
Copyright © 2012 The Authors

http://eprints.gla.ac.uk/73731/

Deposited on: 17 January 2013
Whole Genome Sequencing Reveals Local Transmission Patterns of *Mycobacterium bovis* in Sympatric Cattle and Badger Populations

Roman Biek1, Anthony O’Hare1,*, David Wright2,3, Tom Mallon2, Carl McCormick2, Richard J. Orton1, Stanley McDowell2, Hannah Trewby1, Robin A. Skuce2, Rowland R. Kao1*

1 Boyd Orr Centre for Population and Ecosystem Health and Institute for Biodiversity, Animal Health and Comparative Medicine, University of Glasgow, Glasgow, United Kingdom, 2 Veterinary Sciences Division, Agri-Food and Biosciences Institute, Stormont, Belfast, Northern Ireland, United Kingdom, 3 School of Biological Sciences, Queen’s University Belfast, Belfast, Northern Ireland, United Kingdom

**Abstract**

Whole genome sequencing (WGS) technology holds great promise as a tool for the forensic epidemiology of bacterial pathogens. It is likely to be particularly useful for studying the transmission dynamics of an observed epidemic involving a largely unsampled ‘reservoir’ host, as for bovine tuberculosis (bTB) in British and Irish cattle and badgers. BTB is caused by *Mycobacterium bovis*, a member of the *M. tuberculosis* complex that also includes the aetiological agent for human TB. In this study, we identified a spatio-temporally linked group of 26 cattle and 4 badgers infected with the same Variable Number Tandem Repeat (VNTR) type of *M. bovis*. Single-nucleotide polymorphisms (SNPs) between sequences identified differences that were consistent with bacterial lineages being persistent on or near farms for several years, despite multiple clear whole herd tests in the interim. Comparing WGS data to mathematical models showed good correlations between genetic divergence and spatial distance, but poor correspondence to the network of cattle movements or within-herd contacts. Badger isolates showed between zero and four SNP differences from the nearest cattle isolate, providing evidence for recent transmissions between the two hosts. This is the first direct genetic evidence of *M. bovis* persistence on farms over multiple outbreaks with a continued, ongoing interaction with local badgers. However, despite unprecedented resolution, directionality of transmission cannot be inferred at this stage. Despite the often notoriously long timescales between time of infection and time of sampling for TB, our results suggest that WGS data alone can provide insights into TB epidemiology even where detailed contact data are not available, and that more extensive sampling and analysis will allow for quantification of the extent and direction of transmission between cattle and badgers.

**Introduction**

The application of whole genome sequencing (WGS) technology to infectious bacterial diseases has resulted in unprecedented advances in our ability to resolve epidemic data at the global scale [1,2], provided new insights into within-host replication processes [3], and been used to corroborate the importance of exhaustively identified transmission chains and social drivers of transmission [4,5]. However, evidence of its value when observing fine scale transmission dynamics in a partially sampled population is thus far limited to some tantalising observations [1] with as yet no quantitative evaluation of the underlying contact processes using nonlinear mathematical models. Such evaluations are particularly important where the sampling is biased, such as when the epidemiology involves an unobserved ‘reservoir’ host. Interpretation of sequence data at this scale is further complicated by the lag between the time of transmission and the time of sample collection, which can be considerable, especially for pathogens with extended latent periods. Comparing genetic data with mathematical models based on epidemiological contact data should allow us to develop more general transmission principles, and to compare and contrast the types of information that these different data sources provide. Combining mathematical models and pathogen sequence data has been an area of increased attention in the epidemiology of fast-evolving RNA viruses [6–9], but is as yet largely unexplored for bacteria, particularly for TB and other slow growing mycobacteria, where transmission intervals and routes tend to be more uncertain and evolutionary rates are slower, warranting more extensive sequence information. While this presents a unique set of challenges, WGS now offers promising research solutions to this problem.

*Mycobacterium bovis* is the causative agent of bovine tuberculosis (bTB), an important disease of both livestock and wildlife with...
severe socio-economic consequences and impacts on animal health. Historically, it is believed to have been a major contributor to human TB cases worldwide, and it remains a zoonotic concern in both developed and developing countries [10,11]. While most countries that employ well developed test and slaughter programs have eliminated bTB from their livestock, the control of *M. bovis* data, but cattle movement networks and within-herd transmission patterns generated by mathematical models do not. Thus WGS offers considerable promise for revealing basic principles about bTB maintenance in British cattle and the role of badgers, as well as suggesting that similar approaches combining mathematical models and WGS data could be useful for the study of human TB and other infectious diseases where sampling biases are known to exist.

### Results

WGS from the 31 bