$; $m_{l,t} = 0.0208$; $m_{t,l} = 0.0208$ (B); a model with $\nu_{target} = 0.1$; $\nu_{linked} = 0.01$; and $m_{l,t} = m_{t,l} = 0.0208$ (C); a model with $\nu_{target} = \nu_{linked} = 0.01$; and $m_{l,t} = 0.0208$; and $m_{t,l} = 0.0833$ (approx two migrations per timestep) (D); a model with $\nu_{target} = 0.02$; $\nu_{linked} = 0.01$; $m_{l,t} = 0.0208$; and $m_{t,l} = 0.0833$ (E); and for model with $\nu_{target} = 0.02$; $\nu_{linked} = 0.01$; $m_{l,t} = 0.0833$; and $m_{t,l} = 0.0208$ (F).
The 95% envelope for the relative abundance distributions generated by 500 simulations is shown in red, and the 95% envelope for predictions under neutral theory in blue. Log-scaled absolute abundance is shown on the y-axis and ranked abundance on the y-axis (species ranked in order of decreasing abundance).

This type of structuring of the community therefore appears to generate RADs opposite to the skewed distribution observed in the VNTR-types in NI. Similar findings have also been reported when investigating the effect of spatial structuring of the community through spatially limited dispersal of individuals (Hubbell, 2001, Chapter 6), which was shown to also generate more even distributions of species across the metacommunity than predicted by NTB in the absence of dispersal limitation. Spatial structuring of the *M. bovis* population in NI is evident in the geographical clustering of infection by molecular type (Skuce *et al.*, 2010). I would therefore suggest that a reservoir of infection or other forms of metacommunity structure, including the spatial structuring evident in NI, would in the absence of other factors be expected to draw the RAD in the opposite direction to that seen in the NI VNTR-type data, acting to even out the skewed distribution observed. However, the absence of this finding in the observed data is not sufficient to rule out metapopulation structure, a reservoir of infection in a different species, and/or other forms of population structuring in the epidemiology of bTB in NI. Rather it suggest that, if these factors do play a role in the disease in NI (and as discussed above, it seems very likely that some or all of these are involved in bTB in the British Isles), other processes which push the RAD in the opposite direction must be present and have a strong enough effect to overcome the levelling influence of population structure and generate the skewed RADs observed in these data.

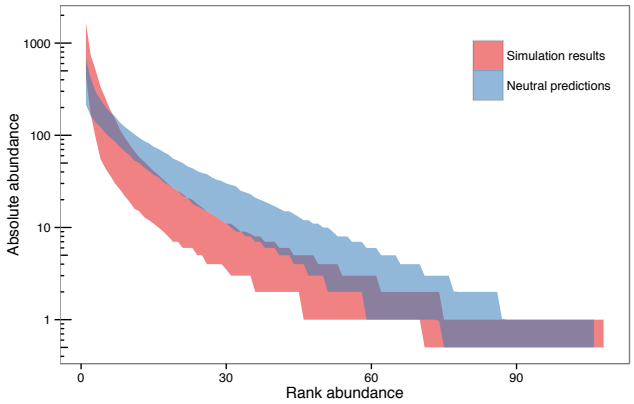


**Figure 2.10: Comparison between Hypothesis 3 simulations (increasing population size) and neutral predictions** - The 95% envelope for the relative abundance distributions generated by 500 simulations is shown in red, and the 95% envelope for predictions under neutral theory in blue. Log-scaled absolute abundance is shown on the y-axis and ranked abundance on the x-axis (species ranked in order of decreasing abundance).

### 2.4.4 Hypothesis 3 - Historical increases in population size

The available data on numbers of cattle testing positive for bTB in NI suggest the prevalence of the disease has increased from the late 1980s to 2002 (Figure 2.6, also Abernethy *et al.* (2006)). One of the key assumptions made by NTB is a constant size population (Hubbell, 2001), and Luciani *et al.* (2008) demonstrated that violating this assumption through an increasing population size could generate the patterns of relative abundances recorded in an outbreak of human TB in California.

The results of these simulations (Figure 2.10) agree with this, demonstrating that an increase in population size approximating historical increases in bTB prevalence in NI is capable of generating a RAD that deviates from NTB predictions in a manner similar to that shown by the observed distribution of NI VNTR-types (Figure 2.7A and B). The increase in simulated population size over time across each of the increasing population simulations is shown in Figure 2.11. However, since 2002 levels of bTB have been declining in NI (Figure 2.6), Abernethy *et al.*, 2006). Model simulations mimicking a declining population indicated that this scenario would have the opposite effect, generating more even RADs than would be expected under NTB (results not shown). Therefore the more recent decrease in bTB prevalence might act to dampen the skew in RAD generated by historical increases in bTB levels.

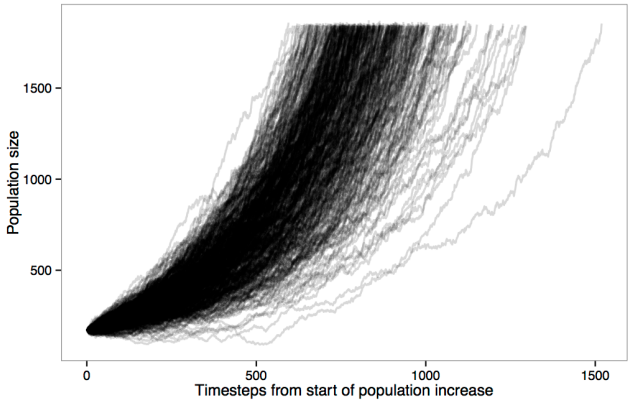


**Figure 2.11: Increases in population size over the course of the simulations of increasing population size** - One model timestep is approximately equivalent to one week.

### 2.4.5 Hypothesis 4 Superspreading

Variation in the number of secondary infections generated by infected individuals, or superspreading, is manifest in the epidemiology of a range of diseases (Lloyd-Smith *et al.*, 2005; Woolhouse *et al.*, 1997) and has been implicated in the spread of bTB in Britain, with a minority of infected cattle herds responsible for a large proportion of the onward transmission of the disease (Brooks-Pollock *et al.*, 2014). Ypma *et al.* (2013) have also demonstrated that this may account for skewed abundances of molecular types of human TB in the Netherlands.

Here, we simulated superspreading by assigning each individual in the model a probability of reproduction per unit time at birth, drawn from a gamma distribution. The results of simulations where individuals' chances of reproduction were distributed with mean of 1 and $k$ values of 1, 0.5, and 0.1 are shown in Figure 2.12A, B and C respectively, and the distribution of the number of offspring per individual over the course of the simulations is given in the inset graphs. These results indicate that superspreading at these levels could also feasibly generate divergences from NTB similar to that seen in the RAD of VNTR-types in NI (Figure 2.7A and B).

### 2.4.6 Conclusions

In this study, we show that neutral ecological drift in isolation cannot account for the observed distribution seen in NI VNTR-types (Figure 2 2.7A and B), although NTB predictions do not appear to differ significantly from the RAD recorded for *M. bovis* in GB (Smith *et al.*, 2003; Figure 2.7E and F), or that of spoligotypes in NI (Figure 2.7C).

Simulation models were then used to identify which, if any, of four specific hypotheses could reproduce the manner in which the observed RAD diverged from NTB predictions. In this analysis, the various alternate hypotheses represented in the simulation models are far from being mutually exclusive, and by contrast were specifically chosen

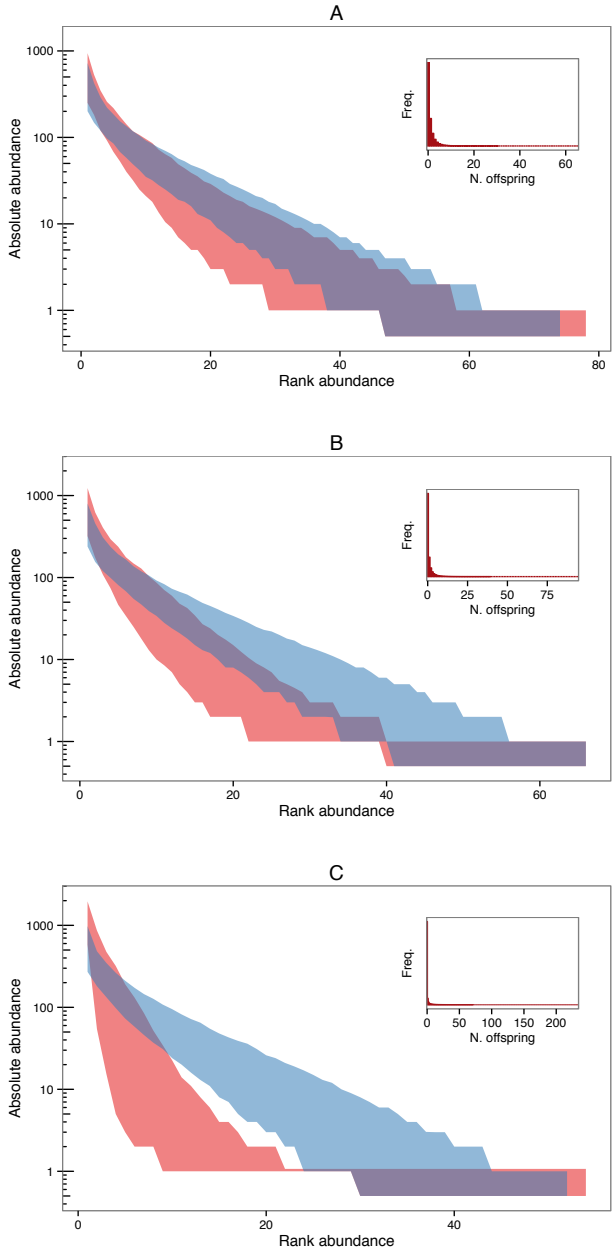


**Figure 2.12: Comparison between Hypothesis 4 simulations (superspreading) and neutral predictions for $k$ values of 1, 0.5, and 0.1 (A, B, and C respectively)** - The 95% envelope for the relative abundance distributions generated by 500 simulations is shown in red, and the 95% envelope for predictions under neutral theory in blue. Log-scaled absolute abundance is shown on the y-axis and ranked abundance on the x-axis (species ranked in order of decreasing abundance). Inset graphs show the number of individuals that give rise to different numbers of offspring, averaged over all simulations, for each value of $k$.

as they are all likely to be involved in the epidemiology and evolution of *M. bovis* in NI. I note that the aim of these simulations was to explore the shape of the distribution expected under the various hypotheses put forward, and compare that to the shape of the RAD expected under the null hypothesis of NTB and that observed in NI VNTR-types. Although the models were parameterised based on the available data on bTB in NI, no explicit model fitting was carried out, and therefore we do not directly compare them to the absolute numbers of breakdowns of each VNTR-type in NI.

I conclude here that VNTR-type homoplasy and the manner in which VNTR-typing loci were selected in NI, taken in isolation, have little impact on the expected abundances of different VNTR-types (Figure 2.8), while structuring of the community, for example due to a reservoir population (Figure 2.9), or through spatial processes such as spatially limited dispersal of individuals (Hubbell, 2001, Chapter 6), is likely to act to even out the RAD. However, in this system the opposite pattern was present and a very uneven distribution was observed, with a small number of VNTR-types present at very high frequency and the majority of types very rare. The results presented here for Hypotheses 3 and 4 indicate that some combination of historical increase in the prevalence of bTB in NI (Figure 2.10) and/or variation in the number of onward transmission events generated by each infected herd (Figure 2.12) are likely to be involved in maintaining the highly skewed RAD evident in NI.

# CHAPTER 3

## Can the spatial-genetic structure of *M. bovis* in Northern Ireland be explained by that of its wildlife host, the Eurasian badger?

# Can the spatial-genetic structure of *M. bovis* in Northern Ireland be explained by that of its wildlife host, the Eurasian badger?

## 3.1 Summary

Where spatial structure is present in a pathogen population, it can potentially give us information on the epidemiological mechanisms involved in the spread of the pathogen. In this chapter, I characterise the spatial structure of the *M. bovis* VNTR-types present in Northern Ireland (NI), showing that in NI as in other regions of Britain and Ireland, there is evidence of spatial clustering of the *M. bovis* population and associations between VNTR-types present in cattle and badgers. I go on to use landscape genetics approaches to evaluate possible correlates of the spatial structure of VNTR-types present in cattle, with the underlying hypothesis that transmission within the badger population is driving the spatial structure of *M. bovis* in cattle. I demonstrate that there is no evidence that major landscape features (elevation and major rivers) correlate with the edges of VNTR-type clusters, but that there is also no sign that these act as a barrier to the genetic mixing of the badger population. Only limited population genetic structure is detected in the badgers, but the observed boundaries between badger populations appear to correlate with the discontinuities in the VNTR-type communities. This may indicate that mixing within the badger population is be driving the spatial structure seen in *M. bovis* isolated from cattle in NI, however without investigating the structure of the NI cattle population it is not possible to rule

out an alternative explanation relating to spread via infected cattle. The approach described here could potentially be used in other multi-host pathogen systems.

## 3.2 Introduction

Infections involving multiple interacting host species are a common problem in disease ecology, and such systems can prove difficult to evaluate and to control (Viana *et al.*, 2014). The landscape genetic structure of a pathogen has the potential to tell us much about the underlying processes and hosts that shape its epidemiology (Biek & Real, 2010). In a multi-host system, each host species may affect the pathogen's genetic signature over differing temporal and spatial scales (Nadler, 1995), and using approaches that combine both pathogen and host genetics with other data may therefore allow us to tease apart the involvement of the various hosts in shaping the spatial structure of the pathogen.

Bovine tuberculosis (bTB) is a disease of cattle and other animals caused by the bacterium *M. bovis*, and is one of the most important diseases currently facing the livestock industry in Britain and Ireland. Despite many decades of bTB control programs in Britain and Ireland the disease persists, and direct transmission of bacteria within the cattle population still plays a role in its epidemiology (Gilbert *et al.*, 2005; Green *et al.*, 2008; Menzies & Neill, 2000). However, as mentioned in previous chapters, control of the disease in this region is further complicated by infection in a wildlife reservoir, the Eurasian badger *Meles meles* (Donnelly & Nouvellet, 2013; Donnelly *et al.*, 2006; Griffin *et al.*, 2005b).

For many years, molecular typing methods such as spoligotyping and VNTR-typing have been employed to sub-type *M. bovis* infections, and these molecular types have been demonstrated to group into geographic clusters in Great Britain (GB, Goodchild *et al.*, 2012; Smith *et al.*, 2006; Woodroffe *et al.*, 2005), NI (Skuce *et al.*, 2010) and the Republic of Ireland (RoI, Olea-Popelka *et al.*, 2005). Three of these studies also demonstrated the spatial co-occurrence of badger and cattle isolates of the same molec-

ular type (Goodchild *et al.*, 2012; Olea-Popelka *et al.*, 2005; Woodroffe *et al.*, 2005), indicating transmission of infection between the two species, but not the direction of this transmission.

Geographic clustering in the *M. bovis* population is likely to point towards the involvement of one or more of the following processes: a) a mechanism of transmissi[illegible]hi is constrained or facilitated by landscape features; b) transmission occurring through a spatially structured pattern of contacts within the host species, e.g. contact within and between badger social groups; and/or c) spread predominantly occurring through transmission over short distances, for example direct transmission between cattle on neighbouring herds or via flooding or local movement of fomites, which could allow spatial structure to arise through chance events alone (Irwin, 2002; Real & Biek, 2007). Hypotheses a) and b) are testable: in the case of a) one would expect the boundaries of pathogen genetic clusters to coincide with landscape features, while in the case of b) the boundaries of pathogen genetic clusters should reflect the genetic structure of the host population, if the contact patterns responsible for the spread of disease correlate with genetic mixing of the host population (plausible for a chronic disease such as bTB).

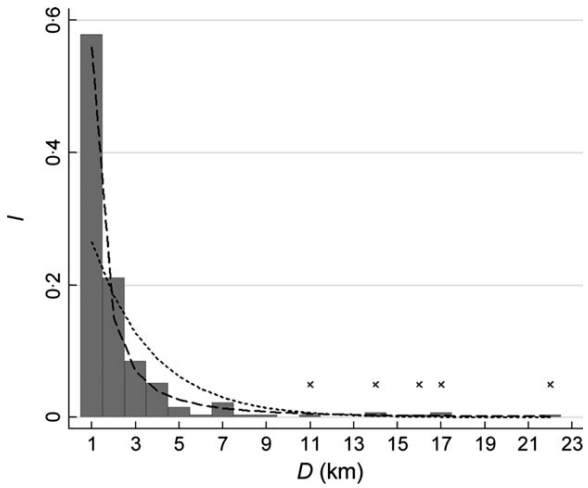


**Figure 3.1: Distribution of distances travelled by badgers in large scale study in the Republic of Ireland** - Predicted inverse power function (dashed line) and the predicted negative exponential function (dotted line) density distribution (I) in relation to the observed proportion of movements (D) over 1 km distance. Crosses denote bins with observed long-distance movements. Reproduced with permission from Byrne *et al.* (2014).

Badgers live in stable territorial groups in Britain and Ireland (Roper, 2010), and the majority of badger movements cover distances of under 5km (although longer distance dispersal events do occur less frequently: Byrne *et al.*, 2014; Pope *et al.*, 2006; Figure 3.1). By contrast, although many of the human-mediated movements of cattle between herds occur over relatively short distances (<40km), they still cover a much greater spatial scale than do badger movements (Mitchell *et al.*, 2005; Figure 3.2). These differences between the two host species lead to specific predictions regarding the spatial-genetic structure each host would be expected to contribute to the pathogen population. If transmission were mediated by infection in badgers in the absence of other influences, this would be expected to generate local spatial organisation reflecting the spatially localised contact and mixing patterns within the badger population. Local transmission between cattle on neighbouring herds would also be expected to result in spatial structure in the *M. bovis* population. By contrast, a predominance of transmission though movements of cattle between herds would be expected to generate more limited spatial structure, with the metapopulation or network organisation of the cattle industry determining the structure of the pathogen population.

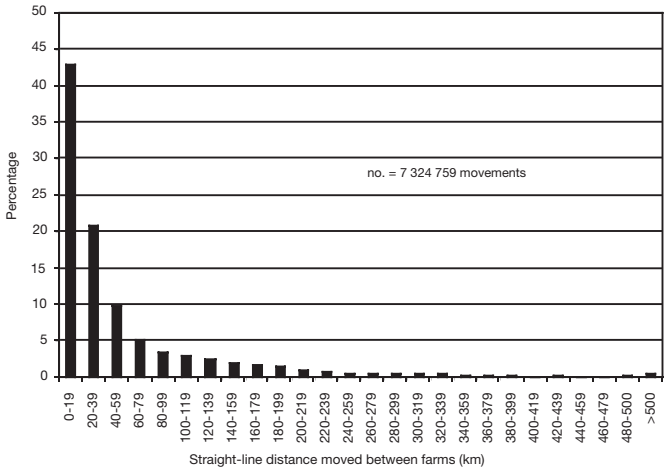


**Figure 3.2: Percentage distribution of straight-line distances moved by cattle in farm-to-farm transfers in Great Britain** - Reproduced with permission from Mitchell *et al.* (2005).

In this chapter, I explore the spatial structuring of molecular types of *M. bovis* in NI, using landscape genetics approaches to test possibilities a) and b) (above), with the underlying hypothesis that infection in the badger reservoir drives the spatial structure

seen in this pathogen population. I use an extensive dataset (>10,000 records) of *M. bovis* Variable Nucleotide Tandem Repeat (VNTR)-types routinely isolated from cattle in NI (Skuce *et al.*, 2010), in combination with badger genotypes and *M. bovis* VNTR-types generated from a survey of badgers killed in road traffic accidents (Abernethy *et al.*, 2011). The specific aims of the chapter are as follows:

- To examine the correlation between *M. bovis* VNTR-types isolated from badgers and cattle, and to characterise the spatial structure present in cattle VNTR-types in NI.
- To characterise the genetic population structure of NI badgers and assess whether there is evidence that landscape features (elevation above sea level and major rivers) act as a barrier to badger gene flow.
- To evaluate whether there is evidence that these landscape features constrain the spatial spread of *M. bovis* VNTR-types isolated from cattle.
- To directly compare badger population-genetic structure with the spatial occurrence *M. bovis* VNTR-types in cattle.

## 3.3 Materials and Methods

All analyses were carried out in R v3.1.1 (R Core Team & R Develoment Core Team, 2014) unless otherwise stated.

### 3.3.1 Spatial distribution of VNTR-types in cattle and badgers

#### 3.3.1.1 Cattle VNTR-type data

As discussed in Chapter 2, *M. bovis* isolates from infected cattle are routinely VNTR-typed in NI. The VNTR-typing dataset provided by Agri-Food and Biosciences Institute (AFBI) NI and owned by the Department for Agriculture and Rural Development

(DARD) NI, was described in the previous chapter and was used again for this analysis. The data consist of those VNTR-type records from cattle that were successfully linked to herd bTB breakdowns between 2003-2010 inclusive, downsampled to include only the first (earliest) VNTR record per outbreak (as multiple molecular typing records per bTB breakdown were only available for the latter part of this period), giving 10049 VNTR-typing records comprising 183 unique VNTR-types (see Chapter 2). The analysis of spatial structuring was conducted on the twenty VNTR-types associated with over 60 herd breakdowns each (VNTR-types 1-9, 11, 13, 15-17, 19, 23, 25, 42, 49, and 117, although VNTR-19 showed no evidence of spatial clustering (see below and Figure 3.3) and was therefore not included in subsequent analyses). These VNTR-types represent 92% of all VNTR-typed breakdowns in the dataset, and as there is likely to be a reduced chance of reliably inferring spatial structure in VNTRs associated with lower numbers of breakdowns, analysis was restricted to these twenty types.

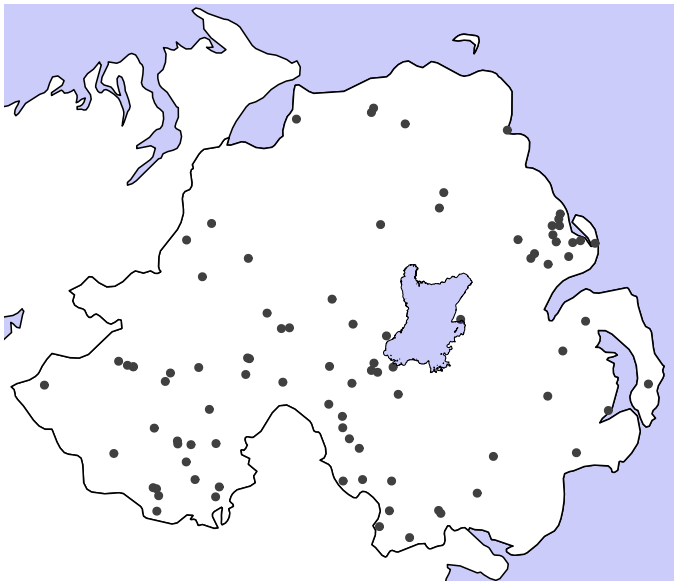

**Figure 3.3: Locations of NI herd breakdowns attributable to VNTR-19**

Further information on all recorded cattle holdings in NI was provided by Stewart McBride (AFBI), including details of the location (easting and northing) of the main farm building, the holding type, the species of animals present, and the number of cattle recorded to be present at routine herd testing for bTB and/or brucellosis. A dataset of cattle herd locations was generated based on herds that were either a)

represented in the VNTR dataset described above, or b) were recorded as enterprise type FARM and species BOV (bovine) in the cattle holding data and for which at least 10 cows had been recorded for at least two years between 2003-2010 in the herd testing data. The latter criteria avoid the inclusion of very small herds, which are less likely to be involved in the epidemiology of bTB, and restrict the dataset to those herds present over a large portion of the study timeframe. The resulting data gave location information for herds both infected and uninfected with bTB, and contained 21798 unique herd references with 20884 unique locations (i.e. some locations were linked to multiple herd references). However, it must be noted that in NI the official recorded locations of farms do not necessarily accurately reflect the locations of cattle associated with these farms, due to high levels of farm fragmentation and use of rented grazing (Abernethy *et al.*, 2006), and cattle movements between different parcels owned by the same person are not recorded.

#### 3.3.1.2 Badger VNTR-type data

Details of VNTR-types were obtained from a total of 265 badgers sampled in a survey of badgers found dead on the roads in NI between the years 1999-2014 (Abernethy *et al.*, 2011) and testing positive for *M. bovis* (data again provided by AFBI and DARD). The correlation between the number of times each VNTR-type was recorded in the badger dataset and the number of cattle breakdowns associated with that VNTR-type was evaluated using Pearsons correlation coefficient.

#### 3.3.1.3 Spatial clustering of VNTR-types: Nearest neighbour distances

A nearest neighbour distance (NND) approach was first used to assess whether *M. bovis* infections in cattle and badgers clustered by VNTR-type. Using the R package RANN, the geographic distance to the nearest infected herd or badger was calculated for infections of the same VNTR-type and compared to the distance to the nearest infection for those of different VNTR-types. If *M. bovis* infections are clustered with

those of the same VNTR-type, the nearest neighbour distance should be lower for comparisons between the same VNTR-type than for comparisons between different VNTR-types.

There were considerably more pairwise comparisons for pairs infected with different VNTR-types than for those infected with the same type, which would be expected to decrease the average NND between infections of different VNTR-types through chance alone. To account for this, I conducted the pairwise comparisons between different VNTR-types on a random sub-sample of the non-target VNTR-type pairs, equal in size to the number of target VNTR-type pairs. This randomisation process was repeated 1000 times, and the mean NND was recorded for each target VNTR-type record based on data for the 20 most common VNTR-types identified above.

NNDs were identified in this way for badger-badger comparisons, for cattle-cattle comparisons, and for badger-cattle comparisons, to assess whether VNTR-types showed spatial clustering in cattle and in badgers independently, and whether badger VNTRs were clustered with cattle VNTRs.

#### 3.3.1.4 Spatial clustering of VNTR-types: Variograms

To further confirm the presence and to assess the scale of spatial clustering of VNTR-types, a variogram approach was taken similar to that described in Goodchild *et al.* (2012). The distances between pairs of infections was identified for: all pairs of VNTR-typed badgers; all pairs of VNTR-typed cattle breakdowns; and for all pairs comprising a badger VNTR-type and a VNTR-typed cattle breakdown. The pairs falling into specified distance classes (0-2.5km, 2.5-5km, 5-10km, 10-20km, 20-30km, 30-40km, 40-50km, 50-75km, 75-100km, and >100km) were then identified, and the proportion of pairs infected with different VNTR-types was calculated for each distance class. If *M. bovis* infections were clustered geographically by VNTR-type, it would be expected that a higher proportion of pairs of infections separated by shorter distance classes would involve the same VNTR-type than those separated by larger distances.

To assess, for each distance class, whether the observed proportions infected with different VNTR-types differed from random, a null distribution of the results expected if there were no association between distance class and proportion of pairs infected with different VNTR-types was generated. This was carried out by randomly permuting the matrix of distances between pairs, and using these randomised pairwise differences to calculate the proportion of pairs in each distance class infected with different VNTR-types. This process was repeated 10000 times for each distance class and the 95% intervals were recorded.

This variogram procedure was also followed separately for each of the twenty most common VNTR-types identified above, for pairs of VNTR-typed cattle breakdowns where one or both members of the pair were associated with the specific VNTR-type. This allowed assessment of whether each of these VNTR-types in turn showed significant levels of geographical clustering.

#### 3.3.1.5 Badger isolates occurring outside of their VNTR-type cluster

Given that badgers only infrequently travel over long distances (see Figure 3.1), the occurrence of a VNTR-type in a badger outside the core cluster for that VNTR is a notable finding. Here, I estimated the core clusters for each of the major VNTR-types using the probability of occurrence for each VNTR-type as generated below. VNTR-typed badger isolates located more than 10km away from this core area were identified and excluded, to try and minimise the chance of including badgers dispersing from the core area. For each of the badger isolates sampled from outside these areas, I also recorded the distance to the nearest cattle isolate of the same VNTR-type that was isolated within two years and five years of that badger isolate.

#### 3.3.1.6 Estimating the spatial locations of VNTR-type clusters

After confirming the presence of spatial structure in *M. bovis* in NI, I went on to identify the locations of the VNTR-type clusters in cattle for nineteen out of the

twenty most common VNTR-types (excluding the single VNTR-type which did not show significant clustering, see Results). For the following analyses it was important to have an objective measure of the locations of VNTR-type clusters. Two other methods have previously been described to evaluate the locations of geographic clusters of different strain-types of *M. bovis*:

Firstly, an empirical method has been optimised to identify the home range of different *M. bovis* strain-types in GB, based on three years presence of a molecular type on at least two different cattle holdings in 5km grid squares over the last 5 years, with a 10km buffer then applied (AHVLA, 2012; Roberts *et al.*, 2014). This method has been successfully used to investigate bTB outbreaks in Britain for many years (e.g. Gopal *et al.*, 2006). However, this is governed by a set spatial scale, which would be inappropriate for this study as: a) NI is considerably smaller than GB; b) different VNTR-typing loci are employed in NI, and so NI VNTR-type clusters may occur on a different spatial scale to the GB home ranges; and c) the farming industry shows differences in NI compared to GB, with higher levels of farm fragmentation and more use of rented pasture (Abernethy *et al.*, 2006).

The second method used a point process approach to determine the spatial risk of occurrence of an *M. bovis* strain-type, relative to the risk of all other strains at that location (Diggle *et al.*, 2005). However, there is no biological reason to assume that different strain-types of *M. bovis* are subject to substantial competitive exclusion: multiple strains are recorded within the same herd breakdown in this dataset (Table 3.1), and multiple molecular types have been observed in a single infected individual for *M. bovis* infection in both deer (Gortazar *et al.*, 2011) and badgers (see Chapter 5), and for *M. tuberculosis* infection in humans (Cohen *et al.*, 2012; Walker *et al.*, 2012). The assessment of discontinuities in spatial structure carried out in this study would have been difficult to determine based on VNTR-type relative risk, and for these reasons the absolute risk of infection for each molecular type was determined to be more useful in this study.

For the reasons outlined above I decided against using either of the previous methods to

**Table 3.1: Number of unique VNTR-types recorded in breakdowns for which multiple cattle infections were VNTR-typed**

| Number of VNTR-types recorded | Frequency |
|---|---|
| 1 | 2042 |
| 2 | 1026 |
| 3 | 159 |
| 4 | 43 |
| 5 | 31 |
| 6 | 17 |
| 7 | 9 |
| 8 | 4 |
| 9 | 1 |
| 11 | 2 |
| 12 | 1 |
| 13 | 3 |
| 18 | 1 |
| 24 | 1 |

estimate the spatial locations of VNTR-types in NI. Instead, I developed a Generalised Additive Modelling (GAM) approach (Wood, 2006) to generate predictions of the spatial occurrence for each of the most common VNTR-types (excluding VNTR-19 which did not show significant geographical clustering, see Results). GAMs are statistical models that enable flexible non-linear smoothed relationships between response and explanatory variables by fitting multiple basis spline functions to different sections of the explanatory variable, joined at points called knots. A penalised likelihood approach is used to establish a trade-off between goodness of fit and smoothness/wiggliness of the model as determined by the spline functions, ensuring that the model is not over-parameterised. This approach is here applied to the interaction between X and Y coordinates, generating a non-linear estimate of the probability of occurrence of the response variable in cartesian space.

Unlike the GB approach described above, this method does not generate simple unambiguous boundaries for the core area for each molecular type, and therefore would not be appropriate for the same applied uses as the GB approach. However, in this study it provided objective estimates of the probability of occurrence of the major VNTR-types, and offered two major advantages over more basic smoothing methods: it allowed smoothing over multiple bandwidths (where indicated by the data); and it

enabled incorporation of the underlying distribution of cattle herd locations. I chose to focus on cattle VNTR-type records only for this analysis, as these represent a high density and relatively unbiased sampling of the *M. bovis* population in cattle, with high-resolution spatial data also available on the underlying cattle population. Compared to this, the badger VNTR dataset was based on passive sampling, with far fewer records available on the VNTR-types present in badgers.

The number of breakdowns associated with each of the VNTR-types was calculated for each of the unique cattle herd locations identified above, and GAMs were formulated using the R package mgcv (Wood, 2006). The number of times each major VNTR-type was recorded at each herd location was modelled as a binomial process with logit link, with the number of trials equal to the eight years covered by the data (each routine herd bTB test being treated as a Bernoulli trial, and bTB testing assumed to be carried out approximately annually). These GAMs aimed to model the spatial autocorrelation present in the data, and therefore the smoothed interaction between the x- and y-coordinates of each herd was used as the sole explanatory variable. The coastline of the north of Ireland and Lough Neagh were included as boundaries using soap film smoothing, which provides a solution to the incorporation of boundaries to the smoothed domain of the GAM (Wood *et al.*, 2008). The gamma parameter, which determines the extent to which extra "wiggliness" in the smoothing parameter is penalised within the model, was set to 1.4 to reduce over-fitting (Kim & Gu, 2004; Wood, 2006). The soap film smooth requires the user to manually generate knots (the points at which basis splines are joined), and for these I used a random sample of 500 herd locations, removing those within 2.2km of a boundary to enable the soap film algorithm to run. The resulting GAM models were then used to generate the probability of occurrence of each VNTR-type at each 1km$^2$ grid square location across NI.

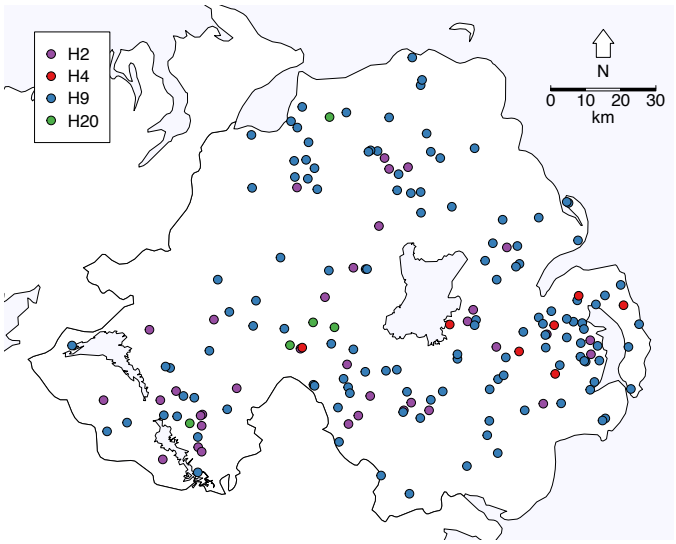


**Figure 3.4: Mitochondrial DNA haplotypes found in NI badgers**

## 3.3.2 Population structure and landscape genetics of NI badgers

### 3.3.2.1 Badger data

Genotype information for 172 badgers taken from the survey of badgers killed by vehicles in NI (Abernethy *et al.*, 2011) was provided courtesy of Adrian Allen (AFBI) and DARD. These badgers were sampled from across NI during the years 2011-2013, and although there is some overlap with the dataset of bTB infected VNTR-typed badgers described above (26 badgers present in both datasets), many of the genotyped badgers were not represented in the VNTR-typing dataset. Genotyped badgers were associated with x-y coordinates, and these locations are shown in Figure 3.4. Badgers were genotyped at 16 microsatellite loci (Mel104, Mel106, Mel109, Mel111, Mel112, Mel117, Mel102, and Mel105; Carpenter *et al.*, 2003), plus a 351bp segment of the mitochondrial DNA D-loop region (O'Meara *et al.*, 2012).

### 3.3.2.2 Badger population structure

Pairwise genetic distances between badger microsatellite genotypes were estimated in the R package ape, using absolute genetic distances (Paradis *et al.*, 2004; Prevosti *et al.*,

1975). Using these distances, a Mantel test and Mantel correlogram were carried out in the R package ecodist (Goslee & Urban, 2007).

The programme STRUCTURE v2.3.4 (Pritchard *et al.*, 2000) was used to determine the number of distinct genetic clusters (k) of badgers within NI. STRUCTURE implements a Bayesian clustering algorithm for multilocus genotypic data, assigning individuals to different populations so as to minimise deviations form Hardy-Weinberg and linkage equilibria. STRUCTURE was run for 100000 MCMC iterations with a burnin of 50000, with ten independent replicates of k values from 1-6. The results was assessed using Evannos $\Delta$k (Evanno *et al.*, 2005), and the ten replicates for the most likely k value were summarised in CLUMPP (Jakobsson & Rosenberg, 2007).

Spatial population structure was also analysed using GENELAND within the R environment (Guillot *et al.*, 2005). GENELAND is based on a similar principle as STRUCTURE, again aiming to minimise deviations from Hardy-Weinberg and linkage equilibria. In contrast to STRUCTURE though, GENELAND also allows explicit inclusion of sampling locations in determining the number of genetic clusters. GENELAND analyses were run for $10^6$ MCMC iterations with a burnin of 1000, for values of k from 1-6.

#### 3.3.2.3 Badger landscape genetics

To assess whether landscape variables were acting as barriers to badger gene flow, resistance surfaces were generated in ArcGIS for several landscape variables hypothesised to be associated with badger population structure. The analysis of badger landscape genetics described in this section was carried out by Dr Jimena Guerrero Flores (University of Glasgow).

A raster of elevation across NI in meters above sea level, cell size 500m$^2$, was downloaded from www.diva-gis.org/gdata (Figure 3.5) and this was also used to generate a raster of slope scaled in degrees. These were translated into resistance surfaces, keep-

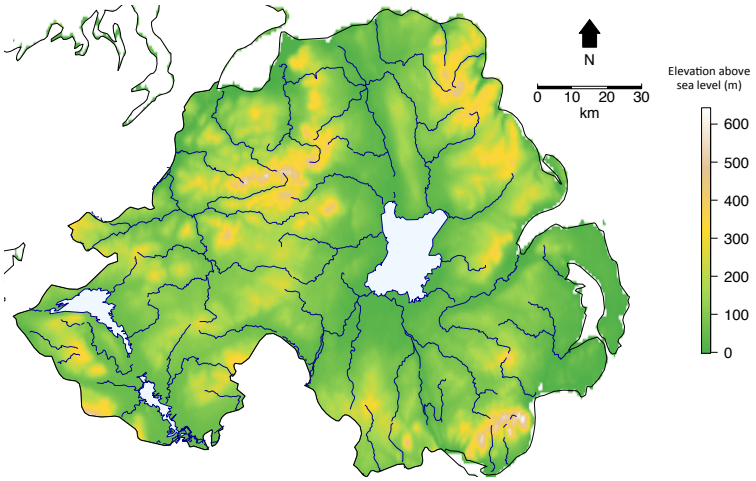


**Figure 3.5: Elevation and major rivers in NI**

ing the original values of both, with the assumption that higher elevations and steeper slopes are likely to give higher resistances to badger gene flow.

Vector files of the major rivers and water bodies in NI were downloaded from www.doeni.gov.uk and combined (Figure 3.5). These were converted into a resistance surface with cell size of 500m$^2$, assigning low resistance values (1) to land cells, and high resistance values (100) to cells containing water.

In addition, a raster of badger habitat suitability across NI (Reid *et al.*, 2011) was provided by Neil Reid and David Wright (Queens University, Belfast), with cell size of 1km$^2$ (Figure 3.6). This was converted directly to a resistance surface, maintaining the original cell values, in this case assuming the higher the cell value, the less the resistance (higher conductance) to gene flow.

These resistance surfaces, along with the locations of the badger samples, were converted to resistance distance matrices in Circuitscape v4.0 (Shah & McRae, 2008). The pairwise distances to the eight nearest neighbour cells of each sampling location were calculated as average resistance (or conductance) for each surface.

Partial Mantel tests were then carried out in the R package ecodist (Goslee & Urban, 2007) to assess whether any of these surfaces were associated with the pairwise genetic

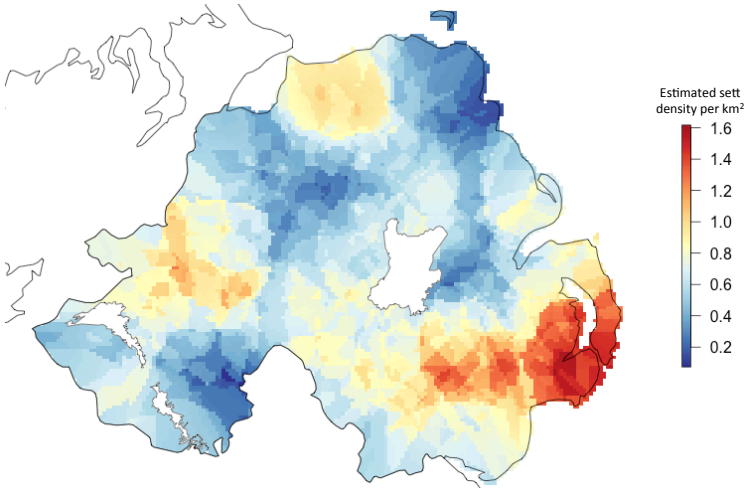


**Figure 3.6: Badger sett density** - estimated by the badger survey of Northern Ireland 2007/2008 (Reid *et al.*, 2011).

distance between badger samples, while accounting for straight-line distance between sampling locations.

### 3.3.3 Correlations between landscape features and VNTR-types in cattle

If a landscape feature, such as elevation or rivers, acts as a barrier to the spread of *M. bovis*, one would expect the presence of this feature to co-occur with breaks in the spatial structure of the pathogen population. Using the spatial probability of occurrence of the most common NI VNTR-types estimated above, I took two approaches to identifying such breaks in the spatial structure of *M. bovis*.

#### 3.3.3.1 Identifying spatial discontinuities in VNTR-types: slope in probability of occurrence

If an area of high probability of occurrence of a VNTR-type represents an area through which the VNTR-type is spreading freely, then regions of decreasing probability of occurrence should denote the edge of this spread (Fortin & Dale, 2005; Womble, 1951). Some of these edges may represent areas in which pathogen spread has been blocked by

a landscape barrier. Alternatively, the edge of an area of high probability of occurrence might represent a snapshot in time of the onward spatial dissemination of that VNTR-type, without the involvement of any barrier.

Given this rationale, the first order rate of change (i.e. the slope) of the probability of occurrence estimated above for each VNTR-type was calculated (scaled in degrees) in the R package raster, and this slope value was used as a proxy for the likelihood of a point being at the edge of the VNTR-type cluster. For each VNTR-type, I then identified grid squares with either a high probability of VNTR-type occurrence or a high slope in probability. Grid squares located outside these regions were removed from the analysis, as areas of low probability of occurrence and low slope in occurrence are not expected to contain any information on the spread of a VNTR-type. This was carried out individually by eye for each VNTR-type, as there was no objective cut-off value that was appropriate across all types. In this instance this procedure was not carried out by two independent observers and therefore incorporates an unknown degree of subjectivity and may not be reproducible.

Many grid square locations were represented multiple times in the resulting dataset, namely those locations for which multiple VNTR-types had either high probability of occurrence or high slope values. For these locations, if a grid square is located at the edge of one VNTR-type cluster (high slope value), and yet is in the centre of another VNTR-type cluster (low slope value), it is likely that this area does not contain a barrier to the spread of the pathogen, as the second VNTR-type appears to be spreading freely at this point. I therefore summarised the results for these locations, using the minimum slope value across all of VNTR-types present at that location. The square root of these minimum slope values was then taken to normalise the distribution of slope values, and these were linked to the relevant spatial and landscape explanatory variables (see below).

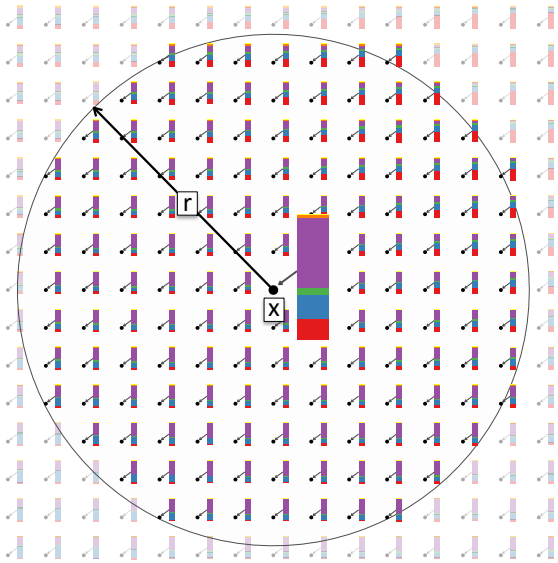


**Figure 3.7: Visual summary of the method for estimating community uniqueness** - The probability of occurrence of each of the most common VNTR-types was extracted at each grid square location (x). The probability of occurrence of the VNTR-types was also extracted for all grid squares within a set radius (r) surrounding location x, and these were summarised, and the resulting vector was compared to the community present at location x.

#### 3.3.3.2 Identifying spatial discontinuities in VNTR-types: comparison of community structure

The second approach to identifying discontinuities in the spatial structure of the *M. bovis* population considered the community of VNTR-types present at each location, relative to the area surrounding that point. If a point is located in an area of spatial discontinuity in *M. bovis* population structure, then it would be expected to show an increased level of divergence compared to at least some of the VNTR-type communities surrounding it.

At each grid square location (x, Figure 3.7), the probability of occurrence of each of the most common VNTR-types as estimated above was extracted, and this vector was taken to represent the VNTR-type community present at that point. The probability of occurrence of the VNTR-types were likewise extracted for all grid squares within a set radius (r, Figure 3.7) surrounding location x, and summarised to generate another vector. This was then compared to the community present at location x using Kullback-

Leibler (KL) divergence, calculated in the R package entropart (Marcon et al., 2014). This process was carried out for each 1km grid square across NI, and for r values of 2.5km, 5km, 7.5km, 10km, 15km and 20km. The log of the resulting values was taken to normalise the results, and these were linked to the relevant explanatory variables (see below) at each location.

#### 3.3.3.3 Comparing discontinuities in M. bovis spatial structure to landscape features

Landscape layers identified as candidate explanatory variables for this analysis (elevation, rivers, sett density) are described in the previous section on badger landscape genetics. Additionally a raster representing the border between the different genetic populations of badgers, as estimated in the Geneland analysis described above, was generated by calculating the slope in the spatial probability of belonging to badger Population 1 (Figure 3.8B), generated as output from the Geneland analysis.

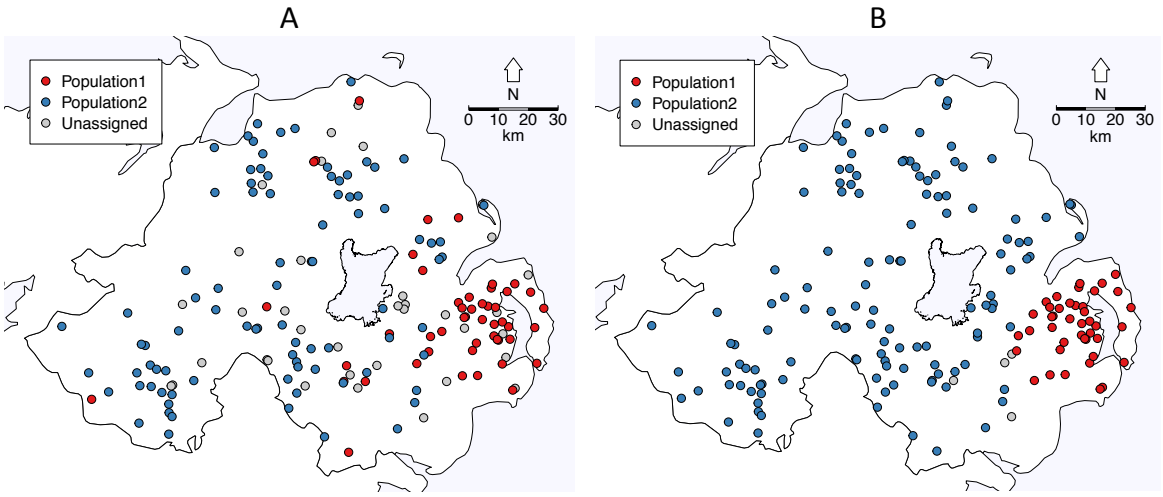


**Figure 3.8: STRUCTURE (A) and Geneland (B) population structures for the NI badgers** - Red and blue points indicate badgers assigned to population 1 or 2 with >75% probability, while grey points show badgers which were assigned with ≤75% probability.

From the map of elevation, I then generated polygons representing areas of land with elevation greater than 300m above sea level, and from the raster of Geneland slope I generated a polygon indicating the border between the two badger populations.

The GAM predictions of probability of VNTR-type occurrence described above were based on the presence of cattle herds, and so the models are expected to become less accurate where cattle herds are sparse. Therefore, for each point in the 1km grid across NI, I identified the distance to the nearest cattle herd. Locations at which the nearest cattle herd was further away than the 99$^{th}$ percentile of all distances to nearest herd were identified and removed from the analysis.

Statistical models were generated using the mgcv package in R (Wood, 2006), allowing incorporation of non-linear interaction between the x- and y-coordinates at each point: in this case, the x-y smooth was included to allow us to account for underlying spatial autocorrelation not explained by the model covariates (Bivand *et al.*, 2008). Two models were generated to assess the association between the explanatory variables above and discontinuities in the spatial structure of the *M. bovis* population, using a Gaussian response with identity link.

The first model took the square root of the minimum slope in probability of VNTR-type occurrence at each location (see above) as response variable. This was modelled as a function of the smoothed x-y coordinates plus the following linear terms: distance to the nearest area of high ground (>300m), distance to the boundary between the two genetic populations of badgers, distance to the nearest river, and estimated badger sett density. The square root was taken of all these linear explanatory variables to normalise their distributions.

The second model used the same explanatory variables. For this model, the response variable was taken as the log of the value generated from the KL comparison between the VNTR-type community composition at each point and that of the surrounding area (see above and Figure 3.7). This model was run for response values estimated considering the VNTR-type community within radii of 2.5km, 5km, 7.5km, 10km, 15km and 20km surrounding the point of interest.

Both models specified 1000 knots for the spatial smooth, and again used a gamma value of 1.4 to avoid overfitting. The large number of knots used in these models

were necessary to take into account the high degree of spatial autocorrelation in both response and explanatory variables, however it also presented a challenge to model fitting. As many of these knots were retained, the resulting models had very high numbers of degrees of freedom, very few of which were related to the linear explanatory variables. Removing linear explanatory variables from the model resulted in alterations in the use of knots, and the subsequent models therefore showed considerable variation in their degrees of freedom. As the Akaike Information Criterion and Likelihood Ratio Test methods of model selection rely on comparison of model degrees of freedom as well as likelihood, this variability in degrees of freedom associated with the spatial smooth was a potential cause for concern. To confirm consistency in the size and direction of the effects estimated by the models for each of the linear predictors, NI was subdivided into two polygons (Figure 3.9), each expected to contain representative samples of the spatial features under investigation. In addition to running each of the models described above for all locations across NI, each model was also run separately for the two polygon subsets, this time using 500 knots.

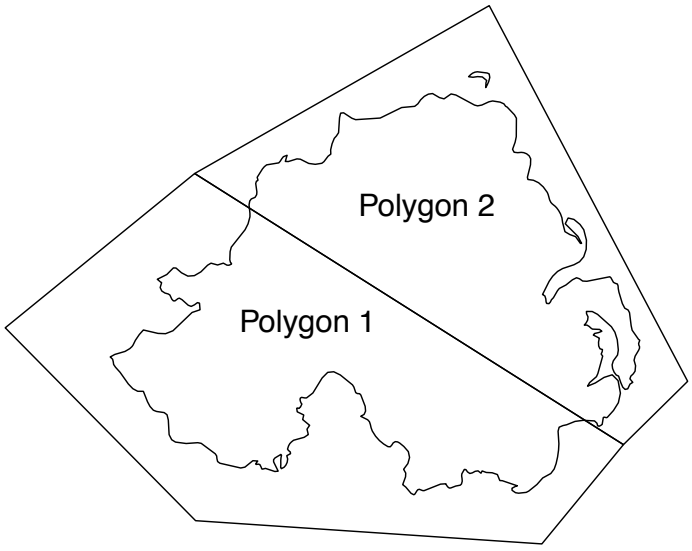


**Figure 3.9: Polygons used to subset the data for the models comparing discontinuities in the spatial structure of the *M. bovis* population to landscape features**

### 3.3.4 Accounting for badger genetic structure using VNTR-type distributions

Three direct comparisons were made to identify associations between badger genetic similarity and the geographic locations of VNTR-types seen in cattle.

Firstly, if mixing between badger populations correlates with the contacts involved in the spread of bTB, the border between badger genetic populations should be associated with breaks in the spatial structure of the VNTR-types. This was tested by the incorporation of the Geneland boundary in the analysis described in the previous section.

The second assessment of the correlation between badger genetic similarity and spatial structure of cattle VNTR-types evaluated whether badger samples separated by the edge of a VNTR-type cluster (as estimated using slope in probability of occurrence, above) were less genetically related than badgers without a VNTR-boundary between them, and was again carried out by Dr Jimena Guerrero Flores. To evaluate this, a raster was generated comprising the sum of all the slope values for each VNTR-type GAM at each grid square across NI. These were reclassified in ArcGIS using the jenks data clustering method (arranging data into new classes by minimising the variance within classes and maximising the variance between classes) to enable the generation of a resistance distance matrix in Circuitscape, as described previously. A partial mantel test was then used to assess whether badgers separated by the edge of a VNTR-type cluster were less closely related after accounting for straight-line distance between locations. This analysis was also run separately for badgers from each of the two Geneland populations identified above. Here, a cut-off value of 75% was used to determine population membership, identifying 42 badgers as belonging to Population 1 and 126 badgers belonging to Population 2 (see Figure 3.8B). Four badgers that did not meet this membership threshold were removed from the analysis.

For a third assessment of the correlation between VNTR-types and badger genetics, the community composition of VNTR-types present in cattle was compared to badger

genetic distances. The probability of occurrence for each major VNTR-type was extracted from the GAM predictions at the location of each genotyped badger sample, and pairwise comparisons between the distribution of VNTR-type probabilities at each point were made using KL divergence in the R program entropart (Marcon *et al.*, 2014). A partial Mantel test was then used to evaluate whether badger genetic similarity correlated with changes in VNTR-type community composition (i.e. KL divergence) after accounting for the straight-line distance between locations.

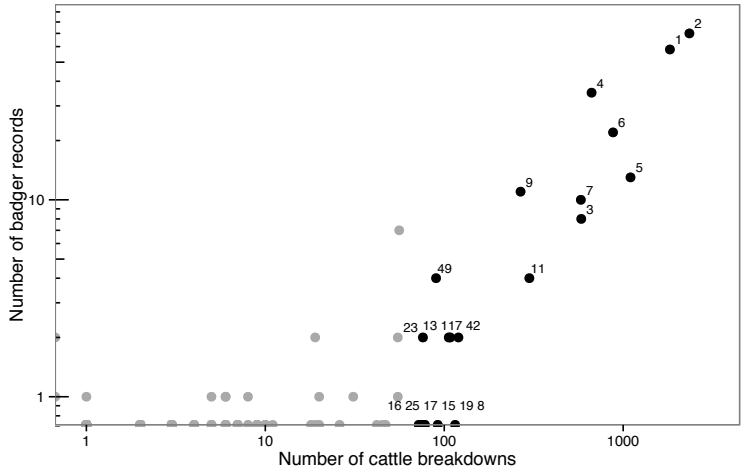


**Figure 3.10: Comparison of the numbers of badger infections (1999-2014) and cattle breakdowns (2003-2010) recorded for different *M. bovis* VNTR-types in NI** - (log-log scale) VNTR-types with more than 60 cattle breakdowns are shown in black and labelled by VNTR-type.

## 3.4 Results

### 3.4.1 Spatial distribution of VNTR-types in cattle and badgers

A total of 183 VNTR-types were recorded in cattle between 2003-2010, with the 20 most common VNTR-types accounting for more than 90% of typed breakdowns over this period. A lower number of VNTR-types (n=28) were recorded in badgers (probably due to lower sampling intensity in the badgers), however their frequency was highly correlated with that in cattle (r= 0.96, p<0.001, Figure 3.10).

Nearest neighbour distances visually confirmed geographic clustering within the major

VNTR-types, both when considering cattle infections and badger infections independently, and when comparing the distance from a badger infection to the nearest cattle breakdown (Figure 3.11), with on average a lower distance to the nearest neighbour infected with the same VNTR-type (blue shading) than the nearest neighbour infected with a different VNTR-type (red shading).

A variogram approach also indicated geographical clustering by VNTR-type. Pairs of VNTR-typed infections were significantly more likely to be infected with the same VNTR-type when separated by shorter distances, for cattle and badger pairs independently, and also for pairs consisting of a cattle breakdown and a badger infection (Figure 3.12). Using this approach separately for each of the twenty most common VNTR-types indicated significant clustering in all major VNTR-types, with the exception of VNTR-19 (Figure 3.13).

The above analyses indicate that cattle breakdowns and badger infections are clustered geographically by VNTR-type, both in cattle and badgers independently and also when considering both species in combination.

Thirteen VNTR-typed badger isolates were recorded to be located >10km out of the core area for their VNTR-type, as estimated using the probability of VNTR-type occurrence below. These isolates were often located within a short distance of cattle isolates of the same VNTR-type sampled within five years of the badger isolates (Table 3.2), raising the possibility of an epidemiological link between them.

GAM predictions of the spatial occurrence of VNTR-types in NI cattle for the major VNTR-types are shown in Figure 3.14. The results for each of the most common VNTR-types (with the exception of VNTR-19, due to lack of clustering mentioned above) indicated clearly localised areas of high probability of occurrence, and these visually correlated well with the locations of the herds infected with each VNTR-type (data not shown).

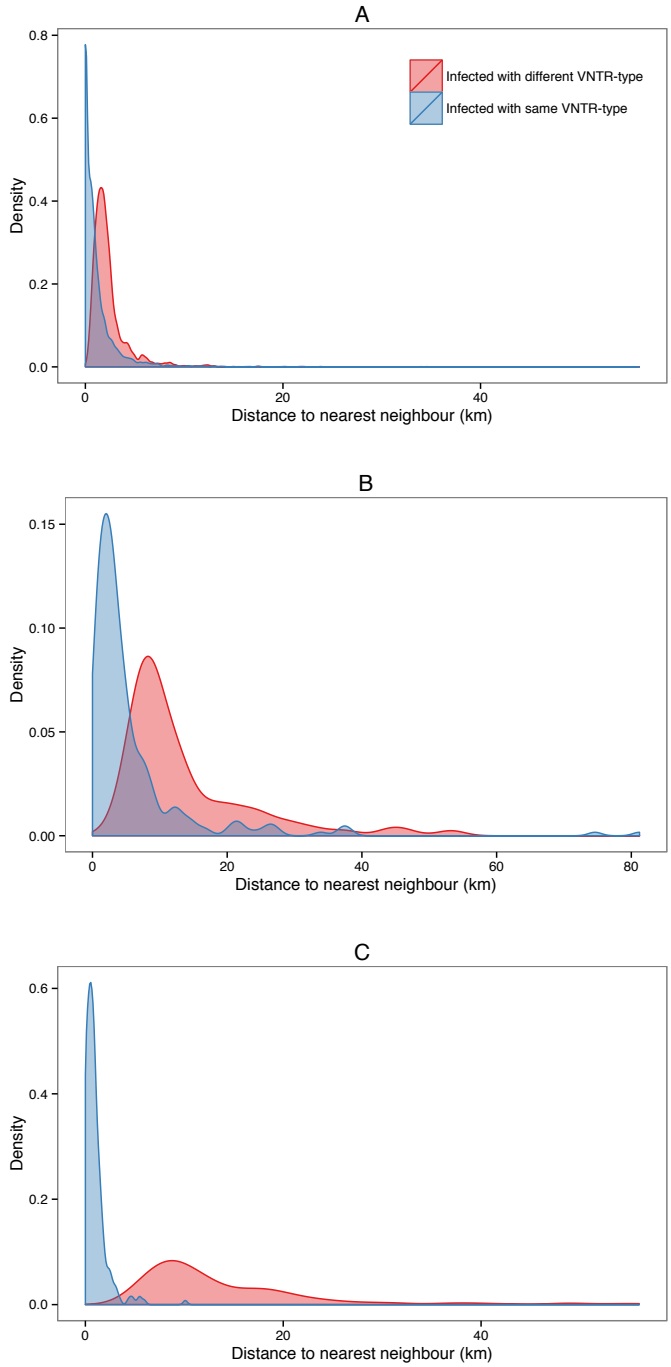


**Figure 3.11: Nearest neighbour distances, showing the distribution of distances between each VNTR-type record and the nearest neighbouring isolate of the same VNTR-type (blue) and of a different VNTR-type (red)** - Comparisons are shown for cattle isolates (A), badger isolates (B) , and badger-cattle comparisons (C).

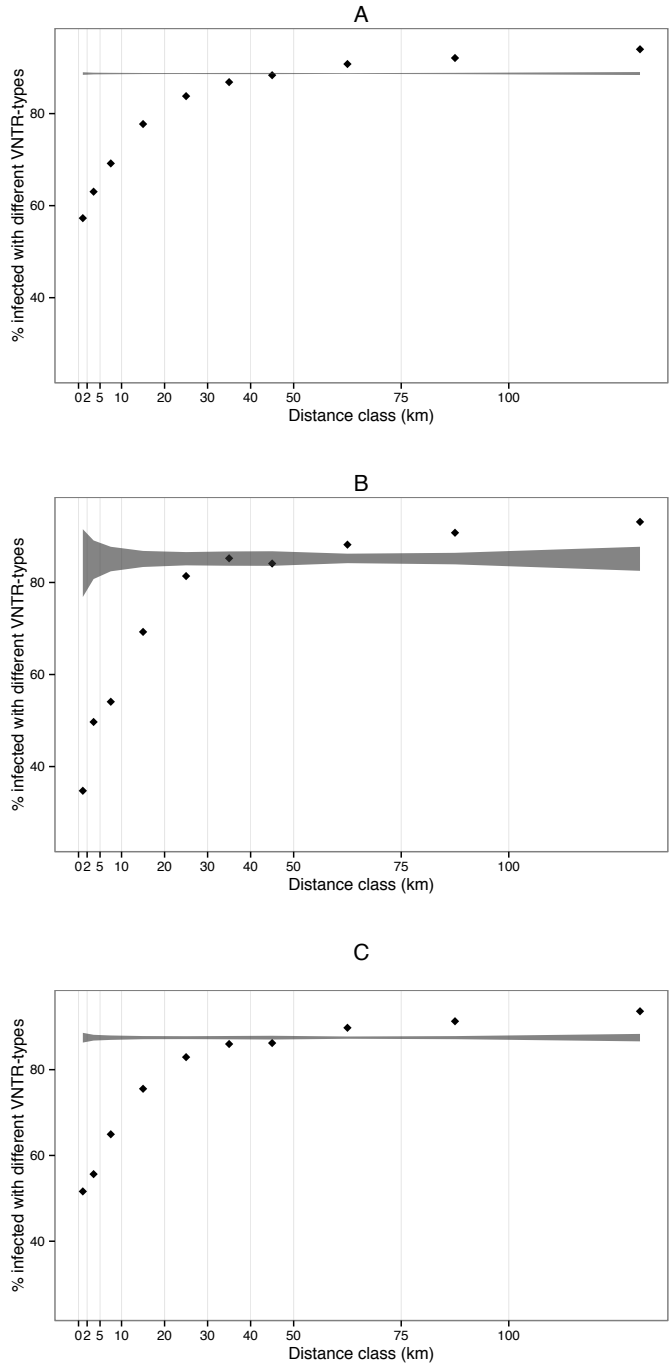


**Figure 3.12: The percentage of pairs separated by different distances, that were infected with different** $M.\ bovis$ **VNTR-types** - for cattle-cattle pairs (A), badger-badger pairs (B), and badger-cattle pairs (C). Points show the observed proportion infected with different VNTR-types for each distance class. Grey shading indicates the 95% interval for the values expected under the null hypothesis of no association between distance class and percentage of pairs infected with different VNTR-types.

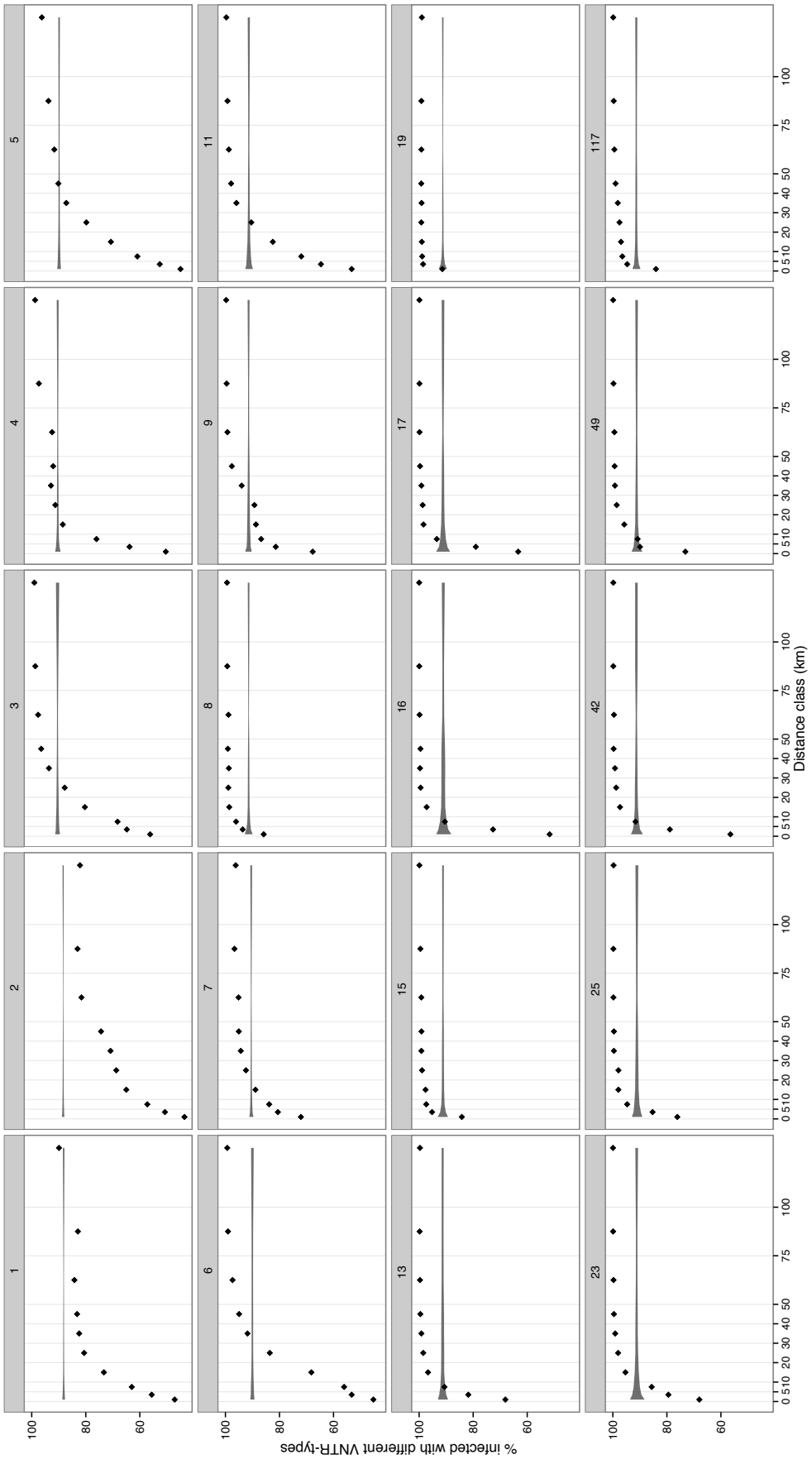


**Figure 3.13: Variograms by VNTR-type** - The percentage of pairs of cattle breakdowns for each VNTR-type, separated by different distances, infected with different VNTR-types. Points show the observed proportion infected with different VNTR-types for each distance class. Grey shading indicates the 95% interval for the values expected under the null hypothesis of no association between distance class and percentage of pairs infected with different VNTR-types.

**Table 3.2: Minimum distances between VNTR-typed badger *M. bovis* isolates >10km outside of the home range for their VNTR-type, and cattle isolates of the same VNTR-type recorded within 2 years and 5 years of the relevant badger isolate.**

| VNTR-type | Min. distance to cattle infection of same VNTR-type <2yrs of badger sample (km) | Min. distance to cattle infection of same VNTR-type <5yrs of badger sample (km) |
|---|---|---|
| 1 | 6.72 | 3.22 |
| 2 | 15.00 | 1.84 |
| 2 | None | 5.47 |
| 3 | 0.67 | 0.51 |
| 4 | 2.94 | 2.94 |
| 4 | 4.42 | 1.42 |
| 4 | 2.61 | 1.42 |
| 4 | 9.47 | 3.06 |
| 5 | 0.32 | 0.32 |
| 5 | 1.61 | 1.61 |
| 6 | 0.92 | 0.57 |
| 7 | 2.55 | 1.80 |
| 49 | 0.00 | 0.00 |

### 3.4.2 Population structure and landscape genetics of NI badgers

Four mtDNA haplotypes were recorded in the NI badger samples (Figure 3.4). Haplotypes H9 and H2 were seen most commonly (recorded in 125 and 35 badgers respectively). Both of these were widespread geographically throughout NI, and both haplotypes were also found to be common in RoI (O'Meara *et al.*, 2012). Haplotype H4 was recorded less frequently in this study (seven badgers) and appears restricted to the south east of NI. This haplotype is common in badgers throughout GB, but had not been identified to date in Ireland. The fourth haplotype, H20 (four badgers), differs from haplotype H9 by a single base pair deletion, and to our knowledge has not previously been recorded.

Results from STRUCTURE analyses of the microsatellite data showed that NI badgers fell into two genetic populations ($\Delta$K of 607.5 for K=2, compared to $\Delta$K <100 for other values of k). Geneland analysis also confirms two genetic populations of badgers (63.8% of the MCMC chain gave two clusters after a burnin of 200, compared to 19.2%, 6.5%, 1.7%, and 8.8% for three, four, five, and six clusters respectively). Both analyses thus broadly agree that badgers from the southeast of NI are genetically distinct from those in the rest of the country (Fig 3.8). Haplotype H4 (above) was more common than

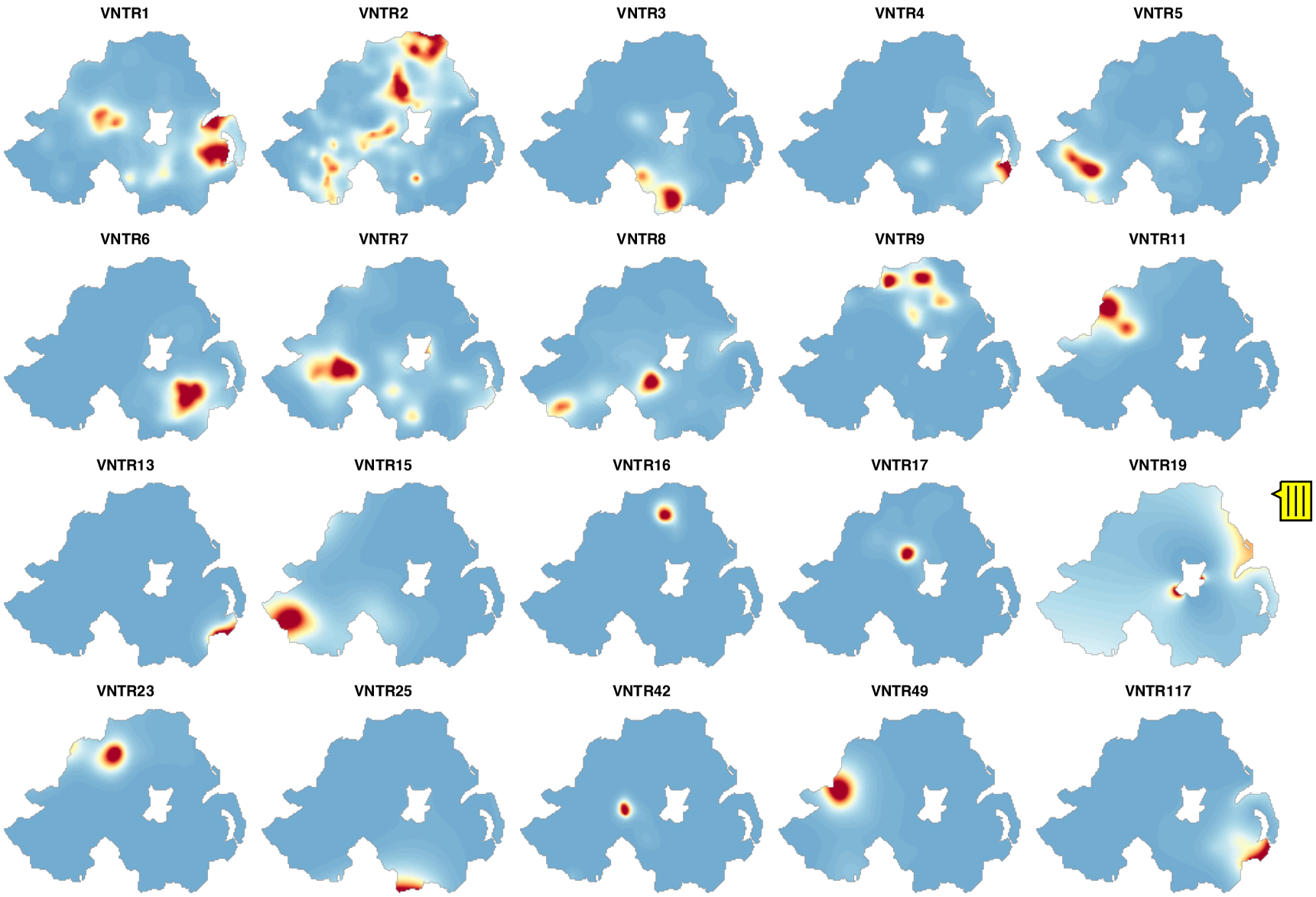


**Figure 3.14: Spatial probability of occurrence for the 20 most prevalent NI VNTR-types, 2003-2010** - (including VNTR-19). Red shading indicates areas of high probability of occurrence and blue shading indicates areas of low probability.

expected in Geneland population 1 (Fishers exact test: p=0.011), while other mtDNA haplotypes showed no significant associations with either population ($p \geq 0.05$).

There was a significant but weak correlation between genetic and geographic distance in the NI badgers (Mantel test: r=0.14, p=0.001; multiple regression on matrices: $r^2$=0.016, p<0.001). A Mantel correlogram confirms this association, showing significant positive autocorrelation over shorter distance classes (Figure 3.15).

Partial Mantel tests showed no significant correlation between badger pairwise genetic distances and landscape resistance, indicating none of the landscape surfaces tested here act as an appreciable barrier to badger gene flow (p-values= 0.81, 0.776, and 0.893 for elevation, slope, and rivers/water bodies respectively).

.

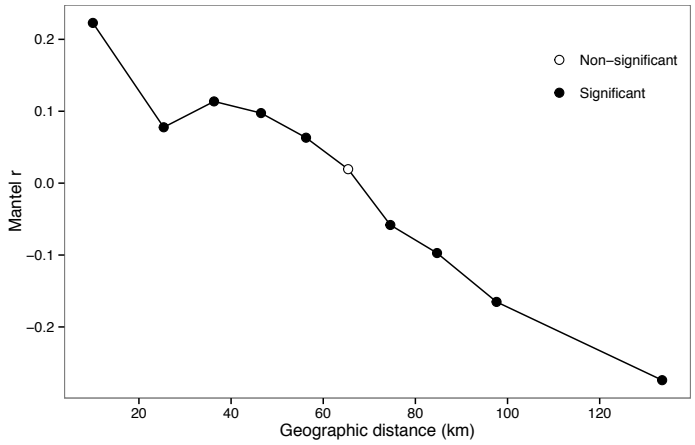


**Figure 3.15: Mantel correlogram for NI badgers** - showing relationship between genetic similarity and geographic proximity over different distance classes

### 3.4.3 Correlations between landscape features and VNTR-types in cattle

To assess whether landscape or badger features were associated with an interruption of the spread of *M. bovis* in NI, two separate models were constructed to compare

discontinuities in the spatial structure of the NI *M. bovis* VNTR-types isolated from cattle, as estimated above, and landscape and badger features.

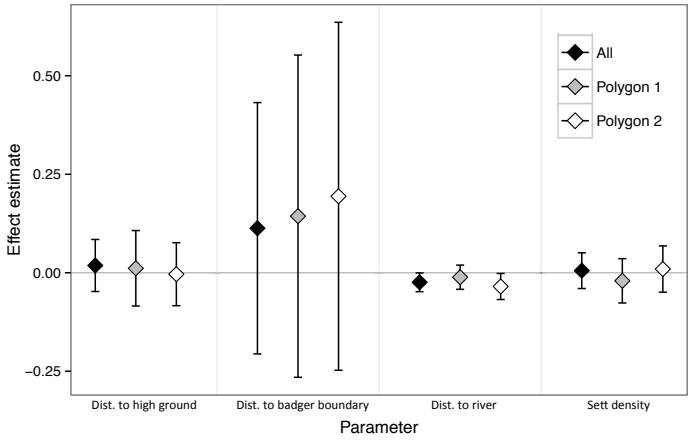


**Figure 3.16: Comparison between minimum slope in VNTR-type probability, and landscape and badger features** - Black points indicate results for model run for all points across NI, and grey and white points indicate results for subsets corresponding to Polygons 1 and 2 respectively (Figure 3.9).

Although the large number of knots required for the smoothed spatial component of these models made formal model comparison challenging (see Materials and Methods), results indicate that neither distance to high ground, distance to the nearest river, nor sett density give consistently significant results. This was true both when considering the minimum slope in the probability of VNTR-type occurrence as the proxy for discontinuities in *M. bovis* spatial structure (Figure 3.16), and when using dissimilarities in the VNTR-type community structure as the proxy (Figure 3.17).

Using minimum slope in VNTR-type probability as response, the distance to the boundary between the two genetic populations of badgers also gave non-significant results (Figure 3.16). By contrast, in the model of VNTR-type community dissimilarity, this proxy for spatial discontinuity was negatively correlated with the distance to the genetic boundary between badger populations when considering all locations (Figure 3.17A) and when modelling the subset containing only the Polygon 1 locations (Figure 3.17B). Although the Polygon 2 subset was also consistently negatively correlated with the border between badger populations, these results were non-significant (Figure 3.17C). However, the untransformed distance to a boundary feature is not likely to be

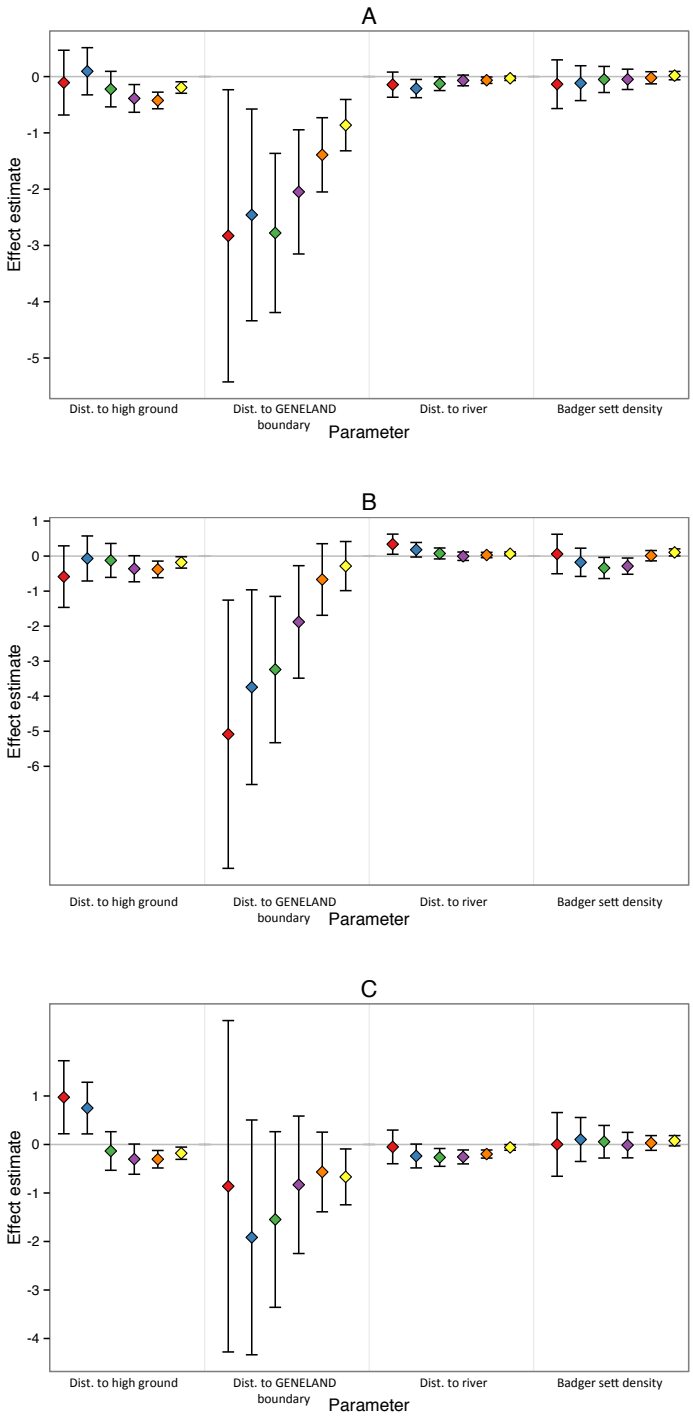


**Figure 3.17: Comparison between landscape and badger features (using distance to feature for high ground, badger population boundary and river) and community uniqueness** - Colours show results for community uniqueness measured using varying radii around the point of interest. A Shows model results for all points across NI, and B and C give results for points within polygons 1 and 2 respectively (Figure 3.9).

a particularly biologically meaningful in this case. Repeating these analyses using a measure that, for example, declines exponentially with increasing distance from the boundary is likely to reflect much more accurately our expectations of the underlying biology.

### 3.4.4 Accounting for badger genetic structure using VNTR-type distributions

The results of the association between the borders between badger genetic populations and discontinuities in the spatial structure of *M. bovis* VNTR-types are described in the section above.

Badger samples separated by the edge of a VNTR-type cluster, as estimated by the slope in probability of VNTR-type occurrence, were more genetically dissimilar than expected (Mantel r=0.14, p=0.0049). However, no significant association was found when this analysis was run for each of the Geneland badger populations separately (Figure 3.8B, Population 1: Mantel r=0.097, p=0.28; Population 2: Mantel r=0.036, p=0.50). This would suggest that the association between badger genetic dissimilarity and the discontinuities in VNTR-type spatial structure is predominantly driven by the boundary between the two badger populations.

By contrast, the comparison between the VNTR-type communities present at the locations of genotyped badgers was not significantly correlated with the genetic similarity between the badgers at these locations (Mantel r =-0.0019, p=0.46).

## 3.5 Discussion

### 3.5.1 Spatial distribution of VNTR-types in cattle and badgers

Two separate approaches both confirm that *M. bovis* is clustered geographically by VNTR-type in NI, validating the results described in Skuce *et al.* (2010). VNTR-

types show significant clustering when considering infections in cattle herd outbreaks or badger infections separately, and also when comparing VNTR-types seen in both species (Figures 3.11 and 3.12), similar to observations from Britain (Goodchild *et al.*, 2012; Woodroffe *et al.*, 2005), and the Republic of Ireland (Olea-Popelka *et al.*, 2005). In addition, there is a clear correlation between the number of times individual VNTR-types have been recorded in badgers and cattle in NI (Figure 3.10). Taken together, these findings indicate that the distribution of *M. bovis* molecular types in cattle and in badgers is closely linked in NI.

The presence of VNTR-types sampled at locations outside of the core cluster for a specific VNTR-type is likely to indicate translocation of that type from the area in which it is established. In cattle, infections outside the core VNTR-type area are seen regularly and are generally ascribed to transport of infected cattle between herds (e.g. Gopal *et al.*, 2006). By contrast, badgers only infrequently travel over large distances (Byrne *et al.*, 2014; Pope *et al.*, 2006; see also Figure 3.1), therefore the finding of several VNTR-typed isolates in badgers outside that VNTR-types core area is more surprising. As many of these badger isolates occur within very short distances of cattle breakdowns of the same VNTR-type (Table 3.2), I suggest that at least some of these badger infections are likely to represent the transmission of the VNTR-type from translocated cattle into the badger population. An alternative possibility is that these VNTR-types are maintained in badgers in these areas without infecting local cattle. Although we cannot out rule the latter possibility given the (relatively) limited sample size of VNTR-typed badger isolates, it seems unlikely given the strong correlations between badger and cattle VNTR-types described above.

The clear spatial clustering described here in cattle VNTR-types remains interesting for two reasons: due to the low sensitivity of the bTB test, the inadvertent transport of infected cattle over considerable distances is likely to be common; and furthermore, in NI high levels of farm fragmentation and the use of rented grazing generate significant discrepancies between the recorded location of a herd and the actual location of the cattle (Abernethy *et al.*, 2006). The observation of geographic clustering despite these

points suggests that the spatial processes driving clustering must be strong enough to overcome them.

### 3.5.2 Population genetic structure of NI badgers

Both STRUCTURE and Geneland analyses of the NI badger samples agree on a genetically distinct population in the south-east of the country (Figure 3.8). The mtDNA haplotype H4 (Figure 3.4), which has not previously been reported in Ireland but which is common in Britain (O'Meara *et al.*, 2012), is also associated with the south-eastern badger group. Therefore a possible explanation of the distinct badger population in southeast NI might be that this population has received badgers immigrating from GB, either naturally or through translocation. Regardless of the specific ancestral history, the boundary between the two populations suggests that they have not yet had sufficient time to mix, or alternatively that admixture is constrained by landscape features not tested in the analysis below.

The badger microsatellite data show weak but significant correlation between genetic and geographic distances between badger samples. Badgers have been reported as showing high levels of relatedness within social groups and between neighbouring social groups (Carpenter *et al.*, 2005; Dugdale *et al.*, 2008; Frantz *et al.*, 2010a), and so only finding a weak signal of isolation by distance in these data was unexpected. It is possible that the intensity of sampling here was not fine enough to capture many of these close relationships between neighbouring setts, and instead over the relatively small area encompassed by NI, rare long-distances dispersal of badgers (Byrne *et al.*, 2014; Pope *et al.*, 2006) was enough to obscure the correlation between genetic and geographic distances. As badger samples were collected in a survey of badgers found dead on the roads (Abernethy *et al.*, 2011), another possibility is that these data include a higher than average proportion of badgers dispersing from their natal area, leading to underestimation of the degree of population structure present in NI badgers.

An landscape genetic analysis of the NI badger genotypes found no evidence that either

elevation or major rivers act as barriers to badger gene flow. As higher elevations have been shown to be less suitable for badger colonisation in NI (Reid *et al.*, 2011), and large rivers have been reported to act as a barrier to badger gene flow in GB (Frantz *et al.*, 2010b), these results are also unexpected. However, the relatively small spatial scale covered by NI (with respect to the dispersal distances covered by badgers), in combination with the limited heterogeneity of the NI landscape, may explain why the landscape features evaluated here were not identified as acting as a barrier to the genetic mixing of NI badgers. Further work may be beneficial to explore whether major roads show any effect on badger gene flow, especially given that sampling of badger genotypes was based heavily on badgers found dead on the roads.

### 3.5.3 Correlations between landscape features, badger population structure, and VNTR-types in cattle

If a spatial feature acts as a barrier or facilitator for the spread of a pathogen, it is expected that the presence of this feature would correlate (positively or negatively, respectively) with any discontinuities in the population structure of the pathogen. Using two measures of discontinuity in the spatial structure of *M. bovis* VNTR-types, both based on the results of the GAM models of probability of occurrence described above (Figure 3.14), we found similar results to those of the badger landscape genetic analysis: neither distance to high ground nor distance to the nearest river appear correlated with breaks in the VNTR-type population structure. Badger sett density was also tested, and likewise showed no significant correlation. Our models of VNTR-type probability of occurrence were based purely on the spatial autocorrelation present in the data, and it is possible that this smoothing process to some degree obscured the spatial signal then used to identify landscape correlations.

By contrast, there were indications that distance to the boundary between the two genetic populations of badgers (Figure 3.8B) was negatively correlated with discontinuities in the population structure of VNTR-types in cattle. The effect was clearly seen in the full dataset and in one part of the split data: in the second polygon a negative

effect was also apparent but confidence intervals included zero (Fig. 3.17). Similarly, genotyped badgers were more genetically dissimilar when a VNTR-type boundary was present between them. These results suggest that, where we detected significant genetic structure in the badger host, this correlates with the spatial distribution of *M. bovis* VNTR-types. If this is a true association, then it may indicate that transmission in badgers plays a role in the spread of *M. bovis* in these areas. If transmission in badgers is not important in determining the structure of the *M. bovis* population, then the only other way in which this association can arise would be if the transmission events responsible for determining the structure of the *M. bovis* population were somehow structured spatially in the same way that the badger population is structured in this region. If transmission of *M. bovis* occurs primarily through infected cattle, for example, then this could occur if contacts between cattle were interrupted in the same spatial are that the barrier between badger populations is located. Further work to compare the structure of *M. bovis* in NI with the cattle population structure and the cattle movement network in NI would be very valuable to explore alternative explanations for the association described here.

The relatively small spatial scale and relatively homogeneous landscape features present in NI mentioned above is likely to have reduced the ability of this study to detect the effects of landscape barriers on the *M. bovis* population, in addition to placing limitations on the genetic structure detectable within the badger population. Replication of this approach over a larger and more variable landscape, for example in Britain or across Ireland as a whole, would be informative and may enable the detection of further correlations. As mentioned above, further work to incorporate the effect of major roads on pathogen structure may also be beneficial.

### 3.5.4 Conclusions

In identifying strong spatial clustering and correlations between the *M. bovis* types in cattle and badgers, this study confirms previous findings suggestive of transmission of infection between cattle and badgers. This does not however allow inference of the

direction of transmission between the two hosts, although results presented here give indirect indication of transmission of infection from cattle to badgers.

This chapter looks further into the spatial structuring of the *M. bovis* population in NI, taking steps towards identifying the processes underlying the clustering of VNTR-types and the links between this and infection in badgers. I show that the boundaries between VNTR-type clusters do not correlate with the landscape predictors available for this study. These landscape predictors similarly did not explain badger population genetic structure, and in fact limited evidence for genetic structure was found in the NI badger population as a whole. However, where discontinuities in badger population structure were detectable, there were signs that these correlated with changes in the *M. bovis* VNTR-type communities present in cattle, indicating that the spread of *M. bovis* strains in cattle may be associated with the contact structure within the badger population.

Although very far from being conclusive, these results may give indication that the spread of disease within the badger population could be responsible for shaping the spatial pattern seen in *M. bovis* VNTR-types in NI cattle. However without further work to compare the *M. bovis* population to the structure present within the cattle industry in NI I cannot rule out alternate, cattle-related explanations for these findings. More generally, the approaches described here should be applicable to other multi-host pathogen systems where hosts differ in population and contact structure.

# CHAPTER 4

# Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis

# Use of bacterial whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis

## 4.1 Summary

High-density bacterial whole genome sequencing (WGS) has only recently become feasible for bacterial pathogens, and has much to offer to our understanding of their epidemiology. In this chapter, I apply WGS in a defined sub-population of *Mycobacterium bovis* isolates sampled from 145 cattle across 66 herd breakdowns, to gain insights into local spread and persistence. I show that despite low divergence among isolates, WGS can in principle expose contributions of under-sampled host populations to *M. bovis* transmission. However, I demonstrate that in these data such a signal is due to molecular type switching, which had been previously undocumented for *M. bovis*. Isolates from farms with a known history of direct cattle movement between them did not show a statistical signal of higher genetic similarity. Despite an overall signal of genetic isolation by distance, genetic distances also showed no apparent relationship with spatial distance among affected farms over distances of under 5km. Using simulations, I find that even over the brief evolutionary timescale covered by our data, Bayesian phylogeographic approaches are feasible. Applying such approaches showed that *M. bovis* dispersal in this system is heterogeneous but slow overall, averaging 2km/year. These results confirm that widespread application of WGS to *M. bovis* will bring novel and important insights into the dynamics of *M. bovis* spread and persistence, but that the current questions most pertinent to control will be best addressed using approaches

that more directly integrate WGS with additional epidemiological data.

## 4.2 Introduction

The increasing availability of bacterial whole-genome sequence (WGS) data now makes it possible to generate sequence datasets for whole bacterial pathogen populations at high sampling densities. Such comprehensive sequencing has yielded impressive advances in outbreak investigation (Eyre *et al.*, 2013; Harris *et al.*, 2010; Walker *et al.*, 2012), and provided new insights into both spatial dissemination (Gray *et al.*, 2011; Holden *et al.*, 2013) and the complexities of multi-host pathogen systems (Mather *et al.*, 2013; Viana *et al.*, 2014). However, even at the genomic scale the rates of evolutionary change estimated for bacteria can be substantially lower than those commonly seen in rapidly evolving pathogens such as RNA viruses (Biek *et al.*, 2015; Bryant *et al.*, 2013b). The extent to which slow evolution constrains the type and scale of epidemiological processes that can be resolved for bacterial pathogens, and which analytical approaches are most appropriate to deal with this, remains unclear for many systems.

*Mycobacterium bovis* is one of a group of closely related bacteria which includes the primary cause of human tuberculosis, *M. tuberculosis*, a pathogen estimated to evolve at a rate of around 0.3-0.5 mutations per genome per year over epidemiological timescales (Bryant *et al.*, 2013b; Walker *et al.*, 2012). *M. bovis* is the causative agent of bovine tuberculosis (bTB), an important disease of cattle and other mammals including man. Herd-to-herd movements of infected cows among farms (Gilbert *et al.*, 2005; Green *et al.*, 2008) and infection in Eurasian badger (*Meles meles*) populations (Delahay *et al.*, 2001; Gallagher & Clifton-Hadley, 2000) have both been implicated in the spread of bTB in Britain and Ireland. While much attention has focused on the relative roles of badgers and cattle in the maintenance of bTB, recent studies emphasising the importance of cattle have highlighted the continuing need for a deeper understanding of the role of cattle-based transmission (Brooks-Pollock *et al.*, 2014; Donnelly & Nouvellet, 2013).

Molecular typing of *M. bovis* isolates based on repeated genetic elements has been advocated for some time to aid in the epidemiology and control of bTB (Cousins *et al.*, 1998; Skuce & Neill, 2001), and in Britain and Ireland these typing methods have shown that *M. bovis* molecular types are maintained within well-defined geographic clusters (Skuce *et al.*, 2010; Smith *et al.*, 2006). While such molecular typing has proved useful for identifying local clustering on larger scales, their power to discriminate within-cluster events involved in fine-scale persistence and spread of bTB is limited.

In a previous study, Biek *et al.* (2012) established the potential of bacterial WGS in investigating the epidemiology of bTB at a local (i.e. farm-to-farm) scale. By sequencing 30 bacterial isolates from a spatially dense cluster of bTB cases within one recently emerged *M. bovis* molecular type (VNTR-10) in Northern Ireland (NI), the study demonstrated i) close relatedness of bacteria isolated from cattle and badgers, ii) persistence of bacterial lineages on the same farm, and iii) that genetic similarity between isolates correlated with geographic distance between sampling locations. The study also showed that, due to slow evolution, even WGS is unlikely to provide sufficient resolution to resolve transmission at the animal-to-animal scale for *M. bovis*, similar to findings in human tuberculosis (Bryant *et al.*, 2013b; Roetzer *et al.*, 2013; Walker *et al.*, 2012), and is more suited to do so at the between-farm scale.

While providing a proof of concept, this previous study was targeted towards a sub-sample of VNTR-10 infected cattle within a small (approx. 5km) spatial radius. This spatially restricted sampling precluded a more systematic investigation of processes occurring on a wider, population level scale within the bacterial strain, including the potential identification of under-sampled reservoirs, the rate and mode of spatial spread, and transmission links between bTB breakdowns. A herd breakdown is defined as the period during which movements of cattle out of a herd are restricted due to the detection of bTB in the herd, starting at the detection of one or more infected animals (either through the tuberculin skin test or through abattoir surveillance for bTB lesions), and ending when the herd has undergone two consecutive negative whole-herd tests at least 60 days apart, or a single negative test where the breakdown was not lab-

oratory confirmed (laboratory culture is performed on all post mortem samples from cattle testing positive for bTB in NI).

Here, we extend the analysis of the Biek *et al.* (2012) study by examining WGS data from all 145 available VNTR-10 isolates in NI since 2003. In NI all cattle herds are tested for bTB on an annual basis, and for over a decade *M. bovis* isolates cultured from test-positive cattle have been extensively typed and stored. These archived isolates therefore gave us the opportunity to target a genetically defined sub-population of *M. bovis* (VNTR-10) for high-density sampling with respect to cattle infections, although VNTR-10 infections in any other population would not be accessible through this sampling strategy. To gain insights into the mode and rates of transmission, we used intensive sampling and WGS of *M. bovis* isolates from cattle to address the following questions:

- Does WGS of VNTR-10 isolates from cattle indicate contributions from another host population which is under-sampled under the above sequencing strategy?
- Does the genetic relatedness between sequenced isolates correlate with recorded movements and/or with spatial distance between premises?
- What are the rate and mode of *M. bovis* dispersal across the landscape at the between-breakdown scale as determined by WGS?

## 4.3 Materials and Methods

### 4.3.1 Molecular-typing of *M. bovis* in NI

In NI, *M. bovis* isolates have been stored and typed since the early 2000s using spoligotyping, more recently combined with Variable Number Tandem Repeat (VNTR) typing, to differentiate molecular types (Skuce *et al.*, 2005). Spoligotyping gives a relatively coarse-grained discrimination of the *M. bovis* population, and is based on the

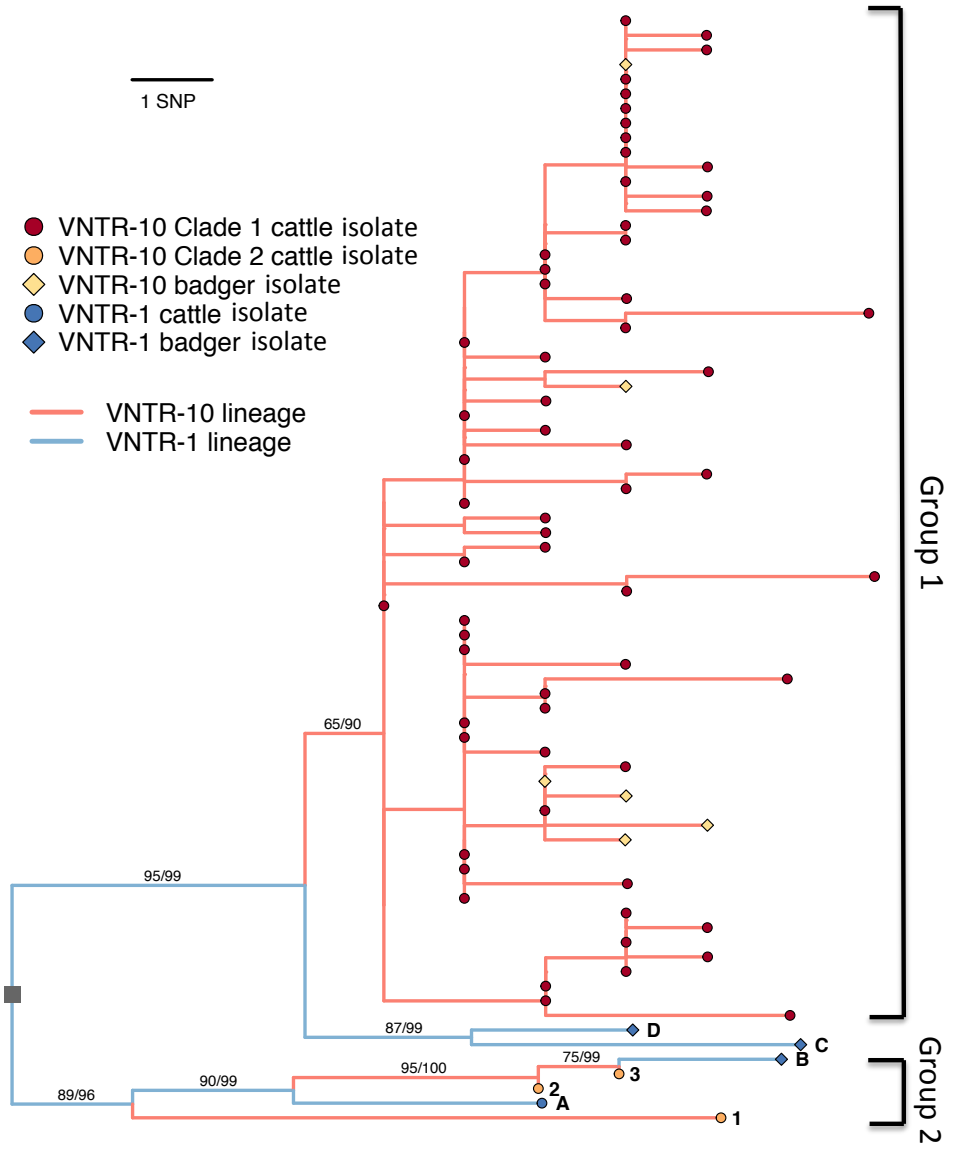


**Figure 4.1: Maximum Likelihood phylogeny of VNTR-1 and -10 isolates** - subsampled to one sequence per outbreak and rooted on the VNTR-4 isolate and *M. bovis* reference sequence (Garnier *et al.*, 2003) (node used to root the phylogeny is indicated by a grey square). Tip colours give details of the isolates: red circles are Group 1 VNTR-10 cattle isolates, orange circles (numbers 1-3) are Group 2 VNTR-10 cattle isolates; yellow diamonds are VNTR-10 badger isolates; blue circle (A) is the VNTR-1 cattle isolate, blue diamonds (B-D) are VNTR-1 badger isolates. Branch colours give the likely VNTR-type of each branch, assuming the most recent common ancestor of the group was VNTR-1. Branch labels show the statistical support for selected nodes: the left-hand value indicates percentage bootstrap support from a Maximum Likelihood phylogeny generated for these isolates, and the right-hand value shows posterior probability of the node in the Bayesian phylogeny generated for these isolates.

presence or absence of multiple spacer oligonucleotides within the Direct Repeat region of the genome (Kamerbeek *et al.*, 1997). VNTR-typing indexes the number of short nucleotide repeats present at several VNTR loci identified within the mycobacterial genome (Mazars *et al.*, 2001), and provides greater discriminatory power than spoligotyping alone, although with a relatively higher chance of homoplasy (i.e. separate lineages converging on the same molecular type). In NI between 2003-2008 one *M. bovis* isolate was VNTR-typed and stored from each herd breakdown for which *M. bovis* was isolated using a panel of 7 VNTR loci optimised for this population of *M. bovis* (Skuce *et al.*, 2010), while from 2009 onwards bacteria were VNTR-typed and stored from all culturable cattle cases, therefore resulting in more intensive sampling in recent years. In addition to these cattle isolates, *M. bovis* has also been typed by spoligotyping and VNTR-typing and archived when isolated from a survey of badgers killed on the roads in NI (Abernethy *et al.*, 2011).

### 4.3.2 Bacterial isolates and sequencing

Cultures of *M. bovis* were isolated from bovine granulomatous tissue using conventional methods (Skuce *et al.*, 2010) by the Agri-Food and Biosciences Institute (AFBI) in NI. Confirmed isolates were grown on LJ slopes to single colonies, following which single colonies were grown up and DNA was extracted using standard CTAB and solvent extraction protocols (van Soolingen *et al.*, 2001), also by AFBI. A total of 144 VNTR-10 *M. bovis* isolates were included in this study, from 66 herd breakdowns (see earlier definition) occurring in 51 herds between 1996 and 2011. The WGS dataset consisted of the raw reads from 31 VNTR-10 isolate originally sequenced in the preceding study (26 cattle and 5 badger isolates; Biek *et al.*, 2012), in addition to 114 VNTR-10 isolates (113 cattle and 1 badger isolate) sequenced for the first time in this study. VNTR-10 is located predominantly within the Newtownards area of NI (178/195 VNTR-10 infections recorded between 1996-2011 were from the Newtownards District Veterinary Office).

To provide broader evolutionary context, we additionally sequenced five isolates from

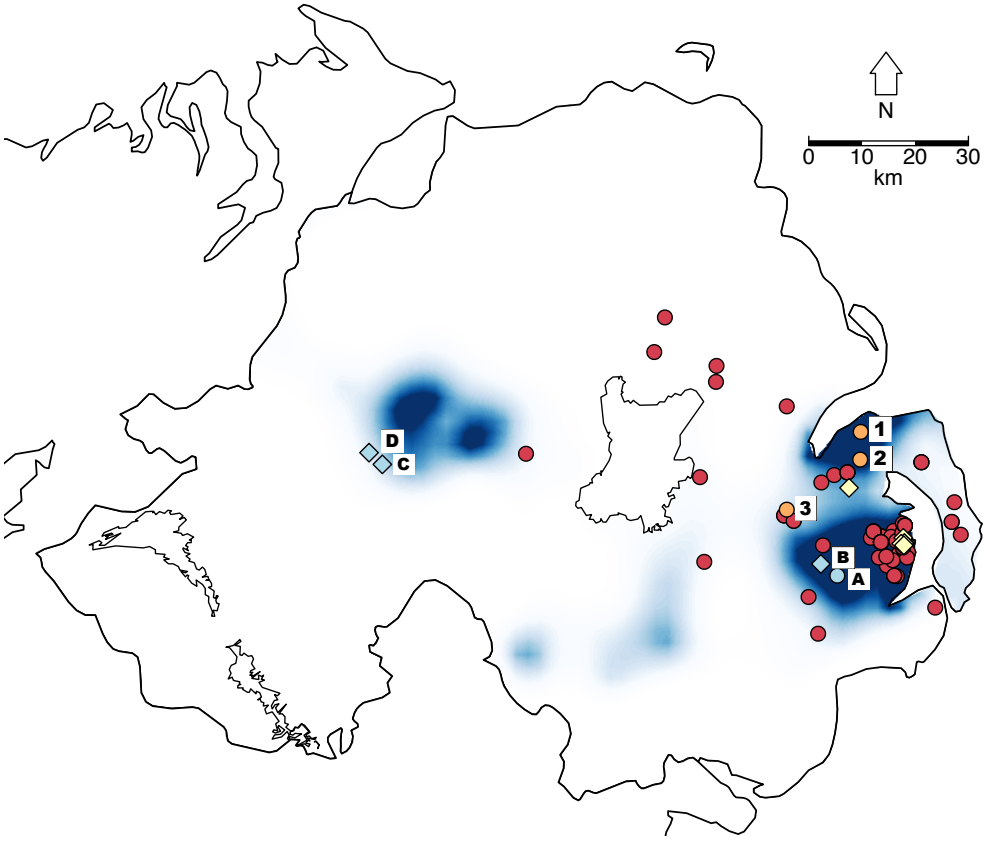


**Figure 4.2: Map of Northern Ireland showing origins of sequenced isolates** - Red circles are Group 1 VNTR-10 cattle isolates; orange circles (isolates 1-3) are Group 2 VNTR-10 cattle isolates, yellow diamonds are VNTR-10 badger isolates, light blue diamonds (badger) and light blue circle (cattle) (isolates A-D) are VNTR-1 isolates, and blue shading indicates the probability of occurrence for the VNTR-1molecular type in cattle, as estimated in Chapter 3.

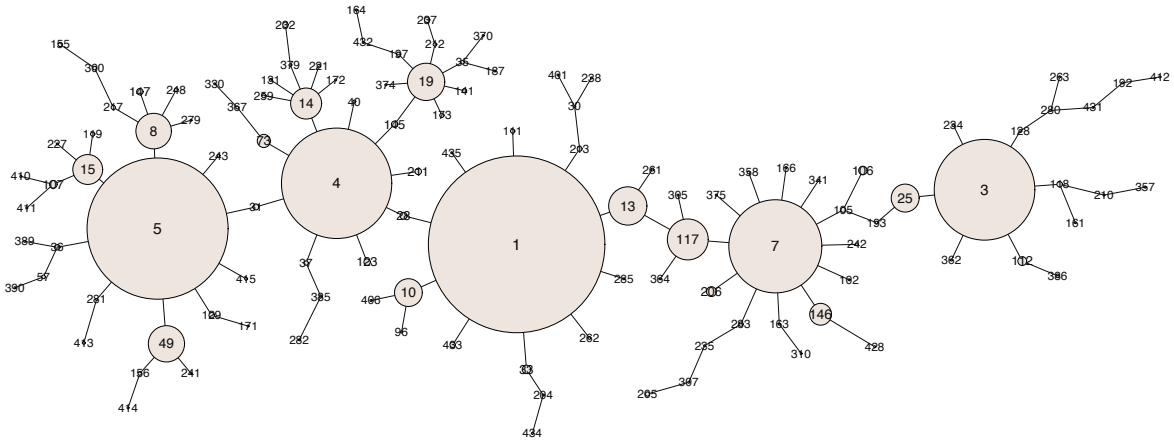


**Figure 4.3: Minimum spanning tree showing relationships between Northern Irish VNTR-types belonging to spoligotype SB0140 (the most common spoligotype in NI)** - generated using the R package Pegas (Paradis, 2010). Circles are scaled proportionate to the number of times the VNTR-type was recorded.

VNTR-types thought to be ancestral to VNTR-10, namely four isolates of VNTR-1 (1 cattle isolate (isolate A), and 3 badger isolates (isolates B-D, Figures 4.1 and 4.2), and one cattle isolate of VNTR-4. VNTR-1 is thought to be the direct ancestor of VNTR-10 based on the following observations: VNTRs -1 and -10 are separated by a single tandem repeat difference; VNTR-1 has been recorded at a high and approximately stable prevalence in NI since routine VNTR-typing commenced (Skuce *et al.*, 2010, 2005), whereas VNTR-10 has been found in low but increasing numbers suggestive of a newly emerged strain; VNTR-1 is found across a wider spatial range than VNTR-10 (Figure 4.2); and a minimum spanning tree of all NI VNTR-types within spoligotype SB0140 shows that VNTR-1 is basal compared to VNTR-10 (Figure 4.3). The VNTR-4 type differs from VNTR-1 by two tandem repeats (Table 4.1), however it is not possible to determine which of them is ancestral. Sequencing was performed by Glasgow Polyomics using the Illumina IIx platform, with the exception of VNTR-1 isolates B-D, sequenced using an Illumina MiSeq.

**Table 4.1: VNTR genotype strings for NI VNTR-types 1, 4 and 10** - shows the number of tandem repeats present at the seven VNTR-loci used in NI

| | MV2163B QUB11B | MV4052 QUB26A | MV2461 ETR B | MV1895 QUB1895 | MV2165 ETR A | MV2163 QUB11A | MV323 QUB3232 |
|---|---|---|---|---|---|---|---|
| VNTR-4 | 4 | 4 | 5 | 4 | 7 | 11 | 7 |
| VNTR-1 | 4 | 4 | 5 | 4 | 7 | 11 | 9 |
| VNTR-10 | 3 | 4 | 5 | 4 | 7 | 11 | 9 |

### 4.3.3 Bioinformatics and SNP calling

Paired end raw reads from 31 *M. bovis* isolates sequenced for the previous study (Biek *et al.*, 2012; Batch 1) were combined with the 116 *M. bovis* isolates sequenced separately as paired ends for this study using an Illumina IIx (Batch 2) and three *M. bovis* VNTR-1 isolates (isolates B-D, Batch 3), also paired ends, sequenced using an Illumina MiSeq. These isolates were processed using the following workflow:

Reads were trimmed based on quality using ConDeTri (Smeds & Kunstner, 2011), then BWA v0.7.5a (Li & Durbin, 2009) was used at the default settings to assemble

the reads to the *M. bovis* reference genome AF2122/97 (GenBank accession number BX248333; Garnier *et al.*, 2003). SAMtools mpileup and bcftools were then applied to identify variant sites within the assembly (Li *et al.*, 2009).

Code was written in Python to filter the raw variants identified above according to: total depth of coverage; depth of high quality coverage (according to the SAMtools vcf file output) on the forward and reverse strands; mapping quality; and percentage of high quality bases that give the dominant allele (*M. bovis* is a slowly mutating, haploid organism, and the sequenced isolates are grown from pure cultures, therefore true variant sites would be expected to show minimal heterozygosity).

The filter code was run using two settings, giving a strict filter and a relaxed filter (Table 4.2). The percentage of high quality bases giving the dominant allele (homozygosity) was kept at 95% for both filters, as lower values appeared to generate homoplasy (see below). The *M. bovis* population as a whole shows very little genetic diversity, especially in Britain and Ireland (Smith *et al.*, 2006), and these isolates represent only one closely-related group within that limited diversity. Therefore, across the whole genome, true biological variation is expected to be rare and any variant identified has a high chance of being the result of sequencing error, and so for our purposes SNP calling needed to be as stringent as possible. Initially the strict criteria, with high filter thresholds, was run to identify sites across the genome where at least one isolate contained a high quality variant. Once these sites containing high quality SNPs had been identified, variants identified at these sites in other isolates have an increased likelihood of being true biological variants, and accordingly, the more relaxed filter with lower thresholds was then used to determine whether each of the isolates pass or fail at each high-quality variant site, and if pass, whether reference or SNP.

After the filtering code was run and sites containing a high quality SNP in at least one isolate were identified, these sites were further filtered and removed if:

- They clustered together with another variant site less than 200bp apart (if mutations were distributed evenly through the genome, we would expect a SNP

**Table 4.2: Filter criteria used to identify variant sites for NI isolates**

| | Batch 1 | | Batch 2 | | Batch 3 | |
|---|---|---|---|---|---|---|
| | Strict | Relaxed | Strict | Relaxed | Strict | Relaxed |
| Total depth of coverage | 50 | 14 | 25 | 7 | 25 | 5 |
| High quality depth of coverage on each strand | 6 | 2 | 2 | 1 | 2 | 1 |
| Mapping quality | 35 | 20 | 35 | 20 | 40 | 35 |
| Homozygosity | 95% | 95% | 95% | 95% | 95% | 95% |

every 14000bp). Clustering of variants may suggest an area of the genome prone to sequencing or mapping error, or alternatively, clustered SNPs might indicate areas of recombination or selection. However, either of these explanations would confound our downstream analysis and therefore these SNPs were removed from the final dataset.

- They were in a repeat region of the genome, as identified previously by artificially fragmenting the reference genome, and using BLAST to identify regions receiving more than the expected number of BLAST hits (Biek *et al.*, 2012).
- More than 10% of isolates failed the relaxed filter criteria at that site.

The pass-reference/pass-variant/fail information given by the relaxed filter was then used to generate sequences of concatenated variant sites for each of the isolates (sites failing the filter were denoted N).

The ideal variant filter thresholds are those that remove all error variants while leaving as many of the true variant sites as possible. Given the clonality of *M. bovis* bacteria (Hershberg *et al.*, 2008; Pepperell *et al.*, 2013; Wirth *et al.*, 2008), all true SNP sites should support one phylogenetic tree, and therefore any homoplasies would indicate that error variants are present in the data. Using this, the filter criteria were sequentially lowered and the optimum filter thresholds (Table 4.2) were determined to be those just high enough to prevent homoplasies. Batch 1 isolates generated considerably more reads, and therefore a higher depth of coverage, than Batches 2 and 3. For

this reason, the total depth and the high quality depth filters were increased for Batch 1 isolates, to the point where they generated on average the same number of variant sites per isolate as the filter criteria used for Batches 2 and 3.

Post assembly, the mean depth of coverage of the *M. bovis* genome was 108 reads for Batch 1 isolates, 53 reads for Batch 2 isolates and 45 for Batch 3, with more than 97% of the genome covered by at least one high quality base in all batches. 335 sites that varied from the reference genome were identified by the strict filter criteria, and 290 of these survived all the filtering steps and were used for downstream analyses; 41 of these sites were informative to the VNTR1/10 group (i.e. shared by two or more isolates within this group).

### 4.3.4 Genetic analysis

A Maximum Likelihood phylogeny was generated in PhyML v3.0 (Guindon & Gascuel, 2003) under the Jukes Cantor model of nucleotide substitution (Jukes & Cantor, 1969), including the *M. bovis* reference sequence as outgroup, and evaluating statistical support for individual nodes based on 1000 non-parametric bootstraps. The Jukes Cantor model represents the most basic model of nucleotide substitution, assuming equal frequencies of different bases and equal mutation rates between different bases. I chose to use it for this analysis in the first instance because, with such a closely related group of bacteria generating so few mutations across the (4.3 megabase) genome, I reasoned that the simplest substitution model would be appropriate. However, formal model selection (jModelTest, results not shown) identified another nucleotide substitution model, the HKY model which differentiates between transitions and transversions between bases and allows unequal base frequencies, as the best fitting model for these data. Although branch lengths do differ slightly for the VNTR-10 phylogeny generated under the HKY model, the overall topology remains the same and the change of model does not significantly affect the conclusions drawn in this chapter (results not shown), likely as a result of the closely related nature of this group of sequences. Results reported below refer to the Jukes Cantor model.

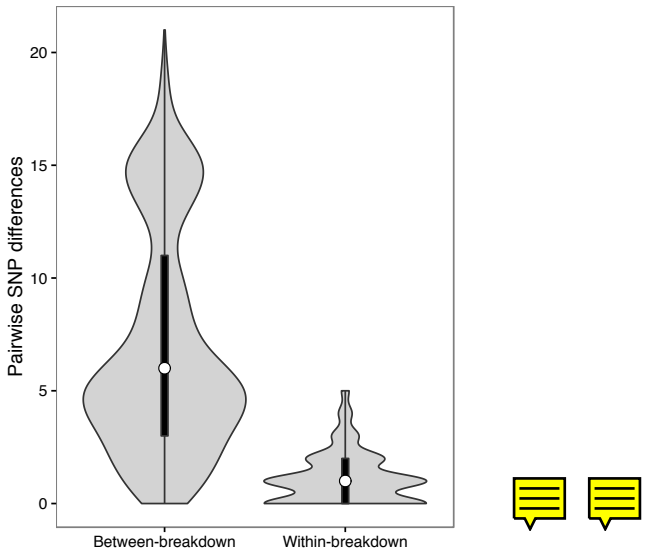


**Figure 4.4: Comparison of pairwise SNP differences between sequences originating from the same breakdown and from different breakdown, for VNTR-10 isolates** - Comparisons are shown for VNTR-10 isolates, and pairwise differences from the same breakdown are labelled "within breakdown" and those from different breakdowns are labelled "between breakdown".

A Bayesian phylogeny was generated in MrBayes (Ronquist & Huelsenbeck, 2003) under the Jukes Cantor model, also including the *M. bovis* reference sequence, and was run for $10^6$ MCMC iterations at which point the standard deviation of split frequencies was below 0.01. Raw pairwise SNP differences between sequenced isolates were calculated in MEGA 5 (Tamura *et al.*, 2011), using pairwise deletions in the event of missing data. Due to the increased level of sampling from 2009 onwards described above, and the low levels of within-breakdown diversity (see Results and Figure 4.4), further analyses were restricted to one representative isolate per herd breakdown.

### 4.3.5 Comparing genetic and epidemiological relationships between breakdowns

In NI, detailed information on the cattle population and movements between herds and bTB test results is recorded by the Department of Agriculture and Rural Development (Houston, 2001). All direct cattle movements between herds with VNTR-10 isolates represented in our dataset (66 sequenced breakdowns and 12793 individual

cattle movements) were made available, and were combined with the location and date of each sequenced isolate using SQL (MySQL v5.5.29; www.mysql.com). All further analyses were conducted in R v3.0.1 (R Core Team & R Develoment Core Team, 2014) unless otherwise stated.

A Mantel test was conducted in the R package ecodist (Goslee & Urban, 2007) to assess correlation between spatial and genetic distance between Group 1 breakdowns, using 10000 permutations to assess significance (see results and Figure 4.1 for definition of Group 1). To further confirm this correlation, a multiple regression on distance matrices was also carried out using the R package ecodist (Lichstein, 2007), assessing the correlation between the matrix of pairwise SNP differences and the matrix of geographic distances between Group 1 breakdowns, and using 10000 permutations to assess significance.

For each sequenced Group 1 breakdown, we recorded whether it was linked to one or more other sequenced breakdowns by potentially infectious cattle moving directly into the original breakdown herd (movement links). I considered movements occurring within 2, 5, and 10 years prior to the official start of each breakdown. Few direct movement links were identified within the 2- and 5-year windows (15 and 53 links, compared to 102 for the 10-year window). I therefore included all movements identified within the most conservative window of 10 years in this analysis. For each sequenced breakdown, I identified any movement links to that breakdown from any other sequenced breakdown. Where links were present, I recorded the minimum pairwise SNP difference between the linked premises, as the minimum SNP difference is most likely to represent a direct transmission event if one exists between breakdowns.

To assess the effect of short-distance transmission mechanisms in a more targeted manner, I identified pairs of sequenced herds located within 2km and 5km of another sequenced herd. This choice was motivated by the observation that the vast majority of badger movements fall within 5km, although larger distances are occasionally recorded (Byrne *et al.*, 2014; Pope *et al.*, 2006), although transmission over short distances could be driven by various mechanisms (including contact with infected cattle in neighbouring

herds) and is not necessarily restricted to badgers. For each sequenced breakdown, I identified the presence of other sequenced breakdowns linked through spatial proximity at these distances, and again recorded the minimum number of SNPs separating it from other spatially-linked breakdowns.

A higher number of epidemiological links between breakdowns is likely to result in a lower minimum SNP difference between breakdowns due to the increased number of comparisons, and therefore the distributions of minimum SNP differences are not directly comparable between different types of links. To statistically assess the significance of the association between epidemiological links and minimum SNP differences, the distribution of the expected number of SNP differences was simulated for movement and proximity links under a null hypothesis of no association between the presence of a link and genetic relatedness between breakdowns. The number of links to each sequenced breakdown, identified above, was kept but these links were effectively rewired by permuting the matrix of linked outbreaks and thus randomising the breakdowns that each sequenced breakdown was linked to, and the minimum SNP difference was again calculated for each breakdown. This was repeated 10,000 times for each type of link (direct recorded movements within 10 years of breakdowns, and spatial proximities of 2km and 5km) to generate null distributions for comparison to the observed distributions.

To formally compare the null simulations with the observed distribution of minimum SNP differences between linked outbreaks, a goodness of fit test was carried out. Ten thousand realisations of a multinomial distribution were generated, with sample size equal to the number of linked breakdowns, and the probability of each category proportional to the expected value of each category (taken as the mean of the values from the null simulations). A Chi-squared test statistic was then calculated for each of these multinomial distributions. This gave a distribution of simulated test statistics to which the Chi-squared statistic calculated from the observed data was compared. A p-value was estimated as the proportion of the null Chi-squared statistics that were greater than the observed Chi-squared value.

### 4.3.6 Bayesian phylogenetic parameters

Phylogenetic and phylogeographic analyses were conducted in the Bayesian phylogenetic program BEAST v1.7.4 (Drummond *et al.*, 2012), using the Jukes Cantor model of nucleotide substitution and the Bayesian skyline model as demographic coalescent prior (Drummond *et al.*, 2005). The prior for the molecular clock rate was given as a normal distribution with the mean corresponding to a rate estimate for human tuberculosis (0.5 SNPs per genome per year; Walker *et al.*, 2012), but with a wider standard deviation (more than 10 SNPs per genome per year) to account for the fact that we are investigating a different bacterial system. Convergence was determined as effective sample sizes (ESS) of greater than 200 for all parameters, as determined in Tracer (v1.5, http://beast.bio.ed.ac.uk/software/tracer/). Log marginal likelihood estimates (MLEs) were calculated to assess model fit using path-sampling and stepping-stone sampling (Baele *et al.*, 2012), run twice to assess MLE convergence.

In order to ascertain the correct clock model to use for the Bayesian phylogeographic analysis, BEAST was used to fit and compare three different molecular clock models: a strict clock model and two relaxed clock models allowing evolutionary rates to differ among branches, with rates drawn from log-normal or exponential distributions (Drummond *et al.*, 2006). The analyses were run for 500 million iterations of the MCMC chain, resulting in convergence.

Log MLEs for the strict molecular clock and the relaxed lognormal clock models were almost identical (-1504.29 and -1504.85 respectively), and both models were slightly preferred over the relaxed exponential clock (log marginal likelihood estimate: 1508.22). These results indicate that a strict molecular clock model is appropriate for these data, and therefore the strict model was used for subsequent BEAST analyses in combination with the other parameter values specified above.

To assess whether the priors chosen for BEAST analyses shaped the results for these analyses, we ran BEAST for 1010 MCMC iterations, using the same settings described above but sampling from the prior distributions only. However, these runs failed to

converge after 500 million iterations of the MCMC chain (combined ESS value of 31 for the tree posterior).

### 4.3.7 Phylogeographic inference

To quantify the spread of *M. bovis* across the landscape, continuous phylogeographic models were applied in BEAST (Lemey *et al.*, 2010). This analysis was restricted to the VNTR-10 clade containing the majority of isolates (Group 1, see Results and Figure 4.1), using one sequence per breakdown (chosen as the sequence from that herd breakdown that was closest to the root of the tree, as this was likely to be the closest representative of the original bacteria that infected the herd). A strict Brownian model of spatial diffusion was compared to a relaxed model allowing diffusion rates to vary among branches, with rates drawn from a Cauchy distribution. A relaxed model with branch rates drawn from a gamma distribution was also tested but failed to converge. Models were run for 500 million iterations, assessed for convergence in Tracer, and model fit was again evaluated using log MLE values. Posterior trees for the best fitting model were combined to find and annotate the Maximum Clade Credibility (MCC) tree. Node locations, branch lengths, and branch-specific rates of geographic dispersal were extracted and evaluated for the MCC tree.

Given that the molecular clock rate of *M. bovis* and other closely related mycobacteria has been shown to be slow and variable (Biek *et al.*, 2012; Bryant *et al.*, 2013b), it was uncertain whether these data would contain enough genetic signal to accommodate phylogeographic analyses. To test this, we simulated a homogeneous spatial diffusion process along the MCC phylogeny generated above, guided by empirical rates, generating a set of spatial coordinates for sampled sequences under a set rate of spatial diffusion along the existing phylogeny. We then evaluated whether phylogeographic analysis in BEAST, using the settings described above and with the simulated coordinates and observed sequences and sampling dates as input, could recover the originally specified diffusion rate for each of 100 simulations.

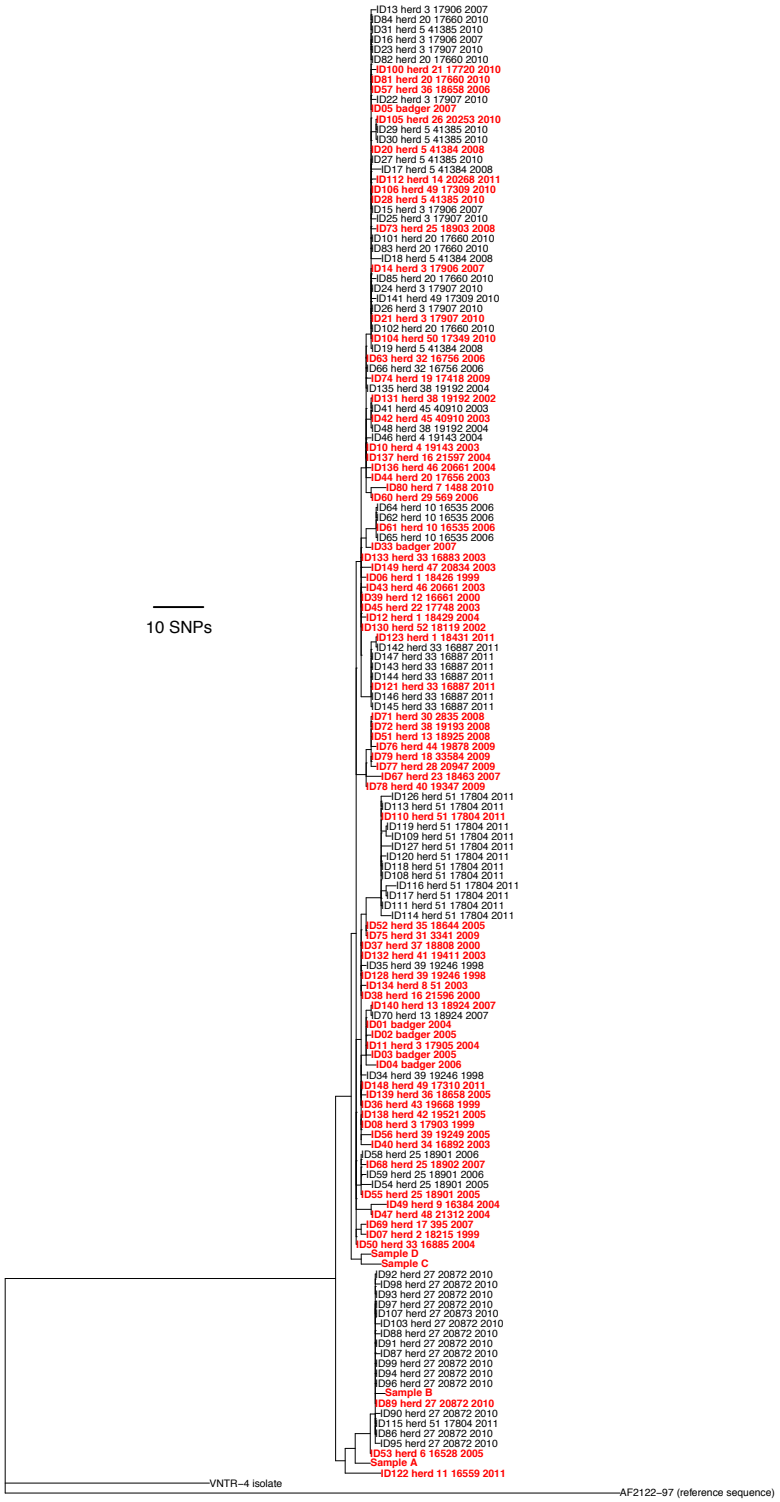


**Figure 4.5: Maximum Likelihood phylogeny of all isolates sequenced in this study** - ML phylogeny of all isolates described in the study, rooted on the *M. bovis* reference sequence AF2122/97 and the VNTR-4 isolate. Phylogeny generated in PhyML v3.0 (Guindon & Gascuel, 2003) using the Jukes-Cantor model of nucleotide substitution. Isolates included in the downstream analyses are shown in red bold font. Tip labels give epidemiological information in the following order: isolate reference, anonymised herd ID, anonymised breakdown ID, year of sampling.

## 4.4 Results

### 4.4.1 VNTR-10 isolates

Among the intensively sampled VNTR-10 isolates, genetic divergence was low, averaging 6.4 SNPs (range 0-19) over the whole group of 145 sequenced VNTR-10 isolates. Despite limited divergence, the group contained 39 shared polymorphisms resulting in a non-star-like phylogenetic structure (Figure 4.1, showing one isolate per herd breakdown).

Regressing genetic distance from the root of the phylogeny against sampling date revealed a moderate positive correlation ($R^2 = 0.32$), indicative of a molecular clock signal within these data (Firth *et al.*, 2010).

Average diversity for sequenced isolates from within the same herd breakdowns (including two breakdowns which were polyphyletic) was low, with mean 0.69 SNPs and range 0-4 SNPs. This was considerably lower than the average minimum SNP differences between different VNTR-10 breakdowns (mean 4.73 SNPs, range 0-17 SNPs). Multiple isolates per breakdown were only available for 19 of the 66 VNTR-10 breakdowns. Given the low level of within-breakdown divergence and the fact that multiple isolates per breakdown were only available after 2008 (see Materials and Methods), we chose to use one sequence per herd breakdown for further analysis in order to focus on the dynamics of *M. bovis* spread at the between-breakdown scale. The full phylogeny including all isolates described here is shown in Figure 4.5.

Both Bayesian and Maximum Likelihood phylogenies indicated that the majority of VNTR-10 sequences fell into a single main group, hereafter referred to as Group 1, including 126 isolates from 63 breakdowns, and including the six badger isolates (Figure 4.1). Nineteen other VNTR-10 isolates, belonging to three individual breakdowns (represented by isolates 1, 2 and 3, Figure 4.1) were not positioned within Group 1 and instead clustered with two of the VNTR-1 isolates (isolates A and B, Figure 4.1 Group 2). Both Group 1 and Group 2 isolates were defined by multiple unique SNPs, however

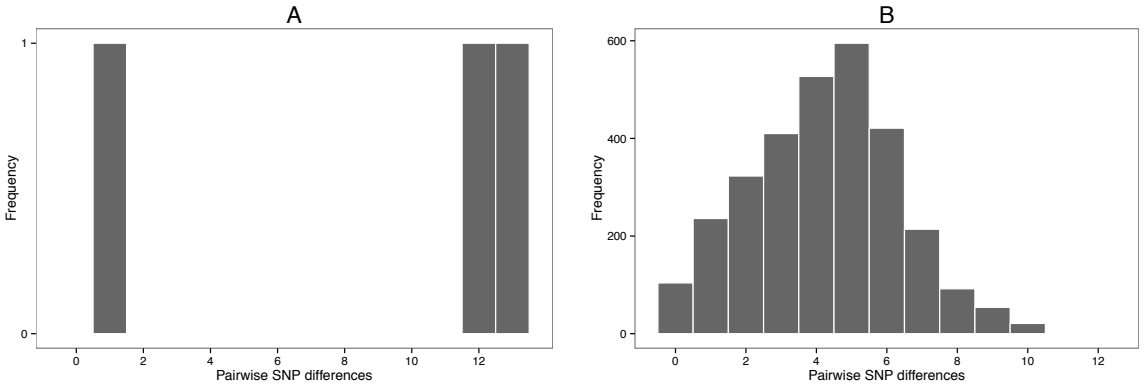


**Figure 4.6: True pairwise genetic distances for Clade 2 sequences (A), and pairwise distances generated from 1000 subsamples of Clade 1 sequences, weighting for sampling year (B)**

statistical support for the nodes defining these two lineages was not consistently high (bootstraps of 65 and 89 respectively, with posterior probabilities of 90 and 96, Figure 4.1).

Despite the much larger number of Group 1 isolates, these two sets of VNTR-10 isolates were similar in terms of maximum genetic divergence among isolates (Group 1 isolates: 15 SNPs; Group 2 isolates: 13 SNPs). Within Group 1, the average pairwise distance was low (mean: 4.4 SNPs), many isolates were genetically indistinguishable, and the majority of mutational steps were represented by one or more sequenced isolates. These observations are indicative of a population that has been comprehensively sampled, consistent with our expectations based on sampling all VNTR-10 outbreaks detected over the study period. In contrast, sequences from the Group 2 VNTR-10 breakdowns were non-identical, more divergent from each other (mean: 8.5 SNPs) and many of the intermediate mutational steps were not represented in the sequenced isolates. This suggests that comprehensive sampling of the lineage represented by these isolates had not been achieved.

To evaluate whether these differences might simply be due to sampling, 1000 random subsamples of three sequences (matching the number of Group 2 VNTR-10 breakdowns) were generated from Group 1 isolates. The observed pairwise differences between Group 2 VNTR-10 isolates (1, 12 and 13 SNP differences) were at the upper end

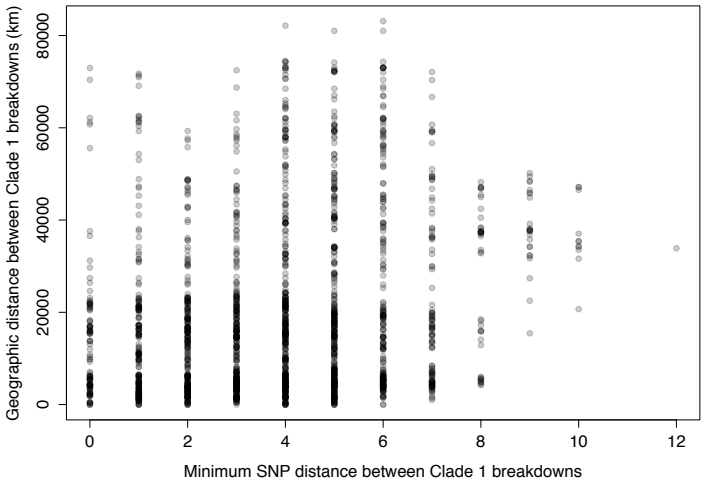


**Figure 4.7: Comparison of pairwise genetic and spatial distances between all Clade 1 herd breakdowns**

of the simulated distribution of pairwise distances among subsamples (mean 4.16 SNPs, range 0-12 SNPs; Figure 4.6). This suggests that Group 2 isolates were sampled from a bacterial population that is at least as large as, or larger than, Group 1 (assuming no significant differences in evolutionary rates between the two lineages), but that a lower proportion of this population has been sampled and sequenced compared to Group 1 despite the consistently high sampling effort for VNTR-10 outbreaks in cattle.

Given that the Group 2 isolates appeared to be drawn from a population with different characteristics compared to Group 1, and given that these comprised only three sampled VNTR-10 breakdowns, we restricted all further analyses to Group 1 isolates unless otherwise indicated.

### 4.4.2 VNTR-types 4 and 1

As anticipated, the single VNTR-4 isolate was genetically highly distinct from the VNTR-10 group, differing by an average of 89 SNPs (Figure 4.5). In contrast, VNTR-1 isolates failed to form a separate clade, as would have been expected for a separate VNTR-type, and instead were found to be nested within the VNTR-10 group. The two VNTR-1 isolates originating in the same geographical area as the majority of the

VNTR-10 isolates (isolates A and B, Figure 4.2) were found to cluster with VNTR-10 Group 2 isolates (Figure 4.1), while those originating from outside the VNTR-10 range formed a sister group to VNTR-10 Group 1 (Figure 4.1, isolates C and D, see also Badger isolates below).

### 4.4.3 Badger isolates

As discussed by Biek *et al.* (2012) using a subset of the data described here, VNTR-10 *M. bovis* isolates from badgers and cattle were highly similar genetically, with a minimum distance of 0-3 SNPs to the most closely related cattle isolate (Figure 4.1), suggestive of recent transmission links between badger and cattle. While the current study included only one additional VNTR-10 badger isolate, our high density sampling of VNTR-10 cattle infections also showed that the cattle isolates most closely related to those from badgers were all found within very close spatial proximity (less than 1.5km) to the locations of these badger isolates.

Of the three newly sequenced VNTR-1 badger isolates, two isolates (C and D) originated from an area outside the distributional range of VNTR-10, and approximately 100km from the area where the majority of VNTR-10 isolates were located (Figure 4.2). These two isolates were closest to VNTR-10 Group 1 sequences, separated by a minimum genetic distance of 5 SNPs (Figure 4.1). As described above, the other VNTR-1 badger isolate (isolate B) originated from the same area as the majority of the VNTR-10 isolates (Figure 4.2), and clustered with Group 2 VNTR-10 isolates (Figure 4.1)

### 4.4.4 Molecular clock rate

Bayesian evolutionary analysis revealed an evolutionary rate of 0.2 substitutions per genome per year (95% HPD: 0.1-0.3), estimating the time of the most recent common ancestor (tMRCA; the estimated date at which the progenitor of a group of sequences

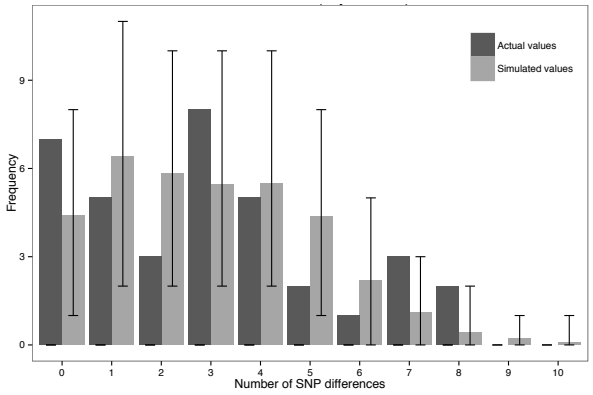


**Figure 4.8: Comparison of observed and simulated SNP differences for movement links** - Observed number of SNP differences between outbreaks linked by movements of cattle within a 10-year timeframe shown in dark grey, and expected SNP differences from $10^4$ simulations of the null hypothesis of no association between presence of a link and genetic similarity are in light grey. Bars show the intervals containing 95% of the results from the null simulations.

existed) of the VNTR-10 group as a whole as 1974 (95% HPD: 1954-1989), and the time of the most recent common ancestor of Group 1 at 1988 (95% HPD: 1979-1995) and Group 2 at 1984 (95% HPD: 1968-1998). Our evolutionary rate estimate is lower than some estimates for *M. tuberculosis* in humans, (95% CIs: 0.3-0.7 substitutions per genome per yr; Roetzer *et al.*, 2013; Walker *et al.*, 2012), whereas other studies reported similar rates (0.13-0.41; Bryant *et al.*, 2013b). While differences in pathogen life history, such as disease latency (Colangeli *et al.*, 2014) or non-replicating persistence of bacteria in the environment (Courtenay *et al.*, 2006; Maddock, 1933; Young *et al.*, 2005), might contribute to rate variation among different myobacteria, further studies will be required to fully explore this possibility.

### 4.4.5 Comparing genetic and epidemiological relationships between breakdowns

A Mantel test showed a significant, though weak, association between the genetic and spatial distances of Group 1 breakdowns (p=0.014, Spearmans rank coefficient=0.20; Figure 4.7), as did multiple regression on matrices (p=0.034, $R^2$=0.031).

The minimum SNP differences observed between pairs of sequenced Group 1 break-

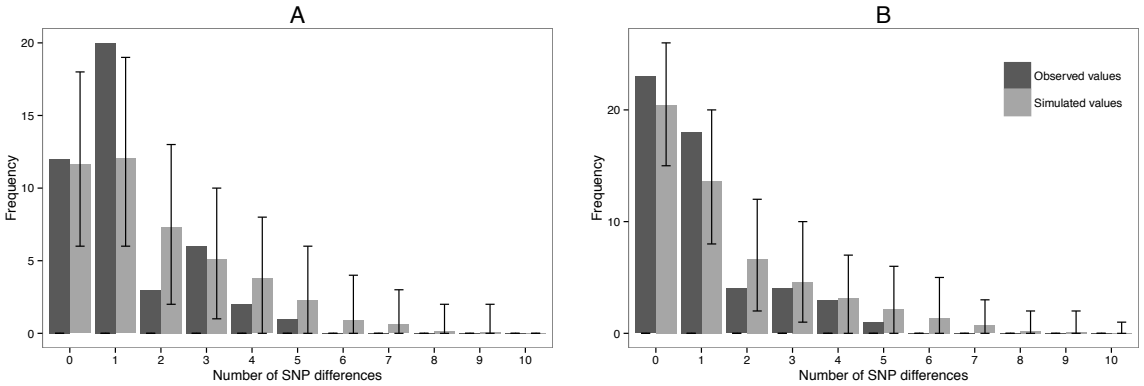


**Figure 4.9: Comparison of observed and simulated SNP differences for spatial links of 2km (A) and 5km (B)** - Observed number of SNP differences between outbreaks linked by movements of cattle within a 10-year timeframe shown in dark grey, and expected SNP differences from $10^4$ simulations of the null hypothesis of no association between presence of a link and genetic similarity are in light grey. Bars show the intervals containing 95% of the results from the null simulations.

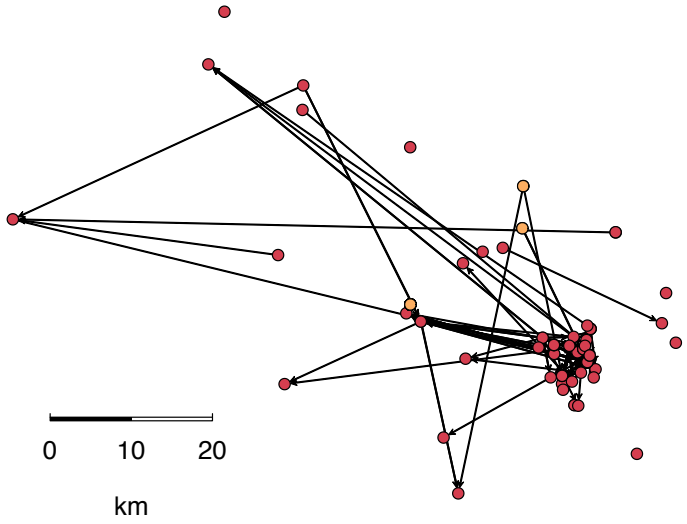


**Figure 4.10: Map showing cattle movement links (black arrows) and spatial relations between cattle premises with VNTR-10 isolates sequenced in this study** - Clade 1 premises shown in red, Clade 2 in orange.

downs linked by direct recorded cattle movements (Figure 4.8, mean 1.29 SNP differences, range 0-11 SNPs) and by spatial proximity of 5km (Figure 4.9B, mean 0.49 SNPs, range 0-4 SNPs) were not significantly different from expectations under the null hypothesis of no association (p-values of 0.350 and 0.338 respectively), whereas the association showed borderline significance for spatial proximities of 2km (Figure 4.9A: mean 0.57 SNPs, range 0-11 SNPs, p=0.048). These analyses were also conducted on a subset of the Group 1 data comprising isolates occurring from 2009 onwards (i.e. after sampling intensity was increased to include all infected cattle in a bTB breakdown), to check for an effect of temporal differences in sampling intensity, and these also gave non-significant results (p>0.05 for all comparisons).

Four of the Group 1 VNTR-10 breakdowns from which isolates were sequenced for this study showed no apparent epidemiological links to any other breakdown; neither through direct recorded movements nor spatial proximity up to 5km (Figure 4.10). One of the Group 2 VNTR-10 breakdowns had no apparent epidemiological links to other VNTR-10 isolates, while the other two Group 2 breakdowns did show epidemiological links, but are only comparatively distantly related to the linked breakdowns (>10 SNP differences).

Data from repeat breakdowns of VNTR-10 within the same herd were available for 11 herds and 27 breakdowns. For nine of these breakdowns, the later breakdown was caused by an isolate closely related to the preceding breakdown (less than 3 SNP differences), indicating the possibility of local persistence even after a cattle herd has been declared free of bTB. The recurrence of infection on the two herds for which the second breakdown differed by 3 SNP differences or more could be due to unsafe buying practices, resulting in multiple imports of different lineages of *M. bovis*.

### 4.4.6 Application of phylogeographic tools

Comparison of different phylogeographic models of bacterial dispersal showed statistical support for a heterogeneous model of spatial diffusion allowing different rates of spread

among phylogenetic branches (log MLEs: -1518 for the relaxed model and -1605 for the homogeneous model). This model further provides information for each phylogenetic branch about the estimated distance travelled over the time period represented by its length. Based on the MCC tree for Group 1 (terminal branches only), the estimated mean diffusion rate was comparatively low at 2km/year, but with higher rates up to 30km/year seen rarely (Figure 4.11). The majority of branches underlying these rates involve distances of <5km travelled over less than 5 years (Figure 4.12).

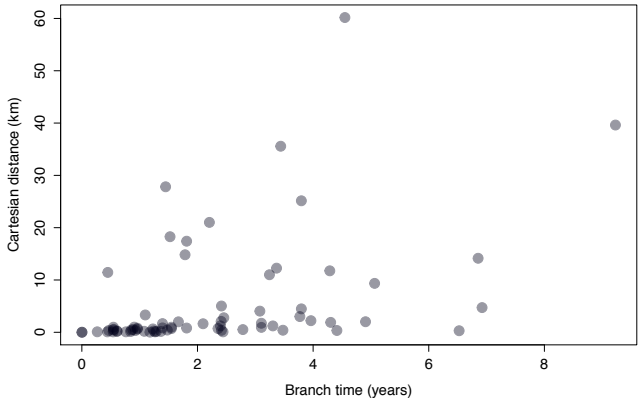


**Figure 4.11: Comparison of branch time and distance travelled for terminal branches for the MCC tree estimated under BEAST continuous phylogeography**

In 95 out of 100 simulations of homogeneous spatial diffusion along the time-stamped VNTR-10 phylogeny, the originally specified diffusion rate could be recovered in BEAST, in that the originally specified value fell within the estimated 95% HPD (Figure 4.13). This suggests that, despite low levels of genetic divergence, our data contain sufficient phylogenetic signal for meaningful phylogeographic inference.

## 4.5 Discussion

Applying WGS to an intensively sampled molecular type of *M. bovis* in cattle has allowed us to explore the potential of this approach to assess the role of cattle movements and spatial proximity in transmission, and to quantify bacterial dispersal across the landscape. These findings demonstrate the potential of WGS as a tool for epidemiological investigation of bTB, but also clearly expose certain limitations.

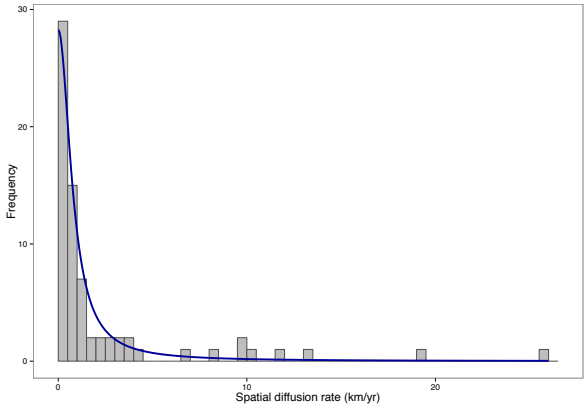


**Figure 4.12: Histogram to show the distribution of branch-specific diffusion rates for terminal branches of the MCC tree** - estimated in BEAST under the Cauchy relaxed random walk model of continuous phylogeography. Superimposed line gives a Cauchy distribution with scale=0.8 and location=0.

### 4.5.1 Differential sampling intensity between VNTR-10 clades and switching of VNTR-type

As expected for a well-sampled population of slowly evolving bacteria, the majority of VNTR-10 isolates (specifically: Group 1, Figure 4.1) were genetically highly similar, often identical and included most of the recent ancestral sequence types that can be inferred from the VNTR-10 phylogeny. In contrast, we found a small number of VNTR-10 isolates (Group 2, Figure 4.1) that were phylogenetically distinct from the Group 1 isolates and showed higher pairwise genetic diversity. Sampling and sequencing effort was even across all VNTR-10-typed isolates from cattle, so the finding of a rare group with more divergent isolates was surprising and suggests that our sampling of Group 2 isolates was less complete compared to Group 1. Observing a much smaller proportion of the overall bacterial diversity might have suggested that this lineage was maintained in a host population largely missed by our sampling, such as a non-cattle reservoir host. However, the sequence data presented here from the closely related VNTR-1 strain indicates that the apparent under-sampling in the Group 2 was caused by switching of VNTR phenotype within this lineage. The phylogenetic placement of VNTR-1 isolates C and D (Figure 4.1) implies that the emergence of VNTR-10 from VNTR-1 occurred independently for the Group 1 clade, whereas the relationship between VNTR-1 isolates A and B, and VNTR-10 isolates in Group 2 (isolates 1-3, Figure 4.1) suggests that in

Group 2 bacteria the VNTR-type has switched VNTR-type multiple times between VNTRs -1 and -10. Because our sampling strategy was reliant on VNTR-typing, and focussed on VNTR-10, lineages that had changed to a different type were almost certain to be missed. Similar evidence for type switching from whole genome data has also recently emerged for human TB (Bryant *et al.*, 2013b; Walker *et al.*, 2012) and in other mycobacteria (Ahlstrom *et al.*, 2015).

Given that we have sequenced so few VNTR-1 isolates, the finding that all of them cluster within the VNTR-10 isolates is significant, however it is difficult to say whether the Group 2 lineage is genuinely more prone to VNTR-switching than Group 1. Although studies have found only limited phenotypic differences between different molecular types of *M. bovis* (Wright *et al.*, 2013a,b), it is possible that a difference between the two groups of VNTR-10 isolates in terms of their propensity for VNTR-switching may be related to functional genetic differences between them. Two out of the four SNPs differentiating the clades show non-synonymous changes in annotated regions, one in the SthA gene and one in a region coding for an unknown hypothetical protein. Additionally, the MV2163B locus which differentiates VNTRs -1 and -10 (Table 4.1) is known to occur within the open reading frame of PPE gene Rv1917c, one of a family of proteins thought to play a role in antigenic variation (Skuce *et al.*, 2002). Therefore a difference in VNTR-switching between clades could also be a response to selective pressures acting on this VNTR locus. However, MV2163B does not show any greater diversity of tandem repeat variation than other NI VNTR-loci, and if anything appears slightly less prone to variation than the other loci used in NI (Skuce *et al.*, 2005).

### 4.5.2 Comparing genetic and epidemiological relationships between breakdowns

An analysis of the association between close spatial proximity among breakdowns (2km and 5km) and genetic similarity gave only marginally significant results (p=0.048) for distances of under 2km, and showed no significance for distances of under 5km (Figure 4.9). A similar evaluation of the correlation between direct recorded movements of

cattle between breakdowns and genetic similarity also showed no significant associations (Figure 4.8).

In contrast to findings for more rapidly evolving bacteria (Eyre *et al.*, 2013), our results suggest that the use of pairwise distances to track transmission and to assess the relative roles of potential transmission mechanisms lacks power for *M. bovis*. The lack of power seen here is likely to be due to the low level of genetic signal in this slowly evolving pathogen as well as the variable, often prolonged, duration of infection within an animal. As a consequence, the genetic distance between breakdowns will be difficult to predict even for breakdowns linked by direct transmission of infection. Adding to this genetic uncertainty, over distances of 2 and 5km the discrepancy between the registered location of a herd and the actual location of the cattle will be brought to the fore (especially in NI where use of rented pasture is common and level of farm fragmentation is high; Abernethy *et al.*, 2006), and recorded movements assessed here do not include indirect movements between herds which may also play a role in transmission.

On a broader scale however, the significant, though weak, correlation between genetic and spatial distances within Group 1 breakdowns demonstrated by the Mantel test and multiple regression on matrices (Figure 4.7) indicates that spatially localised mechanisms are likely involved in the transmission of this lineage. This would be consistent with earlier findings (Biek *et al.*, 2012), as well as the large-scale patterns of spatial expansion of the bTB-endemic areas in the southwest of Britain (Brunton *et al.*, 2015).

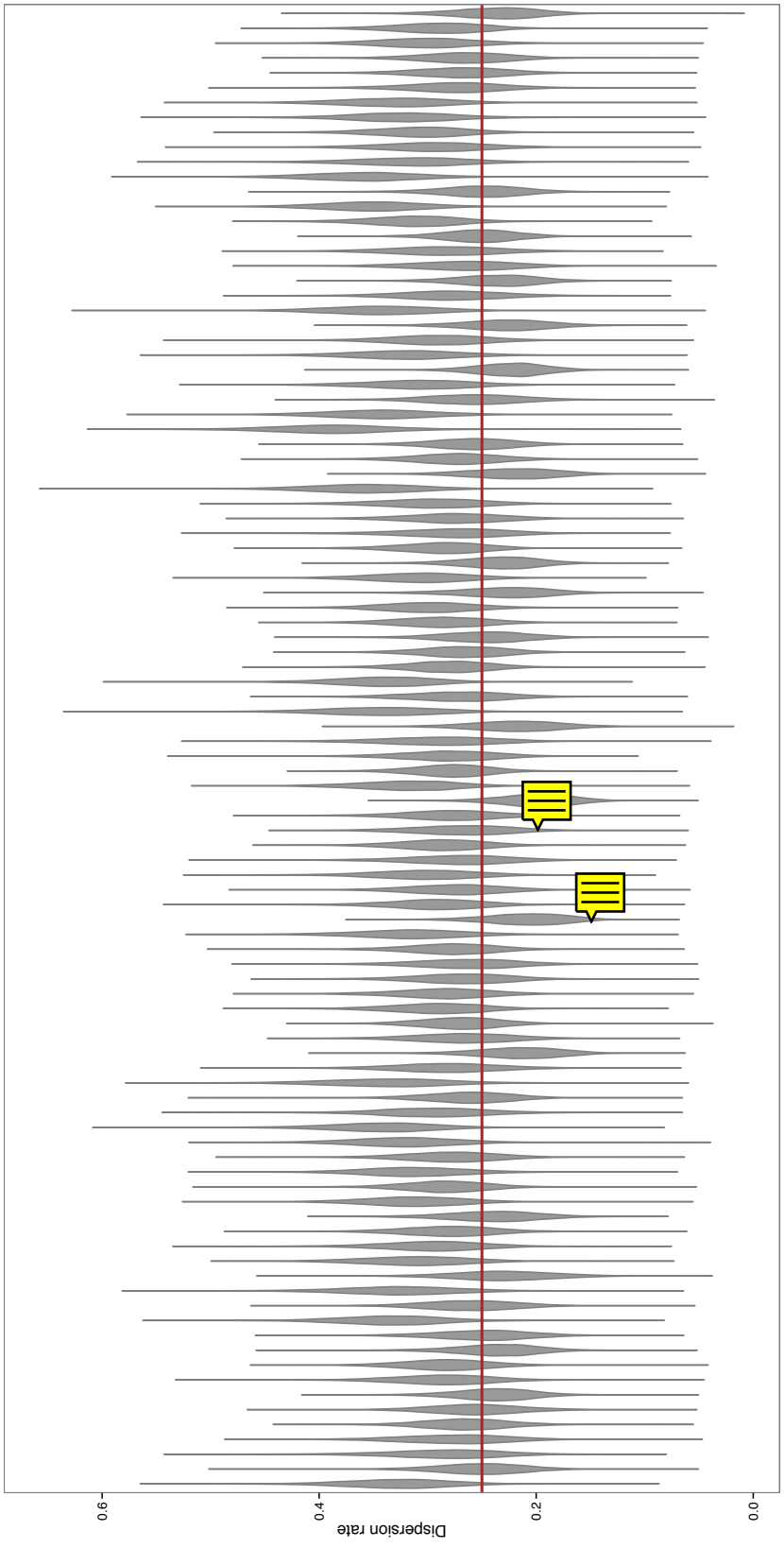


**Figure 4.13: Violin plot of the posterior probabilities of diffusion rates recovered from BEAST phylogeographic analysis on data from 100 simulations of spatial diffusion** - along the time-stamped phylogeny generated from the Clade 1 sequences. Red line indicates the originally-simulated diffusion rate.

### 4.5.3 Application of phylogeographic tools

Despite the limitations described above, phylogenetic data for *M. bovis* can provide insight into the pattern and process of spatial spread. Encouragingly, even taking into account the estimation uncertainties and the low evolutionary rate discussed above, our simulations demonstrate that, genome-wide variation of *M. bovis* contains sufficient information to support meaningful Bayesian phylogeographic analyses over the temporal and spatial scales covered by our data (Figure 4.13).

Based on such an approach, a heterogeneous model of spatial spread fitted the WGS data significantly better than a model assuming a homogeneous diffusion process, implying that the VNTR-10 group has spread across the landscape at a variable rate. Such a pattern might indicate that transmission is underpinned by multiple mechanisms, each associated with a different diffusion process and rate, or it may suggest that transmission is largely driven by a single mechanism involving a variable rate of spatial spread (for example human-mediated movements of cattle between herds). The low mean spatial diffusion rate of VNTR-10 of 2km/year fits the observation of strong spatial clustering characteristic of *M. bovis* in the UK, recorded over different scales and typing methods (Skuce *et al.*, 2010; Smith *et al.*, 2006).

### 4.5.4 Implications for bTB management

Our finding that homoplasy due to VNTR-switching may be common in bTB is significant from an applied point of view since it could confound the epidemiological distinctions made between closely related VNTR-types. However, VNTR-typing in NI is used in conjunction with spoligotyping, and the VNTR loci used have been chosen for optimal discrimination within the NI *M. bovis* population (Skuce *et al.*, 2005), both of which are expected to reduce the impact of VNTR-type homoplasies (Reyes *et al.*, 2012). Of the ten most common VNTR-types in NI (accounting for 85% of all VNTR-typed isolates; Skuce *et al.*, 2010), only one pair is separated by less than two VNTR tandem repeat differences while sharing the same spoligotype. VNTR-

switching is therefore expected to have limited impact on the routine application of VNTR-typing for bTB in NI. Additionally, the VNTR-typing of all culture positive animals in a breakdown (currently standard in NI) will facilitate early detection of VNTR-switching events should they occur.

This study also suggests that it may be possible in principle to use WGS to identify under-sampled populations in *M. bovis*, in this case due to switching of VNTR-type between VNTR-10 and the closely related VNTR-1. However, whether WGS will provide sufficient resolution to characterise the involvement of other types of under-sampled reservoirs for *M. bovis*, such as badger infections, is unclear, and inferences will also be affected by the rate at which transmission occurs between the two host populations (Kao *et al.*, 2014).

Despite the exceptional quality of epidemiological data available for *M. bovis* in NI and the high intensity of bacterial sampling, we found that four Group 1 VNTR-10 breakdowns in this study showed no links to other VNTR-10 breakdowns, neither through direct recorded movements within 10 years of the breakdown nor through spatial proximity of 5km or less. Although in Group 2 isolates such a finding is likely due to the under-sampling of this lineage due to VNTR-switching, the presence of unlinked breakdowns in Group 1 is more surprising, suggesting that the epidemiological links assessed here (direct recorded movements and spatial proximity of under 5km) do not cover all the routes through which infection spreads (for example these do not include unrecorded or indirect movements between herds).

We suggest due to its slow evolutionary rate, some limitations will always be inherent in the application of WGS to *M. bovis* epidemiology and accordingly care must be taken in interpreting results: certain analyses will always remain problematic, for example unambiguous determination of the underlying transmission tree, who infected whom (Didelot *et al.*, 2014). However, we show here that other approaches, such as Bayesian phylogeographic techniques to explore the spatial spread of disease, do appear feasible for this pathogen.

### 4.5.5 Conclusions

Despite a rate of evolution amongst the lowest recorded to date among bacteria, the genomic data presented show a substantial improvement in genetic resolution over previous methods of genetic typing. While WGS data have considerable potential to enhance both our in-depth understanding of bTB epidemiology as well as routine bTB surveillance, the slow and variable evolutionary rate of *M. bovis* does impose a limit to this potential, as has been noted in human tuberculosis (e.g. Didelot *et al.*, 2014). For the future, we suggest that continued advances in mathematical models integrating epidemiological and genetic information will allow a more confident resolution of the factors involved in the spread of bovine tuberculosis, giving a better understanding of the interplay between epidemiological and genetic factors for this important and troubling pathogen.

# CHAPTER 5

# Genetic and genomic characterisation of *M. bovis* strains sampled from a naturally infected badger population over two decades

# Genetic and genomic characterisation of *M. bovis* strains sampled from a naturally infected badger population over two decades

## 5.1 Summary

Understanding the dynamics of multi-host disease systems is notoriously difficult, especially where one of the hosts is a wildlife population. The involvement of badgers in the epidemiology of bTB in Britain and Ireland is a case in point. In this chapter, I use molecular typing and whole genome sequence (WGS) data from an exceptionally well-studied naturally infected badger population to explore the dynamics of *Mycobacterium bovis* infection within individuals and their links to infection in cattle. Several different *M. bovis* spoligotypes were recorded in infected badgers, and although the badger isolates were dominated by a single spoligotype, five occurrences of mixed infections are noted, in which a single badger was infected with more than one spoligotype. The spoligotypes isolated from badger infections correlate with their occurrence in the local cattle population, suggesting transmission of infection between the two species, and indirect evidence indicating transmission of infection from cattle to badgers is also described. WGS of 220 *M. bovis* isolates from this badger population sampled over a >10 year period shows that a single lineage is present in the badgers at high frequency across all sampling years. Isolates from the same badger show reduced diversity compared with isolates originating from different badgers, and despite little superficial sign of a measurably evolving signal in the WGS data, I demonstrate that there is sufficient

temporal-genetic signal to permit the use of Bayesian phylodynamic analysis for these isolates. These data represent a rich, and possibly unique, dataset on *M. bovis* infection in badgers, and continued analysis will generate further information on the evolution and dynamics of *M. bovis* in this wildlife host.

## 5.2 Introduction

The problems of understanding and controlling diseases involving multiple host species (Viana *et al.*, 2014) can become especially challenging where a reservoir of infection exists in a wildlife species, as often only incomplete information available and the options for targeting disease in wildlife are frequently more limited. In these cases, pathogen genetic and genomic information can lead to important improvements in our understanding of the system (Benton *et al.*, 2015).

For bovine tuberculosis (bTB) in Britain and Ireland, the involvement of infection in wild populations of the Eurasian badger *Meles meles* has been a problem for the eradication of the disease in cattle for many decades (Anon, 1975; Krebs *et al.*, 1997). As a consequence of the impact of bTB on the livestock industry, considerable effort and expense have gone into studying the role of the badger in bTB in the region and its impact on the disease in cattle (see for example: Abernethy *et al.*, 2011; Delahay *et al.*, 2001, 2000; Donnelly *et al.*, 2006; Griffin *et al.*, 2005b; Weber *et al.*, 2013). For this reason, in comparison to many wildlife disease systems, a large amount of information has been gathered on badger ecology (see Roper, 2010) and the role of this species in the epidemiology of bTB (Bourne *et al.*, 2007; Corner *et al.*, 2011; Gallagher & Clifton-Hadley, 2000), although despite this there are still many uncertainties surrounding the issue of bTB in badgers, and the subject remains highly controversial.

Bacterial WGS has provided valuable insights into the epidemiology of a number of diseases (e.g. Eyre *et al.*, 2013; Holt *et al.*, 2012; Köser *et al.*, 2012b; Mather *et al.*, 2013), including human tuberculosis (Bryant *et al.*, 2013a,b; Casali *et al.*, 2012; Eldholm *et al.*, 2015; Gardy *et al.*, 2011; Schurch *et al.*, 2010; Walker *et al.*, 2012). The

application of WGS to the bacterial cause of bTB, *M. bovis*, is also beginning to be explored (Biek *et al.*, 2012; Chapter 4 of this thesis). These studies suggest that WGS is likely to prove a useful and exciting addition to the epidemiological toolbox for bTB, enabling us to trace the epidemiology of the disease with far more precision than previous molecular typing methods. However, the slow and variable rate of evolution seen in *M. bovis* is likely to restrict the inferences derived from WGS when compared to other more rapidly evolving pathogens (Biek *et al.*, 2015).

The WGS studies that have been conducted to date for *M. bovis* (Biek *et al.*, 2012 and Chapter 4) have been focussed primarily on cattle infection. Chapter 4 of this thesis demonstrates the use of WGS to identify an undersampled bacterial lineage, raising the possibility that unobserved infection in the badger host might be identifiable through WGS of *M. bovis* samples from cattle infections. However, it is likely that the variation in the the slow and variable rate of evolution in *M. bovis* will also result in too much uncertainty to easily evaluate the involvement of badgers with WGS of cattle isolates alone, and therefore more intense sampling in the badger population will be an important addition to our understanding of the role of this reservoir in the epidemiology of bTB.

Several key questions remain unanswered regarding bTB infection in badgers, such as the diversity of *M. bovis* strains infecting badger populations, the evolution of the bacteria within individual infections, and the rate of transmission between cattle and badgers. Here I present analysis of data from an exceptional study of a naturally infected badger population in the south west of England, with several decades of regular trapping and identification of individual badgers, in addition to isolation and molecular typing of the *M. bovis* isolates infecting them. In this chapter, I evaluate the molecular types and WGS of *M. bovis* present in this badger population, with specific aims as follows:

- To determine the diversity of *M. bovis* strains within the badger population at the spoligotype and whole genome level.

- To quantify the evolution of *M. bovis* in this badger population over time.
- To evaluate the diversity of *M. bovis* detectable within the same badger over time.
- To compare the diversity and abundance of *M. bovis* spoligotypes to that of the surrounding cattle population.

## 5.3 Materials and methods

### 5.3.1 Badger samples

Woodchester Park is a heavily wooded valley of approximately 7km$^2$ in Gloucestershire in the southwest of England, and is surrounded by farmland made up predominantly of a mixture of permanent pasture and ley grassland grazed by beef and dairy cattle and some sheep (Delahay *et al.*, 2006). It is home to a very high-density population of badgers naturally infected with bTB which are free to roam into the surrounding farmland, and for several decades these badgers have been monitored and studied intensively (e.g. Cheeseman, 1979; Cheeseman *et al.*, 1988; Delahay *et al.*, 2000; Rogers *et al.*, 1997; Vicente *et al.*, 2007). The region is comprised of a mix of woodland, grassland and arable land, and the area surrounding the Park is primarily used for cattle farming.

Within Woodchester Park, the spatial extent of badger social territories is identified each spring using bait marking (Delahay *et al.*, 2000), and badgers are trapped at each social group four times per annum. All trapped badgers are anaesthetised, and at their first time of trapping are permanently tattooed. Samples of urine and faeces, tracheal and pharyngeal aspirates, and swabs/aspirates of any bite wounds or abscesses are taken from each badger and subjected to direct culture for *M. bovis*. A post mortem examination is conducted on any badger carcasses found within the Park, and samples from lymph nodes and major organs are cultured for *M. bovis* (Delahay *et al.*, 2013). Positive *M. bovis* cultures isolated from Woodchester Park badgers are generally

subjected to molecular typing (spoligotyping (Kamerbeek *et al.*, 1997) ±VNTR-typing (Frothingham & Meeker-O'Connell, 1998)) at the Animal and Plant Health Agency (APHA), Weybridge at the time of culture. If multiple isolates are available for a single badger they are typed individually. Trapping data and *M. bovis* culture and spoligotyping results spanning the years 1982-2011 were shared for this chapter by Clare Benton and Dez Delahey (APHA Woodchester Park).

## 5.3.2 Whole genome sequencing

### 5.3.2.1 Badger isolates for sequencing

Two hundred and twenty isolates of *M. bovis* from Woodchester Park badgers sampled between 2000-2011 were successfully re-cultured from archived isolates at the Food and Environment Research Agency (FERA), York. DNA was extracted by Adrian Allen and Eleanor Breadon at the Agri Food and Biosciences Institute (AFBI) Northern Ireland, using standard protocols (see Chapter 4), and was sequenced at Glasgow Polyomics on the Illumina MiSeq platform. These isolates originated from 116 individual badgers and of these, bacteria were isolated and sequenced from multiple samples for 47 individuals, either as a result of culture positive samples from multiple anatomical sites at a single trapping date, and/or culture positive samples taken from the same badger on multiple trapping dates. The trapping histories for those badgers with multiple sequenced *M. bovis* isolates are shown in Figure 5.1. Seventy-nine (69.3%) of the 114 badgers included in the sequence dataset were first trapped as cubs and are therefore assumed to have been resident in the Woodchester Park area since birth.

The majority of sequenced isolates (n=213, 96.8%) belonged to Spoligotype 17. Additionally, four Spoligotype 10 isolates, and one isolate each of Spoligotype 9 and Spoligotype 15, were also sequenced.

In the course of going from sampling to sequencing, the Woodchester Park *M. bovis* isolates were processed at several different locations. Unfortunately there are indica-

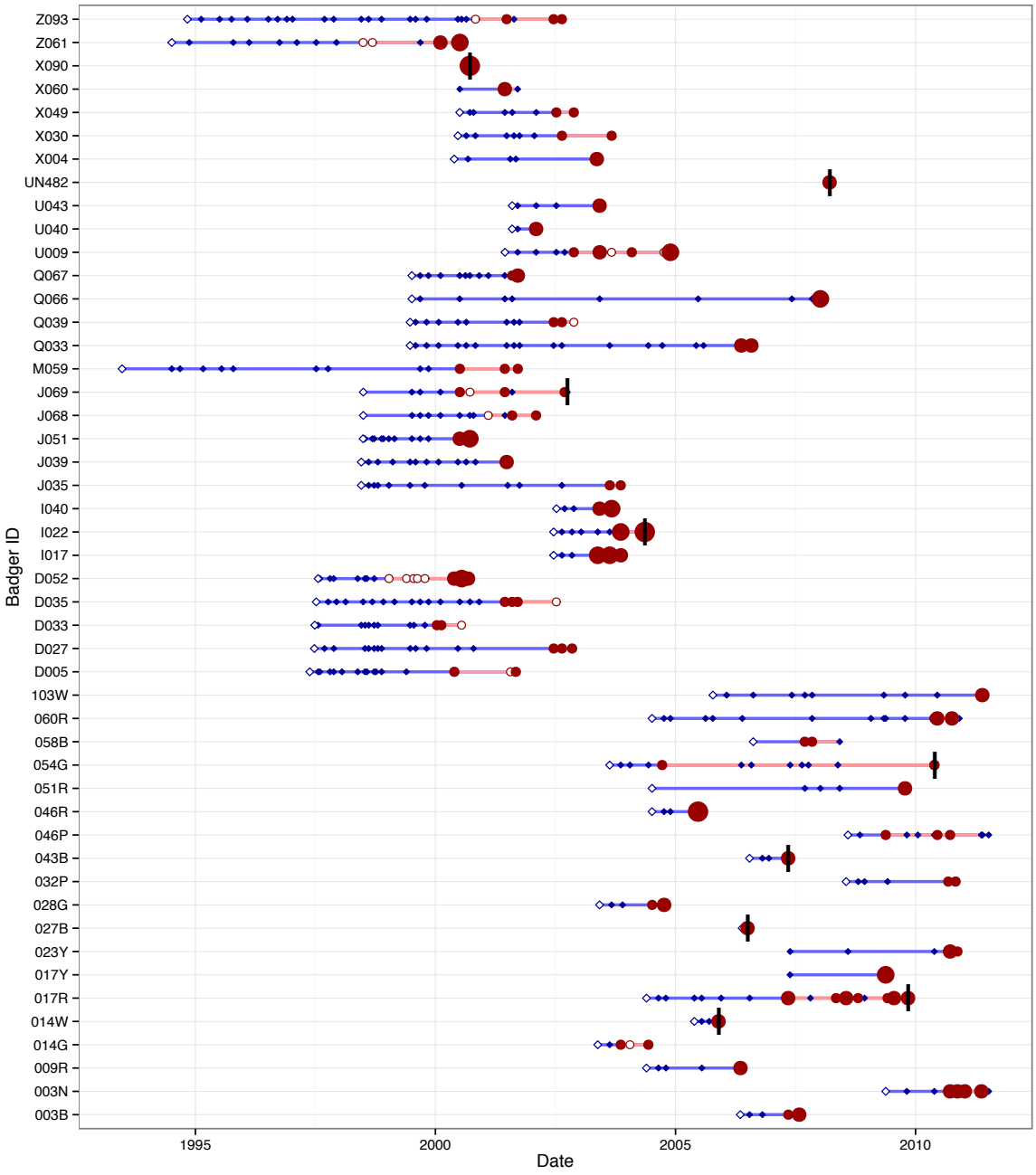


**Figure 5.1: Trapping histories for badgers for which isolates were sequenced from multiple samples** - Blue diamonds show trapping dates for which the badger was culture negative for *M. bovis*. Open blue diamond indicates the badger was first trapped as a cub. Red circles indicate dates at which the badger was culture positive for *M. bovis*, and filled red circles show dates for which bacterial isolates from that badger were sequenced for this study, with the size of circle corresponding to the number of different samples from which isolates were sequenced. Horizontal lines indicate the dates for which we know the badger was alive, and are coloured blue to shows the dates prior to the badger testing culture positive for *M. bovis*, and coloured pink to show the dates after that badger tested positive on culture. Black vertical lines indicate dates of post mortem examinations.

tions that, at some stage in this processing chain, a small number of the isolates (<10) were identified as mislabelled, and it is unclear whether other sequenced isolates were also labelled incorrectly, and if so how many. Therefore, in this chapter I either focus on analyses that do not rely on the metadata attached to each sequenced isolate, or otherwise point out where mislabelling could have impacted the results.

#### 5.3.2.2 Bioinformatics

Sequencing generated paired end reads of approximately 350bp in length, and these were assembled and processed using the bioinformatics pipeline generated for Chapter 4. Briefly, reads were trimmed to remove low quality regions using Trimmomatic (Bolger *et al.*, 2014), mapped to the *M. bovis* reference genome (GenBank accession number BX248333; Garnier *et al.*, 2003) using the BWA mem command (Li & Durbin, 2009). Variant sites were identified using SAMtools (Li et al., 2009) and filtered on mapping quality (>35), heterozygosity (>95% of the reads agreeing with the call at a specific site), number of high quality bases covering that site (at least one in both directions), clustering of variant loci (>50bp between sites), and location relative to repeat regions of the genome as described in the previous chapter. Sites for which more than 10% of the isolates failed the above criteria were then removed, and the resulting sequences were checked for homoplasies as before. These variant sites were concatenated for each isolate, giving genetic sequences in fasta format to be used for downstream analyses.

#### 5.3.2.3 Genetic analysis

Raw pairwise SNP differences between sequences, using pairwise deletions in the event of missing data, were calculated in R v3.1.2 (R Core Team & R Develoment Core Team, 2014) using the package ape (Paradis *et al.*, 2004). Model selection in jModelTest v2.1.7 (Darriba *et al.*, 2012; Guindon & Gascuel, 2003) indicated the HKY nucleotide substitution model (Hasegawa *et al.*, 1985) as the best fit for these data.

The program PhyML v3.0 (Guindon & Gascuel, 2003) was used to generate a maximum likelihood phylogeny for the concatenated sequences, using the HKY nucleotide substitution model and 1000 non-parametric bootstraps to assess node support. A Bayesian phylogeny was also generated using MrBayes v3.2.4 (Ronquist & Huelsenbeck, 2003), again with the HKY model, running two chains for $10^6$ MCMC iterations until a standard deviation of split frequencies of less than 0.01 was reached.

#### 5.3.2.4 Molecular clock rate

The ape package in R was used to identify the correlation between root-to-tip distance and sampling date for each sequenced isolate belonging to Spoligotype 17.

To further test for the presence of a molecular clock signal within the data, the Spoligotype 17 sequences were analysed in BEAST v1.7.4 (Drummond & Rambaut, 2007), and the results of this were compared with BEAST analyses on the same sequences after randomising the tip dates. This latter gives us the null distribution of the BEAST results that would be expected if there were no association between the accumulation of sequence mutations and the date of sampling (Duffy & Holmes, 2009; Firth *et al.*, 2010; Ramsden *et al.*, 2009).

Two independent analyses of the observed data were carried out in BEAST, using the uncorrelated log normal relaxed clock model (Drummond *et al.*, 2006) with a normally distributed prior, truncated at zero, with a mean corresponding to 0.2 SNPs per genome per year (based on the evolutionary rate identified in Chapter 4), and wide standard deviation of approximately 10 SNPs per genome per year. Although the evolutionary rate calculated in Chapter 4 was estimated under a different model of nucleotide substitution (the Jukes Cantor model), the very close genetic relationship between isolates in this study means that the impact of nucleotide substitution model on phylogenetic and temporal estimates should be minimised. However, the wide standard deviation placed on the prior should ensure that even if this is not the case, the results of will not be compromised.

Analyses were run under the HKY model of nucleotide substitution and the Bayesian skyline demographic model (Drummond *et al.*, 2005), for $3x10^8$ MCMC iterations, sampling every 30000 generations. A 10% burnin was removed before the two runs were combined. To generate the null distribution, the tip dates associated with the observed sequences were randomly reassigned twenty times. A single BEAST run was then conducted on each of the twenty sets of randomly dated sequences using the parameters described above.

### 5.3.3 Cattle spoligotype data

Data on bTB testing, bTB breakdowns, and the spoligotypes of *M. bovis* isolates cultured from cattle infections are routinely recorded in Great Britain (GB), and these were made available by the APHA and the Department for Environment, Food and Rural Affairs (Defra) UK. This dataset included information on: bTB breakdowns (breakdown ID, reference of the cattle holding (County-Parish-Herd-Holding (CPHH) number), and start and end dates of the breakdown), bTB testing (CPHH number, test date, and test type), and cattle herds (CPHH number, and x-y coordinates). Records of all spoligotyped isolates sampled from GB cattle were also made available by the APHA and Defra, and included details of spoligotype, CPHH number, breakdown reference, and date of sampling.

Focussing on the decade 2000-2009, the spoligotyping dataset was restricted to include just the first spoligotyping record associated with each bTB breakdown with a start year included in the time period (n=19,237 spoligotyped breakdowns, out of a total of 33,324 bTB breakdowns recorded for this period). 1,287,760 bTB tests were recorded, and 173,585 CPHH numbers were present in the breakdown, test and/or spoligotype datasets for this period (consisting of 152,040 unique x-y coordinates, i.e. some locations were associated with multiple CPHH references).

### 5.3.4 Estimating spatial location of GB spoligotypes in cattle

#### 5.3.4.1 Spatial clustering in GB spoligotypes

This analysis was restricted to the twelve spoligotypes that were associated with over 60 bTB breakdowns in the years 2000-2009, representing over 98% of typed breakdowns (Spoligotypes 9-13, 15, 17, 20-22, 25, and 35). Variograms were constructed to assess whether each of these spoligotypes showed significant spatial clustering. For each of the spoligotypes, pairs of (spoligotyped) breakdowns were identified where one or both members of the pair were infected with the spoligotype of interest. The distance between each of the breakdowns within the pair was identified, and the pairs were grouped into different distance classes (0-25km, 25-50km, 50-75km, 75-100km, 100-125km, 125-150km, 150-175km, 175-200km, 200-300km, and >300km). For each class, the proportion of pairs infected with different spoligotypes was then calculated. If spatial clustering is present, a lower proportion of breakdown pairs will be infected with different spoligotypes for shorter distances classes than would be expected in the absence of spatial clustering.

The null distribution, representing the proportion of pairs of outbreaks expected to be infected with different spoligotypes in the absence of spatial clustering, was also estimated as follows: the matrix of distances between pairs for each spoligotype identified above was randomly permuted, these randomised distances were used to assign each pair to a distance class, and then the proportion of pairs infected with different spoligotypes were calculated for each distance class. This was repeated 1000 times for each spoligotype of interest, and the 95% intervals were recorded and plotted against the observed proportions (above) for each distance class and spoligotype.

#### 5.3.4.2 Estimate of GB spoligotype home ranges

In GB, the method currently optimised to identify the home ranges of different spoligotypes (the geographic location of the core cluster(s) for each spoligotype) is based on

identification of those 5km grid squares in which the spoligotype was present in three out of five years, on at least two unique holdings, and then applying a 10km buffer around those grid squares (AHVLA, 2012; Roberts *et al.*, 2014). Using the information on the dates and the locations of the holdings associated with the spoligotyped *M. bovis* isolates from cattle (this time without restricting to one spoligotype record per breakdown to more accurately represent the procedure used in GB), home ranges were calculated for each of the major spoligotypes (above), for the five-year periods 2000-2004 and 2005-2009. The distance between the Woodchester Park location and the nearest edge of the home range polygons for each major spoligotype was calculated using the R package rgeos. Further, for each spoligotype isolated in the Woodchester badgers, the minimum distance to the nearest cattle record of that spoligotype dated to within five years of the badger isolate(s) was also calculated.

#### 5.3.4.3 Spatial probabilities of spoligotype occurrence

The Generalized Additive Models (GAMs) developed in Chapter 3 were adapted here to estimate the spatial probabilities of occurrence of each of the major GB spoligotypes in cattle. For each unique CPHH location identified above, a count was made of the number of times each of the major clustered spoligotypes was recorded at that holding location. As routine testing for bTB is conducted at different frequencies in different regions of GB, the number of whole herd tests conducted at each location in the years 2000-2009 was also recorded (specific test types and codes included in this are shown in Table 5.1).

For each of the major spoligotypes, the spatial probability of occurrence was then estimated using the R package mgcv (Wood, 2006). The number of times the spoligotype had been recorded at each location was modelled as a binomial response with logit link, with the number of Bernoulli trials equal to the number of whole herd bTB tests (above) recorded at that location. The smoothed x-y coordinates of the holding location were set as the sole explanatory variable. A simplified coastline of GB was included as a boundary using soap film smoothing (Wood *et al.*, 2008), the gamma

**Table 5.1: Test codes and test types included in the GAM model**

| Test code | Test type |
|---|---|
| VE-12M | 12 months post-6M test |
| VE-6M | 6 Month test |
| VE-CON | Contiguous test |
| VE-CON12 | 12 months post CON6-Contiguous test |
| VE-CON6 | 6 months post Contiguous test |
| VE-CT(RTA) | 1st new herd check test |
| VE-CT-HS1 | 2nd new herd check test |
| VE-CT-HS2 | 3rd new herd check test |
| VE-CT-NH1 | 1st reformed herd check test |
| VE-CT-NH2 | 2nd reformed herd check test |
| VE-CT-NH3 | 3rd reformed herd check test |
| VE-CT-RH1 | Carried out outside the normal testing frequency for the herd |
| VE-CT-RH3 | 3rd reformed herd check test |
| VE-CTW1 | For Welsh herds not due a test 1/10/ - 31/12/ |
| VE-CTW2 | Whole herd test for Welsh herds prev. due to reactor herd test 1/10/08-31/12/09 |
| VE-RHT | No testing so restrictions have been imposed |
| VE-SI | Routine surveillance test carried out every 48 months |
| VE-WHT | Whole herd test |
| VE-WHT2 | Yearly test in 2 yearly testing parishes |

parameter was set to 1.4, and knots were generated using a random sample of 500 holding locations, removing those knots within 10km of the boundary lines to allow the soap film algorithm to run.

The predicted probability of occurrence in cattle for each of the major GB spoligotypes was then calculated at the Woodchester Park location using the GAM model outputs. This was compared to the number of Woodchester Park badgers recorded to be infected with that spoligotype (see below) using a generalised linear model with log link, modelling the number of Woodchester badgers infected as Poisson-distributed response variable, with the log probability of occurrence in cattle at Woodchester as explanatory variable.

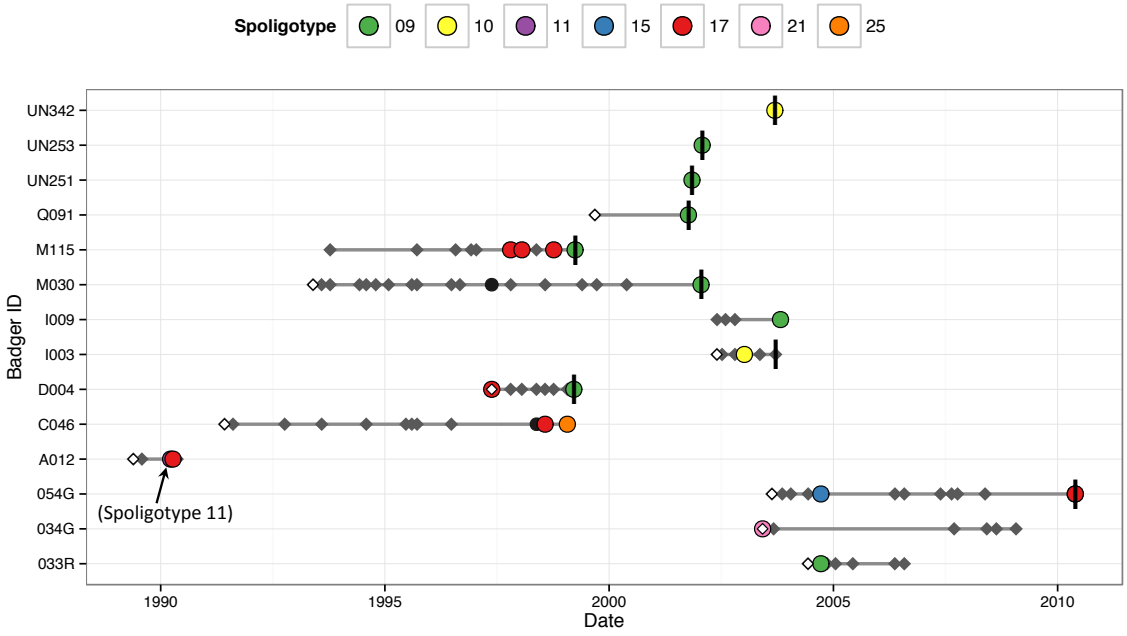


**Figure 5.2: Trapping history for the fourteen badgers infected with spoligotypes other than Spoligotype 17** - Diamonds indicate trapping dates for which the badger was culture negative for M. bovis, and white diamonds indicate the badger was first trapped as a cub. Circles indicate dates at which the badger was culture positive for *M. bovis*, and are coloured by spoligotype (where recorded). Badgers M115, D004, C046, A12, and 054G were additionally infected with Spoligotype 17 (red). Black vertical lines indicate the date of post mortem examinations.

## 5.4 Results

### 5.4.1 Spoligotypes circulating in Woodchester Park

Spoligotyping details had been recorded for a total of 220 badgers, out of the total of 319 badgers recorded as testing positive for *M. bovis* in the trapping dataset. The spoligotyped isolates spanned the years 1990-2011, and a total of seven spoligotypes were recorded. Spoligotype 17 was by far the most common spoligotype, isolated from a total of 211 badgers (95.9%). Spoligotypes 9, 10, 11, 15, 21, and 25 were also recorded in Woodchester badgers at lower frequencies (trapping and culture histories for badgers infected with spoligotypes other than Spoligotype 17 are shown in Figure 5.2, and numbers of badger records for each spoligotype are given in Table 5.2). Of the fourteen badgers infected with spoligotypes other than 17, five badgers (2.3% of the spoligotyped animals) were also recorded as infected with Spoligotype 17 in addition to the other spoligotype. Eight out of the fourteen badgers (57.1%) infected with spoligotypes other than Spoligotype 17 were trapped for the first time as cubs, indicating they have been resident in the park since birth. Of the badgers infected with spoligotype 17 only, 168 out of 209 (80.4%) were trapped first as cubs, although the difference in the proportion trapped as cubs was not significant between the two groups.

**Table 5.2: Details of the most common GB spoligotypes**

| Spoligotype | No. cattle records, 2000-2009 | No. cattle breakdowns 2000-2009 | No. badger records | Distance to home range, 2000-2004 (km) | Distance to home range, 2005-2009 (km) |
|---|---|---|---|---|---|
| 9 | 21271 | 6232 | 8 | 9.5 | 1.5 |
| 10 | 2342 | 613 | 2 | 9.3 | 11.5 |
| 11 | 10286 | 3083 | 1 | 84.2 73.6 | |
| 12 | 1006 | 240 | 0 | 248.8 | 251.9 |
| 13 | 217 | 63 | 0 | - | - |
| 15 | 2021 | 535 | 1 | 198.2 | 191.3 |
| 17 | 15380 | 4422 | 316 | 0 | 0 |
| 20 | 960 | 248 | 0 | 252.7 | 258.9 |
| 21 | 1281 | 344 | 1 | 44.5 | 27.6 |
| 22 | 3081 | 915 | 0 | 19.5 | 19.5 |
| 25 | 5215 | 1646 | 1 | 116.6 | 96.5 |
| 35 | 1848 | 552 | 0 | 38.2 | 40.9 |

### 5.4.2 Sequencing and genetic diversity

Following sequence assembly, the average depth of coverage of the sequenced isolates was 21x, with a mean of 98.1% of the genome per isolate covered by at least one read. After applying the filter criteria described above, a total of 503 high quality single nucleotide polymorphisms (SNPs) were identified across the set of isolates as a whole. Isolates identified as belonging to different spoligotypes were clearly differentiated from each other, with a minimum of 91 SNPs and mean of 126.6 SNPs separating isolates of different spoligotypes. No homoplasies were identified between different spoligotypes.

**Table 5.3: Details of the Spoligotype 17 isolates belonging to each of the different lineages shown in Figure 5.3** - (assumes the metadata associated with the sequenced isolates is correct).

| Lineage | Colour in Figure 5.3 | No. isolates | No. individual badgers | Number badgers first trapped as cubs | Year(s) |
|---|---|---|---|---|---|
| 1 | Purple | 170 | 91 | 80 | 2000-2011 |
| 2 | Yellow | 22 | 9 | 9 | 2003-2010 |
| 3 | Blue | 4 | 1 | 0 | 2000 |
| 4 | Red | 14 | 9 | 7 | 2002-2003 |
| 5 | Green | 2 | 2 | 1 | 2003 |
| 6 | Orange | 1 | 1 | 1 | 2008 |

Spoligotype 17 comprised the majority of sequenced isolates (n=213), and total of 183 SNPs were identified within the Spoligotype 17 group, of which 83 were phylogenetically informative. The mean pairwise distance between Spoligotype 17 isolates was 11.39 SNPs (range: 0-36 SNPs). Closer examination of the Spoligotype 17 phylogeny indicates six different lineages within the Spoligotype 17 group, using a threshold of ten SNP differences between lineages and bootstrap support $\geq 99\%$ (Figure 5.3). The majority of Spoligotype 17 isolates fall into Lineage 1 (Figure 5.3, purple), which was found in all sampling years. Fewer isolates were found in the second lineage (Figure 5.3, yellow), sampled between 2003-2010. The remaining lineages contained between one and fourteen isolates, and none were recorded to persist in the Woodchester Park population for more than twelve months, suggesting that these are transient introduc-

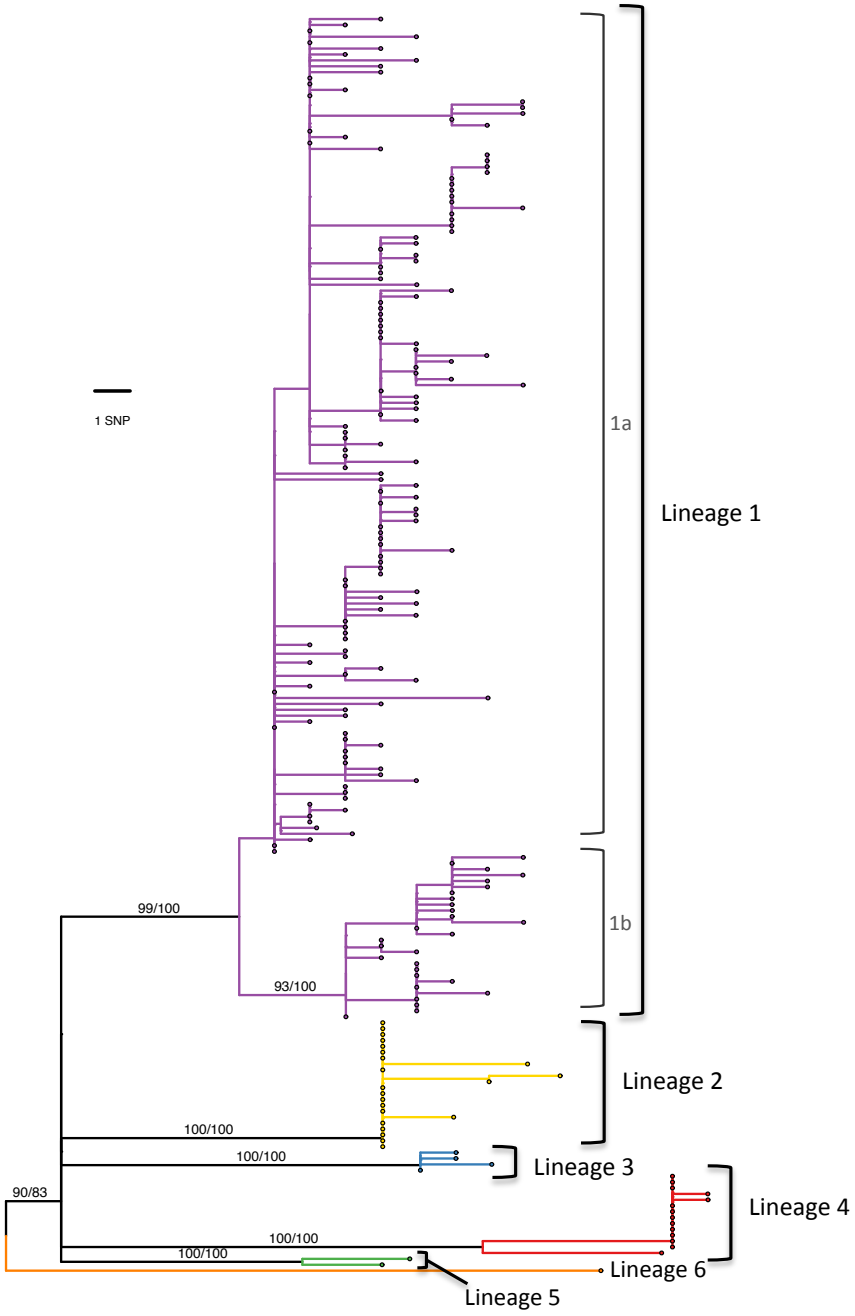


**Figure 5.3: ML phylogeny of the Spoligotype 17 Woodchester Park isolates** - Colours indicate the different lineages and branch labels show the statistical support for selected nodes: the left-hand value indicates percentage bootstrap support from the Maximum Likelihood phylogeny, and the right-hand value shows posterior probability of the node in the Bayesian phylogeny.

tions that have failed to establish in the Woodchester population. Table 5.3 gives the number of isolates and date range of the different lineages. Lineage 3 (blue) appears to represent infection in a single immigrant badger, trapped only once as an adult (assuming the metadata associated with the isolates are correct), whereas other lineages were sampled at least once from badgers trapped as cubs (and therefore assumed to have been resident in the Park since birth).

#### 5.4.2.1 Molecular clock rate

Regressing the distance from the root of the tree against the date of sampling for sequences belonging to the Spoligotype 17 group showed a very low correlation ($R^2$=0.00047, Figure 5.4A) when all Spoligotype 17 sequences were considered, with a somewhat stronger correlation when restricted to the Lineage 1 sequences only ($R^2$=0.048, Figure 5.4B). This is considerably lower than the root-to-tip correlation recorded in Chapter 4 ($R^2$=0.32), and despite the improvement when considering Lineage 1 samples only, this would generally be taken to be poor evidence of a molecular clock like signal within the data (Firth *et al.*, 2010). However, Firth *et al.* (2010) focussed on slowly evolving viruses, and their estimations may not be appropriate for bacteria such as *M. bovis*, and therefore I go on to further test the presence of a temporal signal below.

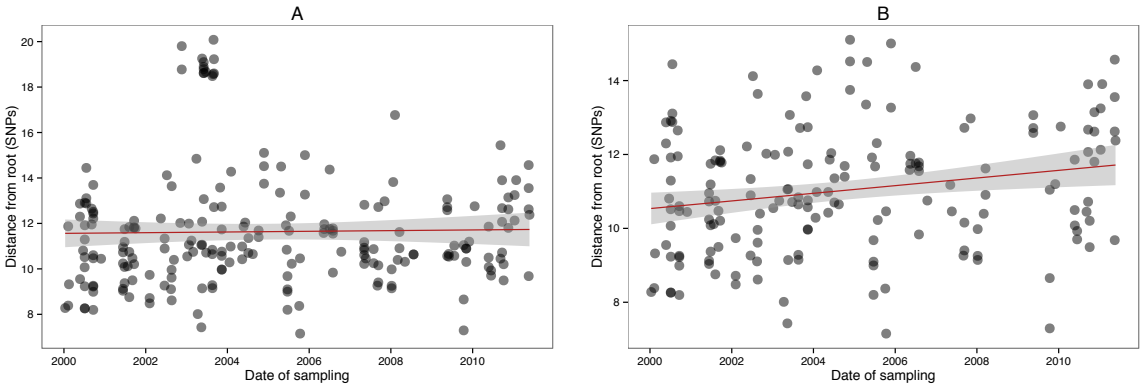


**Figure 5.4: Molecular clock plots** - for all Spoligotype 17 isolates (A) and for isolates belonging to Clade 1 only (B) , showing the accumulation of mutations over time for Spoligotype 17 sequences. Red line shows linear regression line for root-to-tip distance against sampling date, shading indicates 95% confidence intervals.

To further test for the presence of a temporal-genetic signal within the Spoligotype 17 sequences, BEAST analyses were carried out on the observed data and the results compared to those carried out on the sequences after randomly reassigning the dates associated with each of the isolates. All BEAST analyses were determined to have converged based on Estimated Sample Size values of >200 for all parameters. Comparison of the posterior distribution for the molecular clock rate parameter (Figure 5.5) shows that the distribution generated by the observed sequences (dark grey) is significantly different to the distributions generated by the randomised tip dates (light grey), suggesting that, despite the results of the root-to-tip regression above, there actually is an identifiable temporal signal in these data.

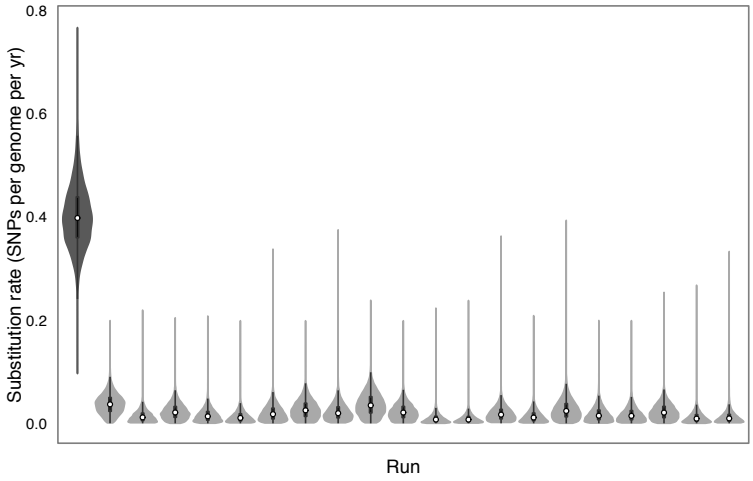


**Figure 5.5: Posterior distributions of clock rate parameter from BEAST runs** - Dark grey shading indicates the combined results from two independent BEAST analyses of the observed data, light grey shading indicates the results of BEAST analyses of the observed sequences with randomly assigned tip dates.

BEAST results estimate a mean evolutionary rate of 0.40 SNPs per genome per year (95% Highest Posterior Density (HPD): 0.29-0.52 SNPs per genome per year) for the Spoligotype 17 isolates, estimating the date of the most recent common ancestor as 1987 for Lineage 1 (95% HPD: 1987-1994), 1998 (95% HPD: 1995-2002) for Lineage 2, 1996 (95% HPD: 1994-1999) for Lineage 3, 1993 (95% HPD: 1986-1999) for Lineage 4, and 1997 (95% HPD: 1993-2002) for Lineage 5 (Spoligotype 17 lineages are shown in Figure 5.3). The date of the most recent common ancestor for the Spoligotype 17 group as a whole was estimated at 1968 (95% HPD: 1953-1983). These analyses do

include multiple isolates per badger, and may therefore be complicated by the rate at which mutations accumulate within a single infection differing from that seen between infections.

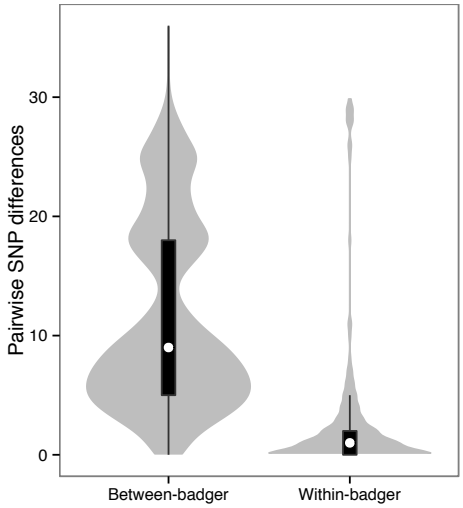


**Figure 5.6: Comparison of pairwise SNP differences between isolates from the same badger and isolates from different badgers for Spoligotype 17 sequences** - Comparisons of isolates from the same badger are labelled "within badger" and those from different badgers are labelled "between badger".

#### 5.4.2.2 Within badger diversity

The pairwise distances between isolates labelled as having been sampled from the same badger were significantly lower than those labelled as coming from different badgers (Figure 5.6), with a mean of 3.67 SNPs (range 0-133) from the same badger compared to a mean of 17.76 SNPs (range 0-145) from different badgers. A generalized linear model confirmed this difference was statistically significant (LRT: $p<0.001$ on one degree of freedom). While this does not rule out the possibility that some of the isolates were mislabelled, the fact that the within-badger pairwise distances are lower than between badger distances indicates that mislabelling has not been so extensive as to obscure this signal.

### 5.4.3 Spatial locations of spoligotypes in cattle

Variograms indicated that all of the twelve most common spoligotypes in GB between 2000-2009 showed significant levels of geographical clustering (Figure 5.7), and GAMs (Wood, 2006) were generated to estimate the spatial probability of occurrence for each of these spoligotypes. The locations of high probability of occurrence for each spoligotype appear to correlate well with the home range areas of the spoligotypes generated using the traditional GB method (Figure 5.8), although better correspondence was found when applying the GB home range methodology after downsampling to one spoligotyping record per herd breakdowns (results not shown). Although significantly clustered (Figure 5.7), no home range was generated for Spoligotype 13, probably due to the lower number of individual records associated with this spoligotype (n=217 records associated with 63 breakdowns, Table 5.2).

### 5.4.4 Comparison between spoligotypes in cattle and in Woodchester Park badgers

Of the spoligotypes found in the Woodchester badgers, Spoligotype 17 was by far the most common, and the Woodchester Park location was also well within the home range area for this spoligotype in cattle (Figure 5.8). Woodchester Park is outside, but reasonably close to, the home range areas, as defined by the APHA methodology for Spoligotypes 9 and 10 (1.5km and 9.3km distances respectively, Table 5.2, Figure 5.8), and these spoligotypes were also recorded sporadically in the Woodchester badgers. The home range for Spoligotype 22 was also less than 20km away from Woodchester and was recorded in high numbers in the cattle population (Table 5.2), however this spoligotype was not recorded at all in the Woodchester badgers.

The home ranges of those spoligotypes recorded only once in the Woodchester badgers (Spoligotypes 11, 15, 21 and 25) were all a minimum of 25km away from the Woodchester location, and in some cases distances were over 200km distant (Table 5.2). Cattle infected with these spoligotypes had been recorded within 20.3km, 10.0km, 11.6km

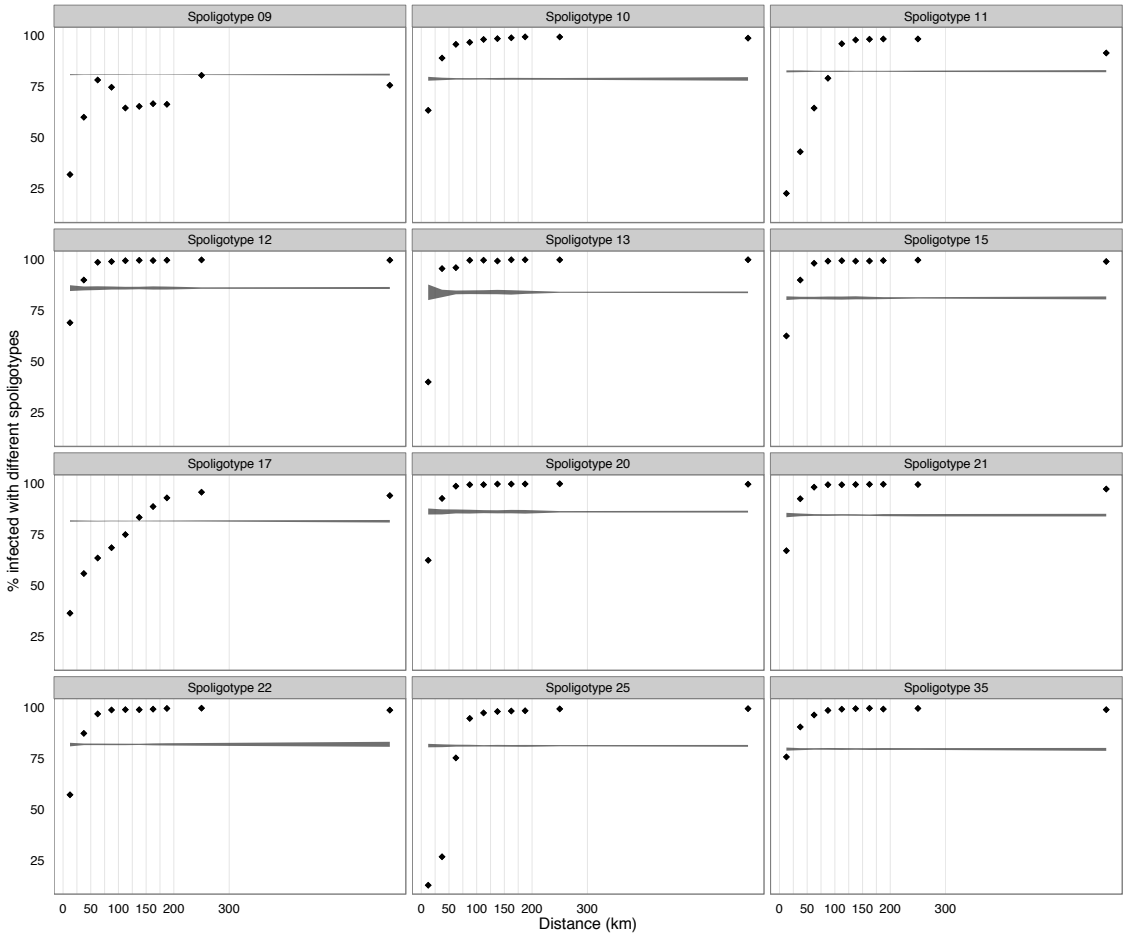


**Figure 5.7: Variograms by spoligotype** - The percentage of pairs of cattle breakdowns for each spoligotype separated by different distances, infected with different spoligotypes. Points show the observed proportion infected with different spoligotypes for each distance class. Grey shading indicates the 95% interval for the values expected under the null hypothesis of no association between distance class and percentage of pairs infected with different spoligotypes.

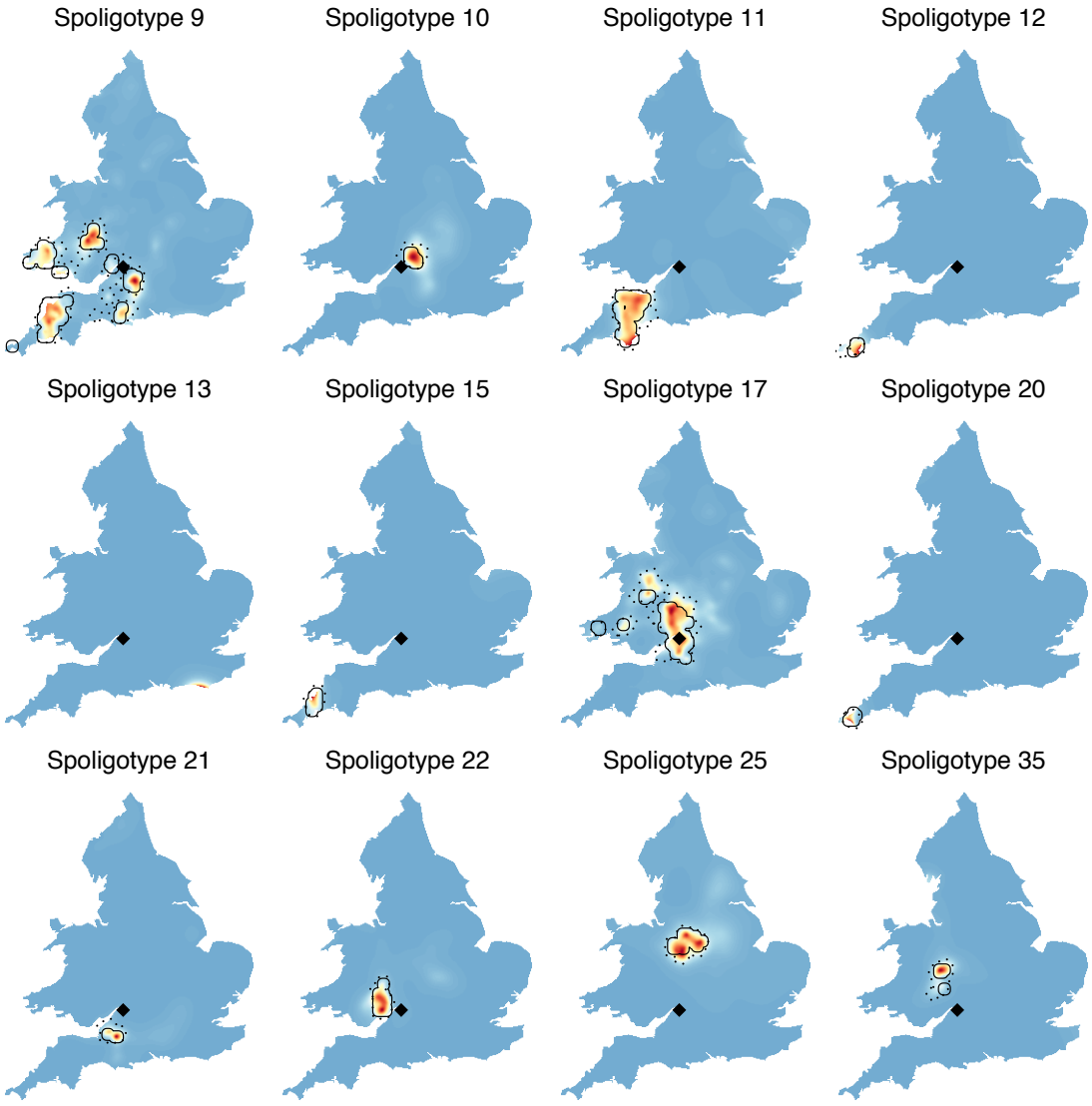


**Figure 5.8: Maps of the spatial locations of the most common GB spoligotypes** - The spatial probability of occurrence for each spoligotype (2000-2009) is shown by coloured shading, with red shading indicating areas of high probability of occurrence, and blue showing low probability of occurrence. The home range area for each spoligotype, estimated using the empirical APHA criteria, is shown by solid lines for the years 2000-2004, and dotted lines for the years 2005-2009. The location of Woodchester Park is plotted as a black diamond.

and 30.0km of Woodchester for Spoligotypes 11, 15, 21, and 25 respectively, within five years of the original badger record for that type.

A GLM also showed a highly significant association between the numbers of Woodchester Park badgers infected with each of the major spoligotypes and the predicted log probability of occurrence of that spoligotype in cattle at the Woodchester location, as estimated from the GAM models (Figure 5.9, Likelihood Ratio Test: $p<0.001$ on 1 degree of freedom).


**Figure 5.9: Comparison between the number of times different spoligotypes were recorded in Woodchester Park badgers and the predicted log probability of occurrence of that spoligotype in cattle at the Woodchester location, as generated by GAM models.** - The results of the GAMs answer the question: if there were a cattle herd present at a given location, what is the spatial probability that this putative herd would have bTB breakdown attributed to that spoligotype? There are no cattle present within Woodchester Park, although cattle are farmed on the surrounding land

## 5.5 Discussion

### 5.5.1 Genetic diversity of *M. bovis* circulating in Woodchester Park badgers

Spoligotype diversity within the Woodchester badgers is limited, with the vast majority of isolates belonging to a single spoligotype (Spoligotype 17), although several other types were present at lower abundance (Table 5.2). This fits the geographically clustered nature of the *M. bovis* population in Britain and Ireland (Goodchild *et al.*,

2012; Olea-Popelka *et al.*, 2005; Skuce *et al.*, 2010; Smith *et al.*, 2006; Woodroffe *et al.*, 2005; Chapter 3 of this thesis), where one *M. bovis* molecular type (here Spoligotype 17) is often seen to dominate in a specific region.

Using WGS to look more deeply into the Spoligotype 17 isolates in the Woodchester badgers reveals that one lineage also dominates at the within-spoligotype resolution, found in 80.5% of the badgers infected with Spoligotype 17 isolates (Figure 5.3, Table 5.3). The low divergence (many isolates are genetically identical) between isolates in this lineage would be consistent with the theory that this particular lineage is often directly transmitted within the Woodchester Park population. Lineage 1 is also the major circulating lineage over the entire time period covered by the sequenced isolates (2000-2011). Lineage 2 is present in multiple years from 2003 onwards, raising the possibility that this lineage had become established in Woodchester Park during the timeframe of the sequenced samples. By contrast, the other lineages were sampled at lower frequency and over less than 12-month timespans, and therefore are likely to represent transient introductions of *M. bovis* into Woodchester Park which have failed to persist in the population.

### 5.5.2 Evolution of *M. bovis* in Woodchester badgers over time

The correlation between root-to-tip distances and sampling dates for Spoligotype 17 isolates in this study is considerably weaker than that recorded in Chapter 4 for *M. bovis* isolates predominantly originating from cattle infections (Figure 5.4). Despite this, further analysis demonstrates that there is actually sufficient temporal signal within these data to allow Bayesian analysis using BEAST (Figure 5.5), justifying the application of powerful phylodynamic approaches (Grenfell *et al.*, 2004) to these data (even given the uncertainty in the metadata associated with the sequenced isolates).

If the increased variability in evolutionary rates recorded here (as compared to that recorded in Chapter 4) is not due to issues of mislabelling, it is possible that it may be the result of infection dynamics in the badgers as compared to cattle. Firstly, the

regular test and slaughter program for bTB in cattle means that few, if any, cattle progress to clinical disease (Liebana *et al.*, 2008), while both asymptomatic and active disease states are routinely recorded in infected badgers (Corner *et al.*, 2011; Gallagher & Clifton-Hadley, 2000). If pre-clinical disease (latency) is associated with a lower mutation rate than active disease for *M. bovis*, as has been suggested for *M. tuberculosis* (Colangeli *et al.*, 2014), then this could act to increase the variability in evolutionary rates in *M. bovis* infections of badgers, as variable portions of the life history of the bacterial lineages will be spent in different disease states, replicating at different rates. Secondly, badger setts are humid, warm environments shielded from UV light (Roper, 1992), and therefore potentially prime sites for environmental survival of *M. bovis* bacteria (Delahay *et al.*, 2000). If non-replicating persistence in the environment plays a role in the epidemiology of bTB in badgers, this will also affect the variability in evolutionary rate in *M. bovis* isolated from badger infections. Notably, opportunities also exist for non-replicating environmental survival on many cattle farms, for example slurry pits and dung piles (Ramirez-Villaescusa *et al.*, 2010).

In addition to increasing heterogeneity in the evolutionary rate, latency and environmental persistence would be both expected to reduce the mean rate of evolution. Contrary to this expectation, I estimated a higher evolutionary rate for this dataset compared to that of the cattle isolates in the previous chapter (0.4 SNPs per genome per year cf. 0.2 SNPs per genome per year in Chapter 4). This argues against a large role for non-replicating environmental persistence of bacteria in the epidemiology of the disease in Woodchester badgers. However, the 95% credible intervals for these two estimates do show slight overlap (0.29-0.52 SNPs per genome per year for this dataset compared to 0.13-0.29 SNPs per genome per year in Chapter 4).

### 5.5.3 Within-badger diversity

Spoligotyping results for Woodchester Park badgers demonstrate several instances of badgers infected with multiple different spoligotypes: in each case the badgers was infected with Spoligotype 17 in addition to one of the less common spoligotypes (Figure

5.2). Spoligotype 17 differs by at least six direct repeats from the other spoligotypes present in the Woodchester Park badgers (Table 5.4), and so these records of multiple infections are unlikely to be due to spoligotyping error or to evolution in spoligotype pattern during a single infection. Given the high levels of Spoligotype 17 circulating in this population it seems probable that other badgers may be infected with multiple strains of Spoligotype 17, which would be undetectable by spoligotyping alone. Infection with multiple separate strains of *M. bovis* has not to our knowledge been described before in badgers, although similar findings have been recorded for bTB in deer (Gortazar *et al.*, 2011), and for human tuberculosis (Cohen *et al.*, 2012).

Figures 5.1 and 5.2 also show that, after testing positive for *M. bovis* at culture, some badgers become culture negative at future trapping events sometimes for prolonged periods of time, after which they may or may not test positive again. It may be that the low sensitivity of bacterial culture for detecting *M. bovis* could explain these events, especially as shedding of *M. bovis* in latent disease is known to be intermittent. An alternative possibility is that some of these badgers may have managed to clear the initial infection, and later positive cultures are due to reinfection from o infected badgers. The finding for badgers I003 and 054G (Figure 5.2) that the bacteria isolated at later timepoints were of a different spoligotype could support this latter hypothesis for these animals, although it is probably more likely that these animals are co-infected with multiple strains of *M. bovis*.

**Table 5.4: Direct repeat patterns for the spoligotypes identified in Woodchester Park badgers** - Spoligotypes 9, 11, 15 and 25 were all isolated from badgers also infected with Spoligotype 17 (Figure 5.2). Direct repeat patterns were downloaded from www.mbovis.org (Smith & Upton, 2012).

| GB ID | International ID | |
|---|---|---|
| 9 | SB0140 | 1 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 |
| 10 | SB0272 | 1 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 0 0 0 0 |
| 11 | SB0274 | 1 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 0 0 0 |
| 15 | SB0275 | 1 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 1 1 0 0 0 0 |
| 17 | SB0263 | 1 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 |
| 21 | SB0130 | 1 1 0 1 1 1 1 1 0 1 0 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 |
| 25 | SB0129 | 1 1 0 1 1 1 1 1 1 0 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 |

Despite the uncertainty in the labels attached to each of the *M. bovis* isolates se-

quenced from Woodchester badger infections, the pairwise genetic distances in isolates originating from within a single badger are significantly lower than those for isolates originating from different animals. This hints that the metadata is correct for many of the isolates sequenced here, although it does not rule out the possibility that the labels are incorrect for some of the sequences. The long tail of sequences sampled from the same badger showing relatively high pairwise SNP differences (Figure 5.6) may be the result of mislabelled isolates (i.e. the isolates were not really from the same badger, they were just labelled as such, thus accounting for the high divergence), or may be due to multiple lineages of *M. bovis* infecting the same badger, as observed in the badger spoligotyping data above. Without an estimate of the expected rate of multiple infection with different lineages in this population, it is not possible to differentiate between these two options.

### 5.5.4 Comparison of spoligotypes found in Woodchester Park badgers with those present in the surrounding cattle population

This study demonstrates a clear correlation between the frequency with which spoligotypes were recorded in badgers in Woodchester Park and the probability of occurrence of those spoligotypes in cattle in the Woodchester area. This is confirmed using the traditional GB method of estimating spoligotype home ranges in cattle: the home range for Spoligotype 17, by far the most common spoligotype isolated from the Woodchester badgers, completely encompasses the Woodchester Park location (Figure 5.8), while for those spoligotypes (9 and 10) which are recorded multiple times but less frequently in the badger isolates, the Woodchester location lies relatively close to the edge of the home range area (Figure 5.8, Table 5.2).

Spoligotypes 11, 15, 21 and 25 were each recorded only once in the Woodchester Park badgers, and Woodchester Park is also located considerable distances from the home range areas for these spoligotypes in cattle. Unless our current understanding of the generally localised nature of badger movements (Byrne *et al.*, 2014; Pope *et al.*, 2006) is incorrect, it is unlikely that these spoligotypes have been translocated from their

home range areas by infected badgers. With the data available, it was not possible to determine whether badgers infected with non-17 spoligotypes were trapped closer to the edge of the Woodchester Park area.

However, cattle are regularly transported over large distances (Mitchell *et al.*, 2005), and therefore it is feasible that cattle infected with the rare spoligotype in the home range area of that spoligotype could be transported to the vicinity of Woodchester Park, and subsequently infect badgers in the area. The low sensitivity of the bTB test and *M. bovis* culture (de la Rua-Domenech *et al.*, 2006; Defra, 2011) means that many bTB infected cattle will not be represented in the spoligotype database. Therefore although the nearest record of cattle infected with Spoligotypes 11, 15, 21, and 25 were 20.3km, 10.0km, 11.6km and 30.0km distant respectively from Woodchester Park, it possible that additional cattle infected with these spoligotypes were also present in the area but not recorded in the spoligotype database. Other evidence from GB also indicates that infection can be transmitted from cattle to badgers (Woodroffe *et al.*, 2006), as do findings in Chapter 3.

An alternative possibility is that these spoligotypes may be circulating in local wildlife without having become established in cattle in the Woodchester area. However, with the exception of Spoligotype 11, a survey of *M. bovis* infecting wildlife in the area around Woodchester Park (Delahay *et al.*, 2007) failed to isolate any of these spoligotypes from the local wildlife (Dez Delahey/APHA, unpublished results).

### 5.5.5 Conclusions

This chapter gives the first in-depth overview of the genetic diversity present in *M. bovis* infecting a well-studied badger population. Additionally, it provides an assessment at the spoligotype level of the relationship between infections in badgers and local cattle, showing clear links between the two populations and raising the suggestion of cattle to badger transmission of bTB. Further WGS of *M. bovis* isolated from local cattle would allow a more detailed picture to be developed of the local interplay of transmission

between these two host species.

Having demonstrated that Bayesian phylodynamic analyses in BEAST are feasible for these data, the application of these approaches to evaluate the spatial, temporal, and population-level spread of *M. bovis* within the Woodchester Park badgers (and between badgers and local cattle if sequences become available) should provide further insights into the dynamics of infection in this population. Additionally, the availability of sequenced isolates from the same badger at multiple timepoints and/or from multiple anatomical sites raises the possibility of understanding the evolutionary events with an infected individual as the infection progresses over time, and as it spreads to different sites within the body. A fine scale evaluation of the level of infection with multiple strains of *M. bovis* will also provide important data, however these last two themes are more heavily dependent on the metadata associated with individual isolates, and therefore greater certainty in these will be necessary to conduct these analyses.

In summary, this chapter gives a broad overview of this particularly rich dataset, describing the *M. bovis* circulating in a particularly well-documented naturally infected badger population. However, these first steps also highlight the exciting prospects that the genomic data sequenced here have to offer for our future understanding of bTB in the badger host.

# CHAPTER 6

# Discussion

# Discussion

Bovine tuberculosis (bTB) is one of the most important diseases currently facing the livestock industry in Britain and Ireland, and despite an intensive control program in cattle the disease has long evaded eradication in this region. The reasons behind this are various, but prime amongst them are the imperfect sensitivity of the current bTB tests, the difficulties associated with vaccination, and the involvement of a wildlife reservoir of infection in the badger population. In this situation we must make use of every tool available to further understand the epidemiological situation and to control the disease.

Genetic epidemiology, or the use of pathogen genetics as an aid to understanding the spread of disease, is one such tool. Molecular typing of *Mycobacterium bovis* isolates has been conducted for several years for routine surveillances of bTB in the UK, while the use of high density whole genome sequencing (WGS) in epidemiology has recently become possible, but has yet to be fully explored for *M. bovis*. In this thesis, I investigate the use of both classes of genetic markers to establish what they can tell us about the underlying processes responsible for the persistence and spread of *M. bovis* in the UK.

## 6.1 Chapter summaries

The first half of the thesis deals with retrospective molecular typing data collected on a national scale, using these exceptionally dense data to investigate the VNTR-types across Northern Ireland (NI) specifically. The patterns present in these data are clear,

and Chapters 2 and 3 go into detail to examine the potential mechanisms underpinning them. The analyses show some interesting results, however they also demonstrate how challenging it can be to extract information on the underlying epidemiological drivers for bTB, even in an region such as NI where large amounts of information are available on both the disease and the cattle population.

In Chapter 2, I investigate the relative abundance distributions (RADs) of the molecular types *M. bovis* in NI. Although these do not contain sufficient information to tease apart the exact roles of the mechanisms that shape them, I use existing ecological theory to demonstrate that the left-skewed RAD shown by *M. bovis* VNTR-types in NI cannot be explained by simple neutral processes. The results from simulation models go on to show that historical increases in bTB prevalence in NI and/or transmission heterogeneity (i.e. superspreading) could be responsible for the observed skew, whereas structuring of the population, for example due to a separate reservoir of infection, would be expected to push the distribution in the opposite direction, levelling out the observed RAD.

Although these results indicate that mechanisms other than population structure must be involved in generating the skewed distributions observed, they by no means imply that population structuring does not play a role in this population, merely that the impact of other processes (such as superspreading and/or an increasing population) on the RAD are strong enough to obscure this; "pulling" the distribution towards the other extreme. In fact, it is highly likely that the *M. bovis* population is structured in various ways in the British Isles: spatial structuring of the *M. bovis* population has been demonstrated throughout the British Isles (Diggle *et al.*, 2005; Goodchild *et al.*, 2012; Jenkins *et al.*, 2007; Olea-Popelka *et al.*, 2005; Woodroffe *et al.*, 2005; Chapter 3); various lines of evidence suggest infection circulates within both badger and cattle populations; and within the badger and cattle populations infection is likely to be further divided by the territorial nature of badger behavior and by the structure of the cattle industry.

Chapter 3 takes a landscape genetics approach to assess the spatial structure in *M.*

*bovis* VNTR-types present in NI. I first characterise the spatial patterns present in the *M. bovis* population, noting the correlation between molecular types in badgers and cattle. Incorporating badger population genetic data I show that, although limited population genetic structure is detectable in badgers, where present it appears to correlate with spatial structure evident in *M. bovis* VNTR-types isolated from cattle. This is an approach that has potential applications to other multi-host systems, and the results suggest that transmission in the badger population could be driving the spatial structure evident in *M. bovis* infections in cattle.

Having examined the use of the low-resolution molecular markers at the countrywide level, in Chapters 4 and 5 I go on to focus in on the use of high-resolution WGS data to study the fine scale spread of *M. bovis* in defined sub-populations of cattle and badgers. This is the first use of in-depth WGS to explore the epidemiology of *M. bovis*.

By conducting WGS on all available isolates of a single closely related VNTR-type, in Chapter 4 I demonstrate the possibilities but also the limitations of WGS as an epidemiological marker in slowly evolving bacteria such as *M. bovis*. The increase in resolution of WGS over VNTR-typing is clear from the results, and I show that sufficient signal exists to conduct Bayesian phylogeographic analyses, and to identify undersampled bacterial populations (here explained by multiple switching of VNTR phenotype in one lineage). However, even across the whole genome there is not sufficient genetic diversity to reliably distinguish between different breakdowns of bTB, and this imperfect resolution results in lack of power in the use of pairwise genetic distances between breakdowns to identify the importance of different transmission mechanisms.

Chapter 5 takes some of the approaches developed in the previous chapters and applies them to an intensively studied badger population. Spoligotyping data again highlight the spatial correlations between the molecular types present in badgers and cattle, and identify several badgers infected with multiple spoligotypes. The dominance of a single spoligotype in the badger population is mirrored at the WGS level by the dominance of a specific lineage within that spoligotype. Very limited temporal signal is seen within the WGS data, however I demonstrate that this is still sufficient to justify the use of

Bayesian phylodynamic approaches. Further analyses of the WGS data described here holds potential for our understanding of the dynamics and evolution of *M. bovis* in this wildlife reservoir host.

## 6.2 Spatial spread and the involvement of badgers

The spatial spread of *M. bovis* can be thought of in terms of spread over long distances as compared to local, short distance transmission. Spread over long distances is generally attributed to movements of infected cattle (Gilbert *et al.*, 2005; Green *et al.*, 2008), but these events are often short-lived and rarely become established in the new location, although the results presented in Chapters 3 and 5 suggest that these translocated infections may persist long enough to generate infections within the badger population. The processes underlying the spread of bTB at the local level are much more difficult to determine, and several possible mechanisms may be involved in short distances transmission, including local movements of cattle, contact between cattle in neighbouring herds, survival of bacteria in the environment, and infection in localised wildlife populations.

The persistence and increasing prevalence of bTB in Britain appears correlated to the spatial expansion of the area of endemic disease in the southwest of England and Wales (Brunton *et al.*, 2015), and therefore it is plausible that that the localised spread of bTB is responsible for driving persistence of the disease and its resistance to eradication efforts in this region. As discussed here and elsewhere (Diggle *et al.*, 2005; Goodchild *et al.*, 2012; Jenkins *et al.*, 2007; Olea-Popelka *et al.*, 2005; Woodroffe *et al.*, 2005), molecular types of *M. bovis* within endemic areas of Britain and Ireland show a geographically clustered pattern, which also supports the hypothesis that local spread is important in driving the epidemiology of the disease. However, it is not clear what ratio of long distance to short distance transmission would still allow the maintenance of the clustered pattern observed. Different lineages of human tuberculosis also shows clustering by region on a global scale (Gagneux & Small, 2007), despite the occurrence

of long distance travel in the human population.

Given this apparent importance of transmission over short distances, the possible indications in Chapter 3 that badgers might contribute to the spatial structuring of the NI *M. bovis* population would support the notion that infection in badgers is important in the maintenance of bTB in NI. However the results of this chapter are far from conclusive and further work is needed to explore the alternative hypothesis that infection in cattle is responsible for the observed spatial structure. If the findings of Chapter 3 are correct, they would correlate with the results of the two large-scale badger culling trials in both Britain and Ireland, which showed that the incidence and clustering of bTB observed in the cattle population were affected by the removal of badgers (Donnelly *et al.*, 2006; Griffin *et al.*, 2005b; Jenkins *et al.*, 2007). However, this does not necessarily indicate that badgers would be capable of maintaining the disease if no infection were present in the cattle population (Bourne *et al.*, 2007). Results described in Chapters 3 and 5 and elsewhere (Woodroffe *et al.*, 2006), indicate that cattle also regularly transmit bTB infection to badgers, therefore it is possible that strictly enforced cattle controls alone will be sufficient to eradicate disease from both badger and cattle populations, as suggested by Brooks-Pollock *et al.* (2014) and Bourne *et al.* (2007).

Although there is still much to learn about the involvement of badgers in the spread of bTB, it is clear from many lines of evidence (including Chapter 3 of this thesis, see also Godfray *et al.*, 2013) that infection in badgers influences the levels of disease in cattle, and therefore combining control strategies to target infection in both species is likely to be more effective than focussing on control measures in cattle alone, although bTB control policies in badgers, especially culling measures, remain highly controversial. If badgers are in fact a maintenance host for bTB in the absence of infection in other hosts, then unless cattle can be isolated from infectious contact with the badger population, controlling infection in this species will indeed be essential to the eradication of the disease.

## 6.3 Whole genome sequencing for bTB epidemiology

Chapters 4 and 5 of this thesis describe the application of WGS to the fine scale epidemiology of bTB. The original hope at the start of this project was that WGS would provide a means to "look inside" the geographical clustering of the molecular types, allowing us to identify and characterise the mechanisms important in local spread and persistence. However, the results presented here have shown that this will be a more complicated and challenging task than was first assumed. WGS generates the maximum amount of resolution possible for a genetic marker: even so, the slow rate of evolution in *M. bovis* is such that many infections are genetically indistinguishable, placing *M. bovis* at the very limits of a "measurably evolving pathogen" at the between-host level (Biek *et al.*, 2015).

### 6.3.1 The use of WGS to identify who infected whom

There has been much interest in the use of pathogen genetics to identify the underlying transmission tree involved in the spread of disease, or who infected whom, based broadly around the idea that the more genetically divergent two infections are, the more likely they are to be separated by multiple transmission events. For this application, the ideal pathogen would generate a "unique [genetic] sequence that is shared by the entire within-host population but, immediately upon transmission, would acquire at least one distinguishing mutation" (Kao *et al.*, 2014).

Unfortunately, the work presented in Chapters 4 and 5 shows that *M. bovis* is very far from this ideal. The slow evolutionary rate for *M. bovis* is such that, even when using the variation present across the whole genome, there is often insufficient diversity in the *M. bovis* population to distinguish between different bTB breakdowns, let alone to differentiate infected individuals. It is therefore clear that one cannot infer the presence of direct transmission from genetic similarity, as even when infections are genetically identical this does not necessarily mean they are linked by transmission. As

with molecular typing though, genetic dissimilarity could still be used to rule out transmission links between isolates. However, I suggest that this is likely to be challenging for bTB for two reasons:

Firstly, as recorded here and elsewhere (Bryant *et al.*, 2013b), the evolutionary rate for *M. bovis* and other mycobacteria is not only slow, but also shows considerable variability. Therefore even if infection events occurred at regular time intervals, and the transmitted bacteria were identical to the sampled bacteria, it would still be difficult to predict the maximum number of SNPs expected between infections linked by direct transmission.

Secondly, the presence of within-host diversity will mean that sequenced isolates are not necessarily representative of the bacteria responsible for transmission of infection. The levels of within-host diversity are affected by (a) the number and diversity of bacteria present in the infecting inoculum; (b) the length of time the bacteria have to diversify within the individual infection; and (c) the presence of mixed infections, or individuals infected multiple times with different strains or substrains of the bacterium. Although based on the biology of the disease, the transmission bottleneck (a) for bTB is expected to be narrow, the chronic nature of the infection means that there may be considerable time for bacteria to diversify within an infected host (b) (although the majority of cattle in Britain and Ireland will be removed relatively soon after infection), and, as shown by this and other work (Gortazar *et al.*, 2011), mixed infections (c) can and do occur for *M. bovis*. Furthermore, the *M. bovis* population within an individual infection is structured into separate lesions in the body, which will also affect the discrepancy between sampled and transmitting isolates. The presence of within-host diversity has been shown to considerably compromise the use of WGS to identify underlying transmission links (Worby *et al.*, 2014).

Given these considerations, it seems unlikely that WGS of *M. bovis* will ever provide enough information to unambiguously identify who infected whom. However, much scope exists in combining WGS with the epidemiological and demographic datasets available on bTB and the cattle population in Britain and Ireland, and although

methodologically challenging, these approaches are likely to represent the way forward.

### 6.3.2 WGS and phylodynamics

Despite the issues outlined above, the results presented in Chapters 4 and 5 show that there is still sufficient temporal-genetic signal within the *M. bovis* WGS data to tap into powerful phylodynamic approaches (Grenfell *et al.*, 2004) such as those implemented in the program BEAST (Drummond *et al.*, 2012), even where at first glance there appears to be very limited evidence of temporal signal in the data (Chapter 5).

Phylodynamic approaches were originally developed for use with rapidly evolving RNA viruses such as influenza virus (e.g. Rambaut *et al.*, 2008; Smith *et al.*, 2009), pathogens with a strong temporal-genetic signal but where the availability of other data is variable. Reflecting this, these analyses primarily make use of the information contained within the genetic and temporal data, and only secondarily focus on the other information associated with isolates (Faria *et al.*, 2013; Lemey *et al.*, 2009, 2010).

For bTB in the UK these strengths and weaknesses are reversed. As discussed above, for various reasons only limited signal exists within the temporal-genetic data for *M. bovis*, while (in cattle at least) an extraordinary amount of information is collected on the disease epidemiology and demographics of the host population. Therefore, although I demonstrate here that BEAST analyses are feasible for *M. bovis* WGS, it is not clear whether the traditional phylodynamics approaches make the best use of the strengths of the bTB dataset - perhaps in such systems the ideal approach would focus on the strength of the epidemiological data in the system. Again, the challenges of data integration and development for these approaches are not trivial, especially for a slowly evolving endemic disease such as bTB, and this is likely to be an exciting area of development for the future.

### 6.3.3 Applied implications and future directions for WGS

As costs decline and sequencing technologies continue to advance, WGS is likely to become the genetic typing methodology of choice for *M. bovis*. Its application to human tuberculosis has indicated that WGS holds all the advantages of molecular typing, and with added advantage of avoiding the homoplasies that arise in, for example, VNTR-typing (Chapter 4, Walker *et al.*, 2012). The possibility of sequencing directly from clinical samples, which has again been explored in human medicine (Brown *et al.*, 2015), would be an added advantage as it would remove the costly and time-consuming culture step. However, it is not clear whether the difference in sample material between human and cattle infections (sputum in humans compared to lesion material or pooled tissue samples in cattle) will affect the success of this technique. Decisions on the most appropriate methods of data processing and storage will also need to be addressed before WGS becomes routine for surveillance of *M. bovis* in Britain and Ireland.

The greater resolution of WGS compared to current molecular typing methods is expected to be of considerable benefit to epidemiological investigations in the field. Given the discussion above, it will be important to determine how to interpret WGS in epidemiological situations, and how best to routinely combine these data with the other information used in breakdown surveillance and investigation. As shown in Chapters 4 and 5, even WGS is not capable of giving full resolution at the level of individual infections, or even for individual herd breakdowns. As sequencing technologies advance it may become possible to sequence across the repeat regions present in the mycobacterial genome, which will add to the genetic diversity captured by WGS (McEvoy *et al.*, 2012). However, as these regions are thought to be involved in antigenicity it is not clear whether they would be appropriate for use as neutral genetic markers.

Over and above its use as an epidemiological marker, WGS generates valuable functional information on the sequenced bacteria themselves. Although the *M. bovis* population in Britain and Ireland has been reported to show limited genetic and phenotypic diversity, routine WGS may still uncover genetic variation that will lead to an improved

understanding of bacterial virulence or host-pathogen interactions, both of which could prove invaluable for the development of better testing and vaccination strategies for control of the disease.

## 6.4 Conclusion

The work outlined in this thesis demonstrates the considerable potential of both population level molecular typing data and more targeted WGS approaches to aid our understanding of *M. bovis* in the UK. The importance of local spatial transmission is noted here, as is the challenge of identifying the roles of the various transmission mechanisms that could be driving it. Although this work shows that WGS approaches will not provide all the answers, they may still hold the key to disentangling the events responsible for the persistence of bTB in the UK. There is much to be gained from their use, and WGS is likely to represent the future for genetic epidemiology in *M. bovis* and in other bacteria. Further methodological integration of WGS with other sources of data is important, and will hold additional promise for their application to epidemiological questions.

# BIBLIOGRAPHY

# Bibliography


Abernethy, D. A., Denny, G. O., Menzies, F. D., McGuckian, P., Honhold, N., & Roberts, A. R. (2006). The Northern Ireland programme for the control and eradication of Mycobacterium bovis. *Veterinary Microbiology*, **112**; pages 231–237. 33, 46, 54, 65, 68, 92, 125

Abernethy, D. A., Upton, P., Higgins, I. M., McGrath, G., Goodchild, A. V., Rolfe, S. J., Broughan, J. M., Downs, S. H., Clifton-Hadley, R., Menzies, F. D., de la Rua-Domenech, R., Blissitt, M. J., Duignan, A., & More, S. J. (2013). Bovine tuberculosis trends in the UK and the Republic of Ireland, 1995-2010. *Veterinary Record*, **172**(12); page 312. ISSN 2042-7670. 14, 46

Abernethy, D. A., Walton, E., Menzies, F., Courcier, E., & Robinson, P. (2011). Mycobacterium bovis surveillance in European badgers (Meles meles) killed by vehicles in Northern Ireland: an epidemiological evaluation. In *International Conference on Animal Health Surveillance*, volume 78, pages 216–218. Association pour l'Étude de l'Épidémiologie des Maladies Animales, Lyon, France. 63, 65, 71, 93, 103, 132

Abraham, M., Allen, C., Bacon, H., Dalzell, F., Eastwood, B., Edwards, R., Elliot, P., Fogle, B., Hale, G., Hale, G., Jones, M., Knight, A., Lewis, J., McGill, I., MacMillan, A., Menache, A., Saunders, R., Southgate, P., & Torgerson, P. (2014). Bovine TB and badger control. *Veterinary Record*, **174**; pages 664–666. 18

Ahlstrom, C. A., Barkema, H. W., Stevenson, K., Zadoks, R. N., Biek, R., Kao, R., Trewby, H., Hendrick, S., Haupstein, D., Kelton, D., Fecteau, G., Labrecque, O., Keefe, G., McKenna, S., & De Buck, J. (2015). Limitations of variable number of tandem repeat typing identified through whole genome sequencing of Mycobacterium avium subsp. paratuberculosis on a national and herd level. *BMC Genomics*, **16**; page 161. 124

AHVLA (2012). SE3257 Milestone 3 - Recommended criteria for identifying M. bovis genotype home ranges. Technical Report 2. 25, 68, 141

AHVLA (2013). SE3257 Milestone11 - Genotype and geography: the geographical localisation of Mycobacterium bovis genotypes in Great Britain. Technical report. 25

Allen, A. R., Minozzi, G., Glass, E. J., Skuce, R. A., McDowell, S. W. J., Woolliams, J. A., & Bishop, S. C. (2010). Bovine tuberculosis: the genetic basis of host susceptibility. *Proceedings of the Royal Society B*, **277**; pages 2737–2745. ISSN 1471-2954. 15

Amanfu, W. (2006). The situation of tuberculosis and tuberculosis control in animals of economic interest. *Tuberculosis*, **86**; pages 330–335. 16, 17

Amos, W., Brooks-Pollock, E., Blackwell, R., Driscoll, E., Nelson-Flower, M., & Conlan, A. J. K. (2013). Genetic predisposition to pass the standard SICCT test for bovine tuberculosis in British cattle. *PLoS One*, **8**(3); page e58,245. ISSN 19326203. 7

Anon (1975). Bovine tuberculosis and badgers. *The Lancet*, **306**(7929); page 316. 12, 14, 132

Anon (2015a). Badger culling to be extended to Dorset. *Veterinary Record*, **177**; page 215. ISSN 0042-4900. 18

Anon (2015b). Fourth year of badger vaccination project starts in Wales. *Veterinary Record*, **176**; page 613. ISSN 0042-4900. 18

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular Biology and Evolution*, **29**(9); pages 2157–2167. ISSN 1537-1719. 112

Barry, C. E., Boshoff, H. I., Dartois, V., Dick, T., Ehrt, S., Flynn, J., Schnappinger, D., Wilkinson, R. J., & Young, D. (2009). The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nature Reviews Microbiology*, **7**(12); pages 845–855. ISSN 1740-1534. 5

Bateson, P., Krebs, J., Bourne, J., Munro, R., Sillero, C., Crispin, S., MacMillan, A., Sainsbury, T., Cheeseman, C., Jones, M., & Bayley, J. (2015). Badger cull is flawed and must now stop. 18

Behr, M. A., Warren, S. A., Salamon, H., Hopewell, P. C., Ponce de Leon, A., Daley, C. L., & Small, P. M. (1999). Transmission of Mycobacterium tuberculosis from patients smear-negative for acid-fast bacilli. *Lancet*, **353**; pages 444–449. ISSN 0140-6736. 5

Bentham, P. F. J. & Broom, D. M. (1989). Interactions between cattle and badgers at pasture with reference to bovine tuberculosis transmission. *British Veterinary Journal*, **145**; pages 226–241. 13

Bentley, S. D., Comas, I., Bryant, J. M., Walker, D., Smith, N. H., Harris, S. R., Thurston, S., Gagneux, S., Wood, J., Antonio, M., Quail, M. A., Gehre, F., Adegbola, R. a., Parkhill, J., & de Jong, B. C. (2012). The genome of mycobacterium Africanum West African 2 reveals a lineage-specific locus and genome erosion common to the M. tuberculosis complex. *PLoS Neglected Tropical Diseases*, **6**(2); page e1552. ISSN 19352727. 25

Benton, C. H., Delahay, R. J., Trewby, H., & Hodgson, D. J. (2015). What has molecular epidemiology ever done for wildlife disease research? Past contributions and future directions. *European Journal of Wildlife Research*, **61**(1); pages 1–16. ISSN 1612-4642. 132

Bermingham, M. L., Bishop, S. C., Woolliams, J. A., Pong-Wong, R., Allen, A. R., McBride, S. H., Ryder, J. J., Wright, D. M., Skuce, R. A., McDowell, S. W. J., & Glass, E. J. (2014). Genome-wide association study identifies novel loci associated with resistance to bovine tuberculosis. *Heredity*, pages 1–9. ISSN 0018-067X. 15

Bersier, L. F. & Sugihara, G. (1997). Species abundance patterns: the problem of testing stochastic models. *Journal of Animal Ecology*, **66**; pages 769–774. 49

Biek, R., O'Hare, A., Wright, D., Mallon, T., McCormick, C., Orton, R. J., McDowell, S., Trewby, H., Skuce, R. A., & Kao, R. R. (2012). Whole genome sequencing reveals local transmission patterns of Mycobacterium bovis in sympatric cattle and badger populations. *PLoS Pathogens*, **8**(11); page e1003,008. ISSN 1553-7374. 26, 27, 28, 100, 101, 103, 105, 107, 113, 118, 125, 133

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., & Didelot, X. (2015). Measurably evolving pathogens in the genomic era. *Trends in Ecology & Evolution*, **30**(6); pages 306–313. ISSN 0169-5347. 99, 133, 166

Biek, R. & Real, L. A. (2010). The landscape genetics of infectious disease emergence and spread. *Molecular Ecology*, **19**; pages 3515–3531. ISSN 1365-294X. 60

Bivand, R. S., Pebesma, E. J., & Gomez-Rubio, V. (2008). *Applied spatial data analysis with R*. Springer, New York, USA. ISBN 978-0-387-78170-9. 78

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15); pages 2114–2120. ISSN 14602059. 137

Bourne, F. J., Cox, D. R., Donnelly, C. A., Gettinby, G., McInerney, J. P., Morrison, W. I., & Woodroffe, R. (2007). Bovine tuberculosis: the scientific evidence. Final report of the independent scientific group (ISG) on cattle TB. Technical report, DEFRA, London. 12, 13, 16, 17, 18, 20, 132, 165

Brooks-Pollock, E., Roberts, G. O., & Keeling, M. J. (2014). A dynamic model of bovine tuberculosis spread and control in Great Britain. *Nature*, **511**; pages 228–231. ISSN 0028-0836. 17, 33, 55, 99, 165

Brosch, R., Gordon, S. V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L. M., Pym, A. S., Samper, S., van Soolingen, D., & Cole, S. T. (2002). A new evolutionary scenario for the Mycobacterium tuberculosis complex. *Proceedings of the National Academy of Sciences*, **99**(6); pages 3684–3689. 3, 21, 22

Brown, A. C., Bryant, J. M., Einer-Jensen, K., Holdstock, J., Houniet, D. T., Chan, J. Z., Depledge, D. P., Nikolayevskyy, V., Broda, A., Stone, M. J., Christiansen, M. T., Williams, R., McAndrew, M. B., Tutill, H., Brown, J., Melzer, M., Rosmarin, C., McHugh, T. D., Shorten, R. J., Drobniewski, F., Speight, G., & Breuer, J. (2015). Rapid Whole Genome Sequencing of Mycobacterium tuberculosis directly from clinical samples. *Journal of Clinical Microbiology*, **53**(7); pages 2230–2237. ISSN 0095-1137. 27, 169

Brunton, L., Nicholson, R., Ashton, A., Alexander, N., Wint, W., Enticott, G., Ward, K., Broughan, J., & Goodchild, A. (2015). A novel approach to mapping and calculating the rate of spread of endemic bovine tuberculosis in England and Wales. *Spatial and Spatio-temporal Epidemiology*, **13**; pages 41–50. ISSN 18775845. 14, 125, 164

Bryant, J. M., Harris, S. R., Parkhill, J., Dawson, R., Diacon, A. H., van Helden, P., Pym, A., Mahayiddin, A. a., Chuchottaworn, C., Sanne, I. M., Louw, C., Boeree, M. J., Hoelscher, M., McHugh, T. D., Bateson, A. L. C., Hunt, R. D., Mwaigwisya, S., Wright, L., Gillespie, S. H., & Bentley, S. D. (2013a). Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. *Lancet Respiratory Medicine*, **1**; pages 786–792. ISSN 2213-2600. 27, 132

Bryant, J. M., Schürch, A. C., van Deutekom, H., Harris, S. R., de Beer, J. L., de Jager, V., Kremer, K., van Hijum, S. A. F. T., Siezen, R. J., Borgdorff, M., Bentley, S. D., Parkhill, J., & van Soolingen, D. (2013b). Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC Infectious Diseases*, **13**; page 110. ISSN 1471-2334. 26, 99, 100, 113, 119, 124, 132, 167

Byrne, A. W., Quinn, J. L., O'Keeffe, J. J., Green, S., Sleeman, D. P., Martin, S. W., & Davenport, J. (2014). Large-scale movements in European badgers: has the tail of the movement kernel been underestimated? *Journal of Animal Ecology*, **83**(4); pages 991–1001. ISSN 1365-2656. 12, 61, 62, 92, 93, 110, 157

Carpenter, P. J., Dawson, D. A., Greig, C., Parham, A., Cheeseman, C. L., & Burke, T. (2003). Isolation of 39 polymorphic microsatellite loci and the development of a fluorescently labelled marker set for the Eurasian badger (Meles meles) (Carnivora: Mustelidae). *Molecular Ecology Notes*, **3**; pages 610–615. ISSN 1471-8278. 71

Carpenter, P. J., Pope, L. C., Greig, C., Dawson, D. A., Rogers, L. M., Erven, K., Wilson, G. J., Delahay, R. J., Cheeseman, C. L., & Burke, T. (2005). Mating system of the Eurasian badger, Meles meles, in a high density population. *Molecular Ecology*, **14**; pages 273–284. ISSN 09621083. 93

Carter, S. P., Chambers, M. A., Rushton, S. P., Shirley, M. D. F., Schuchert, P., Pietravalle, S., Murray, A., Rogers, F., Gettinby, G., Smith, G. C., Delahay, R. J., Hewinson, R. G., & McDonald, R. A. (2012). BCG vaccination reduces risk of tuberculosis infection in vaccinated badgers and unvaccinated badger cubs. *PLoS One*, **7**(12); page e49,833. ISSN 1932-6203. 11

Casali, N., Nikolayevskyy, V., Balabanova, Y., Harris, S. R., Ignatyeva, O., Kontsevaya, I., Corander, J., Bryant, J., Parkhill, J., Nejentsev, S., Horstmann, R. D., Brown, T., & Drobniewski, F. (2014). Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nature Genetics*, **46**(3); pages 279–287. ISSN 1061-4036. 27

Casali, N., Nikolayevskyy, V., Balabanova, Y., Ignatyeva, O., Kontsevaya, I., Harris, S. R., Bentley, S. D., Parkhill, J., Nejentsev, S., Hoffner, S. E., Horstmann, R. D., Brown, T., & Drobniewski, F. (2012). Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Research*, **22**; pages 735–745. ISSN 1549-5469. 132

Cassidy, J. P. (2006). The pathogenesis and pathology of bovine tuberculosis with insights from studies of tuberculosis in humans and laboratory animal models. *Veterinary Microbiology*, **112**; pages 151–161. 4, 5

Chambers, M. A., Carter, S. P., Wilson, G. J., Jones, G., Brown, E., Hewinson, R. G., & Vordermeier, M. (2014). Vaccination against tuberculosis in badgers and cattle: an overview of the challenges, developments and current research priorities in Great Britain. *Veterinary Record*, **175**; pages 90–96. ISSN 0042-4900. 11

Chambers, M. A., Crawshaw, T., Waterhouse, S., Delahay, R., Hewinson, R. G., & Lyashchenko, K. P. (2008). Validation of the BrockTB Stat-Pak assay for detection of tuberculosis in Eurasian badgers (Meles meles) and influence of disease severity on diagnostic accuracy. *Journal of Clinical Microbiology*, **46**(4); pages 1498–1500. ISSN 00951137. 10

Chambers, M. A., Rogers, F., Delahay, R. J., Lesellier, S., Ashford, R., Dalley, D., Gowtage, S., Davé, D., Palmer, S., Brewer, J., Crawshaw, T., Clifton-Hadley, R., Carter, S., Cheeseman, C., Hanks, C., Murray, A., Palphramand, K., Pietravalle, S., Smith, G. C., Tomlinson, A., Walker, N. J., Wilson, G. J., Corner, L. A. L., Rushton, S. P., Shirley, M. D. F., Gettinby, G., McDonald, R. A., & Hewinson, R. G. (2011). Bacillus Calmette-Guérin vaccination reduces the severity and progression of tuberculosis in badgers. *Proceedings of the Royal Society B*, **278**(1713); pages 1913–1920. ISSN 0962-8452. 11

Cheeseman, C. (1979). The behaviour of badgers. *Applied Animal Ethology*, **5**(2); page 193. ISSN 03043762. 134

Cheeseman, C. L., Wilesmith, J. W., Stuart, F. A., & Mallinson, P. J. (1988). Dynamics of tuberculosis in a naturally infected badger population. *Mammal Review*, **18**(1); pages 61–72. 134

Clifton-Hadley, R. S., Sayers, A. R., & Stock, M. P. (1995). Evaluation of an ELISA for Mycobacterium bovis infection in badgers (Meles meles). *Veterinary Record*, **137**; pages 555–558. 10

Clifton-Hadley, R. S., Wilesmith, J. W., & Stuart, F. A. (1993). Mycobacterium bovis in the European badger (Meles meles): epidemiological findings in tuberculous badgers from a naturally infected population. *Epidemiology and Infection*, **111**(1); pages 9–19. ISSN 0950-2688. 6

Cohen, T., van Helden, P. D., Wilson, D., Colijn, C., McLaughlin, M. M., Abubakar, I., & Warren, R. M. (2012). Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. *Clinical Microbiology Reviews*, **25**(4); pages 708–719. ISSN 1098-6618. 24, 68, 156

Colangeli, R., Arcus, V. L., Cursons, R. T., Ruthe, A., Karalus, N., Coley, K., Manning, S. D., Kim, S., Marchiano, E., & Alland, D. (2014). Whole genome sequencing of Mycobacterium tuberculosis reveals slow growth and low mutation rates during latent infections in humans. *PLoS One*, **9**(3); page e91,024. ISSN 1932-6203. 26, 119, 155

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. a., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., & Barrell, B. G. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**; pages 537–544. ISSN 0028-0836. 25

Collins, D. M. (2011). Advances in molecular diagnostics for Mycobacterium bovis. *Veterinary Microbiology*, **151**; pages 2–7. ISSN 1873-2542. 23

Conlan, A. J. K., Brooks Pollock, E., McKinley, T. J., Mitchell, A. P., Jones, G. J., Vordermeier, M., & Wood, J. L. N. (2015). Potential benefits of cattle vaccination as a supplementary control for bovine yuberculosis. *PLoS Computational Biology*, **11**(2); page e1004,038. ISSN 1553-7358. 9

Corner, L. A. L. (2006). The role of wild animal populations in the epidemiology of tuberculosis in domestic animals: How to assess the risk. *Veterinary Microbiology*, **112**; pages 303–312. ISSN 0378-1135. 13

Corner, L. A. L., Murphy, D., & Gormley, E. (2011). Mycobacterium bovis infection in the Eurasian badger (Meles meles): the disease, pathogenesis, epidemiology and control. *Journal of Comparative Pathology*, **144**; pages 1–24. ISSN 0021-9975. 6, 10, 14, 132, 155

Coscolla, M. & Gagneux, S. (2014). Consequences of genomic diversity in Mycobacterium tuberculosis. *Seminars in Immunology*, **26**; pages 431–444. ISSN 10445323. 21, 22

Courtenay, O., Reilly, L. A., Sweeney, F. P., Hibberd, V., Bryan, S., Ul-Hassan, A., Newman, C., Macdonald, D. W., Delahay, R. J., Wilson, G. J., & Wellington, E. M. H. (2006). Is Mycobacterium bovis in the environment important for the persistence of bovine tuberculosis? *Biology Letters*, **2**; pages 460–462. ISSN 1744-9561. 10, 119

Cousins, D. V., Skuce, R. A., Kazwala, R. R., & van Embden, J. D. A. (1998). Towards a standardised approach to DNA fingerprinting of Mycobacterium bovis. *The International Journal of Tuberculosis and Lung Disease*, **2**(6); pages 471–478. 23, 100

DAFM (2015a). Bovine Tuberculosis eradication programme: Department's wildlife policy (badgers). *www.agriculture.gov.ie/ media/migration/ animalhealthwelfare/ diseasecontrols/ tuberculosistbandbrucellosis/ wildlifepolicybadgers/ DepartmentsBadgerPolicyApr13.doc*, (Accessed 2nd Sept 2015). 20

DAFM (2015b). TB and Brucellosis Eradication Schemes. *www.agriculture.gov.ie/ animalhealthwelfarediseasecontrol/ bovinetbbrucellosiseradicationschemes/ diseaseeradicationtbbrucellosis/ diseaseeradicationschemes/*, (Accessed 2nd Sept 2015). 15

Dalley, D., Davé, D., Lesellier, S., Palmer, S., Crawshaw, T., Hewinson, R. G., & Chambers, M. (2008). Development and evaluation of a gamma-interferon assay for tuberculosis in badgers (Meles meles). *Tuberculosis*, **88**; pages 235–243. ISSN 1472-9792. 10

DARD NI (2015a). TB control programme. *www.dardni.gov.uk/index/ animal-health-and-welfare/ animal-diseases/ bovine-tuberculosis/ dards-tb-control-programme.htm*, (Accessed 2nd Sept 2015). 15

DARD NI (2015b). Test and vaccinate or remove (TVR) wildlife intervention research. *www.dardni.gov.uk/ test-and-vaccinate-or-remove.htm*, (Accessed 2nd Sept 2015). 20

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**(8); page 772. ISSN 1548-7091. 137

de la Rua-Domenech, R., Goodchild, A. T., Vordermeier, H. M., Hewinson, R. G., Christiansen, K. H., & Clifton-Hadley, R. S. (2006). Ante mortem diagnosis of tuberculosis in cattle: a review of the tuberculin tests, gamma-interferon assay and other ancillary diagnostic techniques. *Research in Veterinary Science*, **81**(2); pages 190–210. ISSN 0034-5288. 7, 8, 9, 15, 158

Defra (2011). Meta-analysis of diagnostic tests and modelling to identify appropriate testing strategies to reduce M.bovis infection in GB herds. Technical report. 8, 9, 158

Defra (2014). The strategy for achieving Officially Bovine Tuberculosis Free status for England. *www.gov.uk/ government/ uploads/system/ uploads/ attachment_data/ file/ 300447/ pb14088-bovine-tb-strategy-140328.pdf*, (Accessed 17th April 2014). 15, 16, 18

Delahay, R. J., Carter, S. P., Forrester, G. J., Mitchell, a., & Cheeseman, C. L. (2006). Habitat correlates of group size, bodyweight and reproductive performance in a high-density Eurasian badger (Meles meles) population. *Journal of Zoology*, **270**(3); pages 437–447. ISSN 09528369. 134

Delahay, R. J., Cheeseman, C. L., & Clifton-Hadley, R. S. (2001). Wildlife disease reservoirs: the epidemiology of Mycobacterium bovis infection in the European badger (Meles meles) and other British mammals. *Tuberculosis*, **81**; pages 43–49. ISSN 1472-9792. 12, 99, 132

Delahay, R. J., Langton, S., Smith, G. C., & Cheeseman, C. L. (2000). The spatio-temporal distribution of Mycobacterium bovis (bovine tuberculosis) infection in a high-density badger population. *Journal of Animal Ecology*, **69**; pages 428–441. 132, 134, 155

Delahay, R. J., Smith, G. C., Barlow, A. M., Walker, N., Harris, A., Clifton-Hadley, R. S., & Cheeseman, C. L. (2007). 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*, **173**; pages 287–301. ISSN 1090-0233. 13, 158

Delahay, R. J., Walker, N., Smith, G. S., Wilkinson, D., Clifton-Hadley, R. S., Cheeseman, C. L., Tomlinson, A. J., & Chambers, M. A. (2013). Long-term temporal trends and estimated transmission rates for Mycobacterium bovis infection in an undisturbed high-density badger (Meles meles) population. *Epidemiology and Infection*, **141**; pages 1445–56. ISSN 1469-4409. 6, 134

Didelot, X., Gardy, J., & Colijn, C. (2014). Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular Biology and Evolution*, **31**(7); pages 1869–1879. ISSN 1537-1719. 128, 129

Diggle, P., Zheng, P., & Durr, P. (2005). Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society Series C*, **54**(3); pages 645–658. ISSN 0035-9254. 25, 68, 162, 164

Donnelly, C. A. & Nouvellet, P. (2013). The contribution of badgers to confirmed tuberculosis in cattle in high-incidence areas in England. *PLoS Currents Outbreaks*, **1**; pages 1–15. 32, 60, 99

Donnelly, C. A., Woodroffe, R., Cox, D. R., Bourne, F. J., Cheeseman, C. L., Clifton-Hadley, R. S., Wei, G., Gettinby, G., Gilks, P., Jenkins, H., Johnston, W. T., le Fevre, A. M., McInerney, J. P., & Morrison, W. I. (2006). Positive and negative effects of widespread badger culling on tuberculosis in cattle. *Nature*, **439**; pages 843–846. 18, 32, 60, 132, 165

Drewe, J. A., O'Connor, H. M., Weber, N., McDonald, R. A., & Delahay, R. J. (2013). Patterns of direct and indirect contact between cattle and badgers naturally infected with tuberculosis. *Epidemiology and Infection*, **141**(7); pages 1467–1475. ISSN 1469-4409. 13

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, **4**(5); page e88. 112, 138

Drummond, A. J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**; page 214. ISSN 1471-2148. 138

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, **22**(5); pages 1185–1192. ISSN 0737-4038. 112, 139

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**(8); pages 1969–73. ISSN 1537-1719. xviii, 112, 168

Duffy, S. & Holmes, E. C. (2009). Validation of high rates of nucleotide substitution in geminiviruses: Phylogenetic evidence from East African cassava mosaic viruses. *Journal of General Virology*, **90**; pages 1539–1547. ISSN 00221317. 138

Dugdale, H. L., Macdonald, D. W., Pope, L. C., Johnson, P. J., & Burke, T. (2008). Reproductive skew and relatedness in social groups of European badgers, Meles meles. *Molecular Ecology*, **17**; pages 1815–1827. ISSN 1365-294X. 93

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., & Balloux, F. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature Communications*, **6**; page 7119. ISSN 2041-1723. 27, 132

European Food Safety Authority (2013). Scientific opinion on field trials for bovine tuberculosis vaccination. *EFSA Journal*, **11**(12); page 3475. 9, 11

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, **14**; pages 2611–2620. ISSN 09621083. 72

Evans, J. T., Smith, E. G., Banerjee, A., Smith, R. M. M., Dale, J., Innes, J. A., Hunt, D., Tweddell, A., Wood, A., Anderson, C., Hewinson, R. G., Smith, N. H., Hawkey, P. M., & Sonnenberg, P. (2007). Cluster of human tuberculosis caused by Mycobacterium bovis: evidence for person-to-person transmission in the UK. *Lancet*, **369**(9569); pages 1270–1276. ISSN 1474-547X. 16

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**; pages 87–112. 36

Eyre, D. W., Cule, M. L., Wilson, D. J., Griffiths, D., Vaughan, A., O'Connor, L., Ip, C. L. C., Golubchik, T., Batty, E. M., Finney, J. M., Wyllie, D. H., Didelot, X., Piazza, P., Bowden, R., Dingle, K. E., Harding, R. M., Crook, D. W., Wilcox, M. H., Peto, T. E. A., & Walker, A. S. (2013). Diverse sources of C. difficile infection identified on whole-genome sequencing. *The New England Journal of Medicine*, **369**(13); pages 1195–1205. ISSN 1533-4406. 99, 125, 132

Faria, N. R., Suchard, M. A., Rambaut, A., Streicker, D. G., & Lemey, P. (2013). Simultaneously reconstructing viral cross-species transmission history and identifying the underlying constraints. *Philosophical Transactions of the Royal Society B*, **368**; page 20120,196. 168

Firth, C., Kitchen, A., Shapiro, B., Suchard, M. A., Holmes, E. C., & Rambaut, A. (2010). Using time-structured data to estimate evolutionary rates of double-stranded DNA viruses. *Molecular Biology and Evolution*, **27**(9); pages 2038–2051. ISSN 1537-1719. 115, 138, 147

Ford, C. B., Lin, P. L., Chase, M. R., Shah, R. R., Iartchouk, O., Galagan, J., Mohaideen, N., Ioerger, T. R., Sacchettini, J. C., Lipsitch, M., Flynn, J. L., & Fortune, S. M. (2011). Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection. *Nature Genetics*, **43**(5); pages 482–486. ISSN 1546-1718. 5, 26

Fortin, M.-J. & Dale, M. R. (2005). *Spatial analysis: a guide for ecologists*. Cambridge University Press, Cambridge, 4 edition. ISBN 9780521804349. 74

Francis, J. (1947). *Bovine tuberculosis, inculding a contrast with human tuberculosis*. Staples Press Ltd., London. 7

Frantz, A. C., Do Linh San, E., Pope, L. C., & Burke, T. (2010a). Using genetic methods to investigate dispersal in two badger (Meles meles) populations with different ecological characteristics. *Heredity*, **104**; pages 493–501. ISSN 1365-2540. 93

Frantz, A. C., Pope, L. C., Etherington, T. R., Wilson, G. J., & Burke, T. (2010b). Using isolation-by-distance-based approaches to assess the barrier effect of linear landscape elements on badger (Meles meles) dispersal. *Molecular Ecology*, **19**(8); pages 1663–1674. ISSN 1365-294X. 94

Frieden, T. R., Sterling, T. R., Munsiff, S. S., Watt, C. J., & Dye, C. (2003). Tuberculosis. *The Lancet*, **362**; pages 887–899. 4

Frothingham, R. & Meeker-O'Connell, W. A. (1998). Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats. *Microbiology*, **144**; pages 1189–1196. 23, 32, 135

Gagneux, S. & Small, P. M. (2007). Global phylogeography of Mycobacterium tuberculosis and implications for tuberculosis product development. *The Lancet infectious diseases*, **7**; pages 328–337. ISSN 1473-3099. 164

Gallagher, J. & Clifton-Hadley, R. S. (2000). Tuberculosis in badgers; a review of the disease and its significance for other animals. *Research in Veterinary Science*, **69**; pages 203–217. 6, 9, 99, 132, 155

Gardy, J., Loman, N. J., & Rambaut, A. (2015). Real-time digital pathogen surveillance the time is now. *Genome Biology*, **16**; page 155. ISSN 1474-760X. 27

Gardy, J. L., Johnston, J. C., Ho Sui, S. J., Cook, V. J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., Varhol, R., Birol, I., Lem, M., Sharma, M. K., Elwood, K., Jones, S. J. M., Brinkman, F. S. L., Brunham, R. C., & Tang, P. (2011). Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *The New England Journal of Medicine*, **364**(8); pages 730–739. ISSN 1533-4406. 27, 132

Garnier, T., Eiglmeier, K., Camus, J. C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., Simon, S., Harris, B., Atkin, R., Doggett, J., Mayes, R., Keating, L., Wheeler, P. R., Parkhill, J., Barrell, B. G., Cole, S. T., Gordon, S. V., & Hewinson, R. G. (2003). The complete genome sequence of Mycobacterium bovis. *Proceedings of the National Academy of Sciences*, **100**(13); pages 7877–7882. 21, 25, 102, 106, 137

Gilbert, M., Mitchell, A., Bourn, D., Mawdesley, J., Clifton-Hadley, R., & Wint, W. (2005). Cattle movement and bovine tuberculosis in Great Britain. *Nature*, **435**; pages 491–496. 13, 14, 60, 99, 164

Gill, W. P., Harik, N. S., Whiddon, M. R., Reiling, P. L., Mittler, J. E., & Sherman, D. R. (2009). A replication clock for Mycobacterium tuberculosis. *Nature Medicine*, **15**(2); pages 211–214. ISSN 1546-170X. 5

Godfray, H. C. J., Curnow, R. N., Dye, C., Pfeiffer, D., Sutherland, W. L. J., & Woolhouse, M. E. J. (2005). Review of Republic of Ireland 'Four Areas' badger culling trial: Report to Professor Howard Dalton FRS. Technical report. 20

Godfray, H. C. J., Donnelly, C. A., Kao, R. R., Macdonald, D. W., Mcdonald, R. A., Petrokofsky, G., Wood, J. L. N., Woodroffe, R., Young, D. B., McLean, A. R., McDonald, Robbie, A., Petrokofsky, G., Wood, J. L. N., Woodroffe, R., Young, D. B., McLean, A. R., Mcdonald, R. A., Petrokofsky, G., Wood, J. L. N., Woodroffe, R., Young, D. B., McLean, A. R., & Fann, B. (2013). A restatement of the natural science evidence base relevant to the control of bovine tuberculosis in Great Britain. *Proceedings of the Royal Society B*, **280**; page 20131,634. 14, 165

Gomez, J. & McKinney, J. D. (2004). M. tuberculosis persistence, latency, and drug tolerance. *Tuberculosis*, **84**; pages 29–44. ISSN 14729792. 5

Good, M., Clegg, T. A., Duignan, A., & More, S. J. (2011). Impact of the national full herd depopulation policy on the recurrence of bovine tuberculosis in Irish herds, 2003 to 2005. *Veterinary Record*, **169**; page 581. ISSN 0042-4900. 16

Goodchild, A. V., Watkins, G. H., Sayers, A. R., Jones, J. R., & Clifton-Hadley, R. S. (2012). Geographical association between the genotype of bovine tuberculosis in found dead badgers and in cattle herds. *Veterinary record*, **170**; page 259. ISSN 2042-7670. 25, 32, 60, 61, 66, 92, 153, 162, 164

Gopal, R., Goodchild, A., Hewinson, G., de la Rua Domenech, R., & Clifton-Hadley, R. (2006). Introduction of bovine tuberculosis to north-east England by bought-in cattle. *Veterinary Record*, **159**; pages 265–271. 25, 68, 92

Gortazar, C., Torres, M. J., Acevedo, P., Aznar, J., Negro, J. J., de la Fuente, J., & Vicente, J. (2011). Fine-tuning the space, time, and host distribution of mycobacteria in wildlife. *BMC microbiology*, **11**(1); page 27. ISSN 1471-2180. 68, 156, 167

Goslee, S. C. & Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal Of Statistical Software*, **22**(7); pages 1–19. ISSN 15487660. 72, 73, 110

Gray, R. R., Tatem, A. J., Johnson, J. A., Alekseyenko, A. V., Pybus, O. G., Suchard, M. A., & Salemi, M. (2011). Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant Staphylococcus aureus ST239 genome-wide data within a Bayesian framework. *Molecular Biology and Evolution*, **28**(5); pages 1593–1603. ISSN 1537-1719. 99

Green, D. M., Kiss, I. Z., Mitchell, A. P., & Kao, R. R. (2008). Estimates for local and movement-based transmission of bovine tuberculosis in British cattle. *Proceedings of the Royal Society B*, **275**; pages 1001–1005. ISSN 0962-8452. 60, 99, 164

Greenwald, R., Esfandiari, J., Lesellier, S., Houghton, R., Pollock, J., Aagaard, C., Andersen, P., Hewinson, R. G., Chambers, M., & Lyashchenko, K. (2003). Improved serodetection of Mycobacterium bovis infection in badgers (Meles meles) using multiantigen test formats. *Diagnostic Microbiology and Infectious Disease*, **46**; pages 197–203. ISSN 07328893. 10

Grenfell, B. T., Pybus, O. G., Gog, J. R., Wood, J. L. N., Daly, J. M., Mumford, J. a., & Holmes, E. C. (2004). Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, **303**; pages 327–332. ISSN 1095-9203. 26, 154, 168

Griffin, J. F. T. & Mackintosh, C. G. (2000). Tuberculosis in deer: Perceptions, problems and progress. *The Veterinary Journal*, **160**; pages 202–219. 13

Griffin, J. M., More, S. J., Clegg, T. A., Collins, J. D., OBoyle, I., Williams, D. H., Kelly, G. E., Costello, E., Sleeman, D. P., OShea, F., Duggan, M., Murphy, J., & Lavin, D. P. T. (2005a). Tuberculosis in cattle: the results of the four-area project. *Irish Veterinary Journal*, **58**(11); pages 629–636. ISSN 0368-0762. 12, 20

Griffin, J. M., Williams, D. H., Kelly, G. E., Clegg, T. A., O'Boyle, I., Collins, J. D., & More, S. J. (2005b). The impact of badger removal on the control of tuberculosis in cattle herds in Ireland. *Preventive Veterinary Medicine*, **67**(4); pages 237–266. ISSN 0167-5877. 32, 60, 132, 165

Groenen, P. M., Bunschoten, A. E., van Soolingen, D., & van Embden, J. D. (1993). Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Molecular Microbiology*, **10**(5); pages 1057–1065. ISSN 0950-382X. 23

Guillot, G., Mortier, F., & Estoup, A. (2005). GENELAND: A computer package for landscape genetics. *Molecular Ecology Notes*, **5**; pages 712–715. ISSN 14718278. 72

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5); pages 696–704. ISSN 1063-5157. 108, 114, 137, 138

Hall, S. (2010). Official bovine tuberculosis-free status in Scotland. *Veterinary Record*, **166**(8); pages 245–246. ISSN 2042-7670. 14

Hankin, R. K. S. (2007). Introducing untb, an R package for simulating ecological drift under the unified neutral theory of biodiversity. *Journal of Statistical Software*, **22**(12); pages 1–15. 35, 36, 39

Harris, S. R., Feil, E. J., Holden, M. T. G. J., Quail, M. a., Nickerson, E. K., Chantratita, N., Gardete, S., Tavares, A., Day, N., Lindsay, J. A., Edgeworth, J. D., de Lencastre, H., Parkhill, J., Peacock, S. J., Bentley, S. D., & Traveres, A. (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science*, **327**(469); pages 469–474. ISSN 1095-9203. 99

Hasegawa, M., Kishino, H., & Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**; pages 160–174. ISSN 00222844. 137

Hershberg, R., Lipatov, M., Small, P. M., Sheffer, H., Niemann, S., Homolka, S., Roach, J. C., Kremer, K., Petrov, D. A., Feldman, M. W., & Gagneux, S. (2008). High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS Biology*, **6**(12); page e311. ISSN 1545-7885. 21, 107

Holden, M. T. G., Hsu, L.-y., Kurt, K., Weinert, L. A., Mather, A. E., Harris, S. R., Strommenger, B., Layer, F., Witte, W., de Lencastre, H., Skov, R., Westh, H., Zemlickova, H., Coombs, G., Kearns, A. M., Hill, R. L., Edgeworth, J., Gould, I., Gant, V., Cooke, J., Edwards, G. F., McAdam, P. R., Templeton, K. E., McCann, A., Zhou, Z., Castillo-Ramirex, S., Feil, E. J., Hudson, L. O., Enright, M. C., Balloux, F., Aanensen, D. M., Spratt, B. G., Fitzgerald, J. R., Parkhill, J., Achtman, M., Bentley, S. D., & Nubel, U. (2013). A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic. *Genome Research*, **23**; pages 653–664. 99

Holt, K. E., Baker, S., Weill, F.-X., Holmes, E. C., Kitchen, A., Yu, J., Sangal, V., Brown, D. J., Coia, J. E., Kim, D. W., Choi, S. Y., Kim, S. H., da Silveira, W. D., Pickard, D. J., Farrar, J. J., Parkhill, J., Dougan, G., & Thomson, N. R.

(2012). Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics*, **44**(9); pages 1056–1059. ISSN 1546-1718. 132

Houston, R. (2001). A computerised database system for bovine traceability. *Revue Scientifiique et Technique de l'Office Internationale des Epizooties*, **20**; pages 652–661. 35, 109

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, New Jersey. ISBN 0-691-02129-5. xix, 27, 31, 36, 39, 48, 53, 54, 57

Hunter, P. R. & Gaston, M. A. (1988). Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. *Journal of Clinical Microbiology*, **26**(11); pages 2465–2466. 43

Irwin, D. E. (2002). Phylogeographic breaks without geographic barriers to gene flow. *Evolution*, **56**(12); pages 2383–2394. ISSN 0014-3820. 61

Jakobsson, M. & Rosenberg, N. A. (2007). CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**(14); pages 1801–1806. ISSN 13674803. 72

Jenkins, H. E., Woodroffe, R., Donnelly, C. A., Cox, D. R., Johnston, W. T., Bourne, F. J., Cheeseman, C. L., Clifton-Hadley, R. S., Gettinby, G., Gilks, P., Hewinson, R. G., Mcinerney, J. P., & Morrison, W. I. (2007). Effects of culling on spatial associations of Mycobacterium bovis infections in badgers and cattle. *Journal of Applied Ecology*, **44**; pages 897–908. ISSN 00218901. 6, 162, 164, 165

Jones, G. J., Whelan, A., Clifford, D., Coad, M., & Vordermeier, H. M. (2012). Improved skin test for differential diagnosis of bovine tuberculosis by the addition of Rv3020c-derived peptides. *Clinical and Vaccine Immunology*, **19**(4); pages 620–622. ISSN 15566811. 9

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protien molecules. In Munro, H., editor, *Mammalian Protien Metabolism*, pages 21–132. Academic Press, New York, USA, iii edition. 108

Kamerbeek, J., Schouls, L., Kolk, A., van Agterveld, M., van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., & van Embden, J. (1997). Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and epidemiology. *Journal of Clinical Microbiology*, **35**(4); pages 907–914. 23, 32, 103, 135

Kao, R. R., Haydon, D. T., Lycett, S. J., & Murcia, P. R. (2014). Supersize me: How whole-genome sequencing and big data are transforming epidemiology. *Trends in Microbiology*, **22**(5); pages 282–291. ISSN 18784380. 26, 128, 166

Karolemeas, K., de la Rua-Domenech, R., Cooper, R., Goodchild, A. V., Clifton-Hadley, R. S., Conlan, A. J. K., Mitchell, A. P., Hewinson, R. G., Donnelly, C. A., Wood, J. L. N., & McKinley, T. J. (2012). Estimation of the relative sensitivity of the comparative tuberculin skin test in tuberculous cattle herds subjected to depopulation. *PloS One*, **7**(8); page e43,217. ISSN 1932-6203. 17

Kim, Y. J. & Gu, C. (2004). Smoothing spline Gaussian regression: more scalable computation via efficient approximation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **66**(2); pages 337–356. ISSN 13697412. 70

King, D., Roper, T., Young, D., Woolhouse, M., Collins, D., & Wood, P. (2007). Bovine tuberculosis in cattle and badgers. A report by the Chief Scientific Adviser, Sir David King. Technical report, Defra, London. 20

King, H. C., Murphy, A., James, P., Travis, E., Porter, D., Sawyer, J., Cork, J., Delahay, R. J., Gaze, W., Courtenay, O., & Wellington, E. M. (2015). Performance of a non-invasive test for detecting Mycobacterium bovis shedding in European badger (Meles meles) populations. *Journal of Clinical Microbiology*, **53**(7); pages 2316–2323. ISSN 0095-1137. 10

Köser, C. U., Ellington, M. J., Cartwright, E. J. P., Gillespie, S. H., Brown, N. M., Farrington, M., Holden, M. T. G., Dougan, G., Bentley, S. D., Parkhill, J., & Peacock, S. J. (2012a). Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathogens*, **8**(8); page e1002,824. ISSN 1553-7374. 27

Köser, C. U., Holden, M. T. G., Ellington, M. J., Cartwright, E. J. P., Brown, N. M., Ogilvy-Stuart, A. L., Hsu, L. Y., Chewapreecha, C., Croucher, N. J., Harris, S. R., Sanders, M., Enright, M. C., Dougan, G., Bentley, S. D., Parkhill, J., Fraser, L. J., Betley, J. R., Schulz-Trieglaff, O. B., Smith, G. P., & Peacock, S. J. (2012b). Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *The New England journal of medicine*, **366**(24); pages 2267–75. ISSN 1533-4406. 132

Krebs, J. R., Anderson, R., Clutton-Brock, T., Morrison, I., Young, D., & Donnelly, C. (1997). Bovine tuberculosis in cattle and badgers: Report to the Rt Hon Dr Jack Cunningham MP. Technical report, Ministry of Agriculture, Fisheries and Food, London. 12, 14, 18, 132

Kruuk, H. H. & Parish, T. (1987). Changes in the size of groups and ranges of the European badger (Meles meles) in an area of Scotland. *Journal of Animal Ecology*, **56**; pages 351–364. 12

Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, **5**(9); page e1000,520. ISSN 1553-7358. 168

Lemey, P., Rambaut, A., Welch, J. J., & Suchard, M. a. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, **27**(8); pages 1877–85. ISSN 1537-1719. 113, 168

Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**(14); pages 1754–1760. ISSN 1367-4811. 105, 137

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16); pages 2078–2079. ISSN 1367-4811. 106

Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, **188**; pages 117–131. ISSN 13850237. 110

Liebana, E., Johnson, L., Gough, J., Durr, P., Jahans, K., Clifton-Hadley, R., Spencer, Y., Hewinson, R. G., & Downs, S. H. (2008). Pathology of naturally occurring bovine tuberculosis in England and Wales. *The Veterinary Journal*, **176**; pages 354–360. 6, 13, 155

Lin, P. L. & Flynn, J. L. (2010). Understanding latent tuberculosis: A moving target. *Journal of Immunology*, **185**; pages 15–22. ISSN 1550-6606. 3, 5

Little, T. W., Naylor, J. W., & Wilesmith, J. W. (1982). Laboratory study of Mycobacterium bovis infection in badgers and calves. *Veterinary Record*, **111**; pages 550–557. 13

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., & Getz, W. M. (2005). Superspreading and the impact of individual variation on disease emergence. *Nature*, **438**; pages 355–359. ISSN 1476-4687. 47, 55

Luciani, F., Francis, A. R., Tanaka, M. M., Faso, B., & Republic, C. (2008). Interpreting genotype cluster sizes of Mycobacterium tuberculosis isolates typed with IS6110 and spoligotyping. *Infection, Genetics and Evolution*, **8**; pages 182–190. ISSN 1567-1348. 32, 34, 54

Maddock, E. C. G. (1933). Studies on the survival time of the bovine tubercle bacillus in soil, soil and dung, in dung and on grass, with experiments on the preliminary treatment of infected organic matter and the cultivation of the organism. *Journal of Hygiene*, **33**(1); pages 103–117. ISSN 0022-1724. 119

Magurran, A. E. & Magurran, a. E. (2005). Species abundance distributions: pattern or process? *Functional Ecology*, pages 177–181. ISSN 0269-8463. 31

Marcon, E., Scotti, I., Herault, B., Rossi, V., & Lang, G. (2014). Generalization of the partitioning of Shannon diversity. *PLoS One*, **9**(3); page e90,289. 81

Mather, A. E., Reid, S. W. J., Maskell, D. J., Parkhill, J., Fookes, M. C., Harris, S. R., Brown, D. J., Coia, J. E., Mulvey, M. R., Gilmour, M. W., Petrovska, L., de Pinna, E., Kuroda, M., Akiba, M., Izumiya, H., Connor, T. R., Suchard, M. A., Lemey, P., Mellor, D. J., Haydon, D. T., & Thomson, N. R. (2013). Distinguishable epidemics of multidrug-resistant Salmonella Typhimurium DT104 in different hosts. *Science*, **341**(6153); pages 1514–1517. ISSN 1095-9203. 99, 132

Mazars, E., Lesjean, S., Banuls, A. L., Gilbert, M., Vincent, V., Gicquel, B., Tibayrenc, M., Locht, C., & Supply, P. (2001). High-resolution minisatellite-based typing as a portable approach to global analysis of Mycobacterium tuberculosis molecular epidemiology. *Proceedings of the National Academy of Sciences*, **98**(4); pages 1901–1906. ISSN 0027-8424. 23, 103

McDonald, J. L., Smith, G. C., McDonald, R. A., Delahay, R. J., & Hodgson, D. (2014). Mortality trajectory analysis reveals the drivers of sex-specific epidemiology in natural wildlifedisease interactions. *Proceedings of the Royal Society B*, **281**; page 20140,526. 6

McEvoy, C. R. E., Cloete, R., Müller, B., Schürch, A. C., van Helden, P. D., Gagneux, S., Warren, R. M., & van Pittius, N. C. G. (2012). Comparative analysis of Mycobacterium tuberculosis pe and ppe genes reveals high sequence variation and an apparent absence of selective constraints. *PLoS One*, **7**(4); page e30,593. ISSN 1932-6203. 169

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., Hurlbert, A. H., Magurran, A. E., Marquet, P. a., Maurer, B. a., Ostling, A., Soykan, C. U., Ugland, K. I., & White, E. P. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, **10**(10); pages 995–1015. ISSN 1461023X. 31

Medlar, E. M. (1940). Pulmonary tuberculosis in cattle. The location and type of lesions in naturally acquired tuberculosis. *American Review of Tuberculosis*, **41**; pages 283–306. 5

Menzies, F. D. & Neill, S. D. (2000). Cattle-to-cattle transmission of bovine tuberculosis. *Veterinary Journal*, **160**; pages 92–106. ISSN 1090-0233. 12, 60

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, **11**(1); pages 31–46. ISSN 1471-0064. 25

Mitchell, A., Bourn, D., Mawdsley, J., Wint, W., Clifton-Hadley, R., & Gilbert, M. (2005). Characteristics of cattle movements in Britain an analysis of records from the Cattle Tracing System. *Animal Science*, **80**; pages 265–273. ISSN 1357-7298. 62, 158

Monaghan, M. L., Doherty, M. L., Collins, J. D., Kazda, J. F., & Quinn, P. J. (1994). The tuberculin test. *Veterinary Microbiology*, **40**; pages 111–124. ISSN 0378-1135. 8

Mouillot, D., George-Nascimento, M., & Poulin, R. (2003). How parasites divide resources: A test of the niche apportionment hypothesis. *Journal of Animal Ecology*, **72**; pages 757–764. ISSN 00218790. 49

Muirhead, R. H., Gallagher, J., & Burns, K. J. (1974). Tuberculosis in wild badgers in Gloucestershire: epidemiology. *Veterinary Record*, **95**; pages 552–555. 12

Nadler, S. A. (1995). Microevolution and the genetic structure of parasite populations. *The Journal of Parasitology*, **81**(3); pages 395–403. ISSN 0022-3395. 60

Namouchi, A., Didelot, X., Schöck, U., Gicquel, B., & Rocha, E. P. C. (2012). After the bottleneck: Genome-wide diversification of the Mycobacterium tuberculosis complex by mutation, recombination, and natural selection. *Genome Research*, **22**; pages 721–734. ISSN 1549-5469. 21

Nardell, E. A. (2004). Catching droplet nuclei: Toward a better understanding of tuberculosis transmission. *American Journal of Respiratory and Critical Care Medicine*, **169**; pages 553–554. ISSN 1073-449X. 13

Neill, S. D., Bryson, D. G., & Pollock, J. M. (2001). Pathogenesis of tuberculosis in cattle. *Tuberculosis*, **81**(1/2); pages 79–86. 5

Nugent, G. (2011). Maintenance, spillover and spillback transmission of bovine tuberculosis in multi-host wildlife complexes: A New Zealand case study. *Veterinary Microbiology*, **151**; pages 34–42. ISSN 1873-2542. 3, 13

O'Connor, C. C. M., Haydon, D. D. T., Kao, R. R. R., OConnor, C. M., Haydon, D. D. T., & Kao, R. R. R. (2011). An ecological and comparative perspective on the control of bovine tuberculosis in Great Britain and the Republic of Ireland. *Preventive Veterinary Medicine*, **104**(3-4); pages 185–197. ISSN 01675877. 19, 20

Okafor, C. C., Grooms, D. L., Bruning-Fann, C. S., Averill, J. J., & Kaneene, J. B. (2011). Descriptive epidemiology of bovine tuberculosis in michigan (1975-2010): lessons learned. *Veterinary Medicine International*, pages 1–13. ISSN 2042-0048. 3, 13

Olea-Popelka, F. J., Flynn, O., Costello, E., McGrath, G., Collins, J. D., O'Keeffe, J., Kelton, D. F., Berke, O., & Martin, S. W. (2005). Spatial relationship between Mycobacterium bovis strains in cattle and badgers in four areas in Ireland. *Preventive Veterinary Medicine*, **71**; pages 57–70. ISSN 01675877. 25, 32, 60, 61, 92, 154, 162, 164

O'Meara, D. B., Edwards, C. J., Sleeman, D. P., Cross, T. F., Statham, M. J., Mcdowell, J. R., Dillane, E., Coughlan, J. P., O'Leary, D., O'Reilly, C., Bradley, D. G., & Carlsson, J. (2012). Genetic structure of Eurasian badgers Meles meles (Carnivora: Mustelidae) and the colonization history of Ireland. *Biological Journal of the Linnean Society*, **106**; pages 893–909. ISSN 00244066. 71, 86, 93

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, **26**(3); pages 419–420. ISSN 1367-4811. 104

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**(2); pages 289–290. ISSN 1367-4803. 71, 137

Pepperell, C. S., Casto, A. M., Kitchen, A., Granka, J. M., Cornejo, O. E., Holmes, E. C., Birren, B., Galagan, J., & Feldman, M. W. (2013). The role of selection in shaping diversity of natural M. tuberculosis populations. *PLoS Pathogens*, **9**(8); page e1003,543. ISSN 1553-7374. 21, 107

Pollock, J. M., McNair, J., Bassett, H., Cassidy, J. P., Costello, E., Aggerbeck, H., Rosenkrands, I., & Andersen, P. (2003). Specific delayed-type hypersensitivity responses to ESAT-6 identify tuberculosis-infected cattle specific delay. *Journal of Clinical Microbiology*, **41**(5); pages 1856–1860. 5, 7

Pope, L. C., Domingo-Roura, X., Erven, K., & Burke, T. (2006). Isolation by distance and gene flow in the Eurasian badger (Meles meles) at both a local and broad scale. *Molecular Ecology*, **15**(2); pages 371–386. ISSN 0962-1083. 12, 62, 92, 93, 110, 157

Prevosti, A., Ocana, J., & Alonso, G. (1975). Distances between populations of Drosophila subobscura, based on chromosome arrangements and frequencies. *Theoretical and Applied Genetics*, **45**; pages 231–241. 71

Pritchard, D. G. (1988). A century of bovine tuberculosis 1888-1988: Conquest and controversy. *Journal of Comparative Pathology*, **99**; pages 357–399. 2

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**; pages 945–959. ISSN 0016-6731. 72

R Core Team & R Develoment Core Team (2014). R: A Language and Environment for Statistical Computing. 34, 63, 110, 137

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., & Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature*, **453**; pages 615–619. ISSN 0028-0836. 168

Ramirez-Villaescusa, A. M., Medley, G. F., Mason, S., & Green, L. E. (2010). Risk factors for herd breakdown with bovine tuberculosis in 148 cattle herds in the south west of England. *Preventive Veterinary Medicine*, **95**; pages 224–230. ISSN 01675877. 155

Ramsden, C., Holmes, E. C., & Charleston, M. A. (2009). Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. *Molecular Biology and Evolution*, **26**(1); pages 143–153. ISSN 07374038. 138

Real, L. A. & Biek, R. (2007). Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. *Journal of the Royal Society Interface*, **4**; pages 935–948. ISSN 1742-5689. 61

Reid, N., Etherington, T. R., Wilson, G. J., Montgomery, W. I., & McDonald, R. A. (2011). Monitoring and population estimation of the European badger Meles meles in Northern Ireland. *Wildlife Biology*, **18**; pages 46–57. ISSN 0909-6396. 12, 73, 74, 94

Reilly, L. M. O. & Daborn, C. J. (1995). The epidemiology of Mycobacterium bovis infections in animals and man: a review. *Tubercle and Lung Disease*, **76**; pages 1–46. 3

Reviriego Gordejo, F. J. & Vermeersch, J. P. (2006). Towards eradication of bovine tuberculosis in the European Union. *Veterinary Microbiology*, **112**; pages 101–109. ISSN 0378-1135. 3, 15, 16, 17

Reyes, J. F., Chan, C. H. S., & Tanaka, M. M. (2012). Impact of homoplasy on variable numbers of tandem repeats and spoligotypes in Mycobacterium tuberculosis. *Infection, Genetics and Evolution*, **12**; pages 811–818. ISSN 15671348. 24, 33, 42, 49, 127

Reynolds, D. (2006). A review of tuberculosis science and policy in Great Britain. *Veterinary Microbiology*, **112**; pages 119–126. 2

Roberts, T., O'Connor, C., Nuñez-Garcia, J., de la Rua-Domenech, R., & Smith, N. H. (2014). Unusual cluster of Mycobacterium bovis infection in cats. *Veterinary Record*, **174**; page 326. ISSN 2042-7670. 27, 68, 141

Roche, P. W., Triccas, J. A., & Winter, N. (1995). BCG vaccination against tuberculosis: past disappointments and future hopes. *Trends in Microbiology*, **3**(10); pages 397–401. ISSN 0966842X. 10

Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., Supply, P., Kalinowski, J., & Niemann, S. (2013). Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. *PLoS Medicine*, **10**(2); page e1001,387. ISSN 1549-1676. 26, 27, 100, 119

Rogers, L. M., Cheeseman, C. L., Mallinson, P. J., & Clifton-Hadley, R. (1997). The demography of a high density badger (Meles meles) population in the west of England. *Journal of Zoology*, **242**; pages 705–728. ISSN 09528369. 134

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**(12); pages 1572–1574. ISSN 1367-4803. 109, 138

Roper, T. J. (1992). Badger Meles meles setts - architecture, internal environment and function. *Mammal Review*, **22**(1); pages 43–53. ISSN 1616-5047. 155

Roper, T. J. (2010). *Badger*. HarperCollins Publisher, London, 1st edition. ISBN 9780007320417. 12, 13, 62, 132

Rosindell, J., Hubbell, S. P., & Etienne, R. S. (2011). The unified neutral theory of biodiversity and biogeography at age ten. *Trends in Ecology and Evolution*, **26**(7); pages 340–348. ISSN 0169-5347. 31, 36, 48

Rosindell, J., Hubbell, S. P., He, F., Harmon, L. J., & Etienne, R. S. (2012). The case for ecological neutral theory. *Trends in Ecology and Evolution*, **27**(4); pages 203–208. ISSN 0169-5347. 31

Russell, D. G. (2011). Mycobacterium tuberculosis and the intimate discourse of a chronic infection. *Immunological Reviews*, **240**; pages 252–268. ISSN 1600-065X. 5

Russell, D. G., Cardona, P.-J., Mi-Jeong, K., Allain, S., & Altare, F. (2009). Foamy macrophages and the progression of the human tuberculosis granuloma. *Nature Immunology*, **10**(9); pages 943–948. ISSN 1529-2916. 4

Ryan, T. J., Buddle, B. M., & de Lisle, G. W. (2000). An evaluation of the gamma interferon test for detecting bovine tuberculosis in cattle 8 to 28 days after tuberculin skin testing. *Research in Veterinary Science*, **69**(1); pages 57–61. ISSN 0034-5288. 8

Schurch, A. C., Kremer, K., Kiers, A., Daviena, O., Boeree, M. J., Siezen, R. J., Smith, N. H., & van Soolingen, D. (2010). The tempo and mode of molecular evolution of Mycobacterium tuberculosis at patient-to-patient scale. *Infection, Genetics and Evolution*, **10**; pages 108–114. 132

Shah, V. B. & McRae, B. H. (2008). Circuitscape : A Tool for Landscape Ecology. In *Proceedings of the 7th Python in Science Conference*, pages 62–65. 73

Skuce, R. A., Mallon, T. R., McCormick, C. M., McBride, S. H., Clarke, G., Thompson, A., Couzens, C., Gordon, A. W., & McDowell, S. W. J. (2010). Mycobacterium bovis genotypes in Northern Ireland: herd-level surveillance (2003 to 2008). *Veterinary Record*, **167**; pages 684–689. 25, 32, 34, 53, 60, 63, 91, 100, 103, 105, 127, 154

Skuce, R. A., McCorry, T. P., McCarroll, J. F., Roring, S. M. M., Scott, A. N., Brittain, D., Hughes, S. L., Hewinson, R. G., & Neill, S. D. (2002). Discrimination of Mycobacterium tuberculosis complex bacteria using novel VNTR-PCR targets. *Microbiology*, **148**; pages 519–528. ISSN 1350-0872. 124

Skuce, R. A., McDowell, S. W., Mallon, T. R., Luke, B., Breadon, E. L., Lagan, P. L., McCormick, C. M., McBride, S. H., & Pollock, J. M. (2005). Discrimination of isolates of Mycobacterium bovis in Northern Ireland on the basis of variable numbers of tandem repeats (VNTRs). *Veterinary Record*, **157**; pages 501–504. 24, 32, 33, 34, 43, 49, 101, 105, 124, 127

Skuce, R. A. & Neill, S. D. (2001). Molecular epidemiology of Mycobacterium bovis: exploiting molecular data. *Tuberculosis*, **81**(1/2); pages 169–175. 24, 25, 100

Sleeman, D. P., Davenport, J., More, S. J., Clegg, T. A., Collins, J. D., Martin, S. W., Williams, D. H., Griffin, J. M., & OBoyle, I. (2009). How many Eurasian badgers Meles meles L. are there in the Republic of Ireland? *European Journal of Wildlife Research*, **55**; pages 333–344. ISSN 1612-4642. 12

Smeds, L. & Kunstner, A. (2011). ConDeTri - A content dependent read trimmer for Illumina data. *PLOS One*, **6**(10); page s26,314. 105

Smith, G. J. D., Vijaykrishna, D., Bahl, J., Lycett, S. J., Worobey, M., Pybus, O. G., Ma, S. K., Cheung, C. L., Raghwani, J., Bhatt, S., Peiris, J. S. M., Guan, Y., & Rambaut, A. (2009). Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, **459**; pages 1122–1125. ISSN 1476-4687. 168

Smith, N. H., Dale, J., Inwald, J., Palmer, S., Gordon, S. V., Hewinson, R. G., & Smith, J. M. (2003). The population structure of Mycobacterium bovis in Great Britain: clonal expansion. *Proceedings of the National Academy of Sciences*, **100**(25); pages 15,271–15,275. ISSN 0027-8424. 23, 32, 33, 38, 49, 50, 55

Smith, N. H., Gordon, S. V., de la Rua-Domenech, R., Clifton-Hadley, R. S., & Hewinson, R. G. (2006). Bottlenecks and broomsticks: the molecular evolution of Mycobacterium bovis. *Nature Reviews Microbiology*, **4**; pages 670–681. ISSN 1740-1534. 21, 24, 25, 32, 33, 60, 100, 106, 127, 154

Smith, N. H. & Upton, P. (2012). Naming spoligotype patterns for the RD9-deleted lineage of the Mycobacterium tuberculosis complex; www.Mbovis.org. *Infection, Genetics and Evolution*, **12**; pages 873–876. ISSN 1567-7257. 156

Spatharis, S., Mouillot, D., Chi, T. D., Danielidis, D. B., & Tsirtsis, G. (2009). A niche-based modeling approach to phytoplankton community assembly rules. *Oecologia*, **159**; pages 171–180. ISSN 0029-8549. 49

Stamp, J. T. (1948). Bovine pulmonary tuberculosis. *Journal of Comparative Pathology and Therapeutics*, **58**(1); pages 9–23. 6

Supply, P., Lesjean, S., Savine, E., Kremer, K., van Soolingen, D., & Locht, C. (2001). Automated high-throughput genotyping for study of global epidemiology of Mycobacterium tuberculosis based on mycobacterial interspersed repetitive units. *Journal of Clinical Microbiology*, **39**(10); pages 3563–3571. ISSN 0095-1137. 32

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, **28**(10); pages 2731–2739. ISSN 1537-1719. 109

Tompkins, D. M., Ramsey, D. S. L., Cross, M. L., Aldwell, F. E., de Lisle, G. W., & Buddle, B. M. (2009). Oral vaccination reduces the incidence of tuberculosis in free-living brushtail possums. *Proceedings of the Royal Society B*, **276**; pages 2987–2995. 11

Torgerson, P. R. & Torgerson, D. J. (2009). Public health and bovine tuberculosis: what's all the fuss about? *Trends in Microbiology*, **18**(2); pages 67–72. ISSN 1878-4380. 16

Tsairidou, S., Woolliams, J. a., Allen, A. R., Skuce, R. a., McBride, S. H., Wright, D. M., Bermingham, M. L., Pong-Wong, R., Matika, O., McDowell, S. W. J., Glass, E. J., & Bishop, S. C. (2014). Genomic prediction for tuberculosis resistance in dairy cattle. *PLoS ONE*, **9**(5). ISSN 19326203. 15

van Soolingen, D. (2001). Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. *Journal of Internal Medicine*, **249**; pages 1–26. ISSN 0954-6820. 24

van Soolingen, D., de Haas, P. E. W., & Kremer, K. (2001). Restriction fragment length polymorphism typing of Mycobacteria. *Methods in Molecular Medicine*, **54**; pages 165–203. 103

Viana, M., Mancy, R., Biek, R., Cleaveland, S., Cross, P. C., Lloyd-Smith, J. O., & Haydon, D. T. (2014). Assembling evidence for identifying reservoirs of infection. *Trends in Ecology and Evolution*, pages 1–10. ISSN 01695347. 60, 99, 132

Vicente, J., Delahay, R. J., Walker, N. J., & Cheeseman, C. L. (2007). Social organization and movement influence the incidence of bovine tuberculosis in an undisturbed high-density badger Meles meles population. *Journal of Animal Ecology*, **76**; pages 348–360. ISSN 0021-8790. 134

Vordermeier, M., Gordon, S. V., & Hewinson, R. G. (2011). Mycobacterium bovis antigens for the differential diagnosis of vaccinated and infected cattle. *Veterinary Microbiology*, **151**(1-2); pages 8–13. ISSN 03781135. 9

Walker, T. M., Kohl, T. a., Omar, S. V., Hedge, J., Del Ojo Elias, C., Bradley, P., Iqbal, Z., Feuerriegel, S., Niehaus, K. E., Wilson, D. J., Clifton, D. a., Kapatai, G., Ip, C. L. C., Bowden, R., Drobniewski, F. a., Allix-Béguec, C., Gaudin, C., Parkhill, J., Diel, R., Supply, P., Crook, D. W., Smith, E. G., Walker, a. S., Ismail, N., Niemann, S., Peto, T. E. A., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., Eyre, D. W., Wilson, D. J., Hawkey, P. M., Crook, D. W., Parkhill, J., Harris, D., Walker, a. S., Bowden, R., Monk, P., Smith, E. G., & Peto, T. E. A. (2012). Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *The Lancet Infectious Diseases*, **3099**(12); pages 1–10. ISSN 1474-4457. 26, 27, 68, 99, 100, 112, 119, 124, 132, 169

Weber, N., Carter, S. P., Dall, S. R., Delahay, R. J., McDonald, J. L., Bearhop, S., & McDonald, R. a. (2013). Badger social networks correlate with tuberculosis infection. *Current Biology*, **23**(20); pages R915–R916. ISSN 09609822. 132

Welsh Government (2015). History of bovine TB controls. 15

White, P. C. L., Böhm, M., Marion, G., & Hutchings, M. R. (2008). Control of bovine tuberculosis in British livestock: there is no 'silver bullet'. *Trends in Microbiology*, **16**(9); pages 420–7. ISSN 0966-842X. 16

White, P. C. L. & Whiting, S. J. (2000). Public attitudes towards badger culling to control bovine tuberculosis in cattle. *Veterinary Record*, **147**; pages 179–184. 17

WHO (2011). Global tuberculosis control 2011. Technical report, World Health Organsiation, Geneva. 3

Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C., Brisse, S., Meyer, A., Supply, P., & Niemann, S. (2008). Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathogens*, **4**(9); page e1000,160. ISSN 1553-7374. 21, 107

Womble, W. H. (1951). Differential systematics. *Science*, **114**; pages 315–322. ISSN 0036-8075. 74

Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*. Chapman and Hall/CRC. 69, 70, 78, 141, 150

Wood, S. N., Bravington, M. V., & Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society B*, **70**(5); pages 931–955. ISSN 13697412. 70, 141

Woodroffe, R., Donnelly, C. a., Jenkins, H. E., Johnston, W. T., Cox, D. R., Bourne, F. J., Cheeseman, C. L., Delahay, R. J., Clifton-Hadley, R. S., Gettinby, G., Gilks, P., Hewinson, R. G., McInerney, J. P., & Morrison, W. I. (2006). Culling and cattle controls influence tuberculosis risk for badgers. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(40); pages 14,713–14,717. ISSN 0027-8424. 14, 158, 165

Woodroffe, R., Donnelly, C. A., Johnston, W. T., Bourne, F. J., Cheeseman, C. L., Clifton-Hadley, R. S., Cox, D. R., Gettinby, G., Hewinson, R. G., le Fevre, A. M., McInerney, J. P., & Morrison, W. I. (2005). Spatial association of Mycobacterium bovis infection in cattle and badgers Meles meles. *Journal of Applied Ecology*, **42**; pages 852–862. ISSN 00218901. 3, 18, 25, 32, 60, 61, 92, 154, 162, 164

Woodroffe, R., Donnelly, C. A., Wei, G., Cox, D. R., Bourne, J. F., Burke, T., Butlin, R. K., Cheeseman, C. L., Gettinby, G., Gilks, P., Hedges, S., Jenkins, H. E., Johnston, W. T., McInerney, J. P., Morrison, W. I., & Pope, L. C. (2009). Social group size affects Mycobacterium bovis infection in European badgers (Meles meles). *Journal of Animal Ecology*, **78**; pages 818–827. ISSN 1365-2656. 13

Woolhouse, M. E. J., Dye, C., Etard, J. F., Smith, T., Charlwood, J. D., Garnett, G. P., Hagan, P., Hii, J. L. K., Ndhlovu, P. D., Quinnell, R. J., Watts, C. H., Chandiwana, S. K., & Anderson, R. M. (1997). Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proceedings of the National Academy of Sciences*, **94**; pages 338–342. ISSN 00278424. 55

Worby, C. J., Lipsitch, M., & Hanage, W. P. (2014). Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Computational Biology*, **10**(3); page e1003,549. ISSN 1553-7358. 167

Wright, D. M., Allen, A. R., Mallon, T. R., McDowell, S. W. J., Bishop, S. C., Glass, E. J., Bermingham, M. L., Woolliams, J. A., & Skuce, R. A. (2013a). Detectability of bovine TB using the tuberculin skin test does not vary significantly according to pathogen genotype within Northern Ireland. *Infection, Eenetics and evolution*, **19**; pages 15–22. ISSN 1567-7257. 21, 33, 124

Wright, D. M., Allen, A. R., Mallon, T. R., McDowell, S. W. J., Bishop, S. C., Glass, E. J., Bermingham, M. L., Woolliams, J. A., & Skuce, R. A. (2013b). Field-isolated genotypes of Mycobacterium bovis vary in virulence and influence case pathology but do not affect outbreak size. *PloS One*, **8**(9); page e74,503. ISSN 1932-6203. 21, 33, 124

Wu, H., Wang, Y., Zhang, Y., Yang, M., Lv, J., Liu, J., & Zhang, Y. (2015). TALE nickase-mediated SP110 knockin endows cattle with increased resistance to tuberculosis. *Proceedings of the National Academy of Sciences*, **In press**; pages E1530–E1539 —. ISSN 0027-8424. 15

Young, J. S., Gormley, E., Wellington, E. M. H., Elizabeth, M. H., & Wellington, E. M. H. (2005). Molecular detection of Mycobacterium bovis and Mycobacterium bovis BCG (Pasteur) in soil. *Applied and Environmental Microbiology*, **71**(4); pages 1946–1952. 119

Ypma, R. F., Altes, H. K., van Soolingen, D., Wallinga, J., & van Ballegooijen, W. M. (2013). A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. *Epidemiology*, **24**(3); pages 395–400. 32, 33, 34, 38, 49, 50, 55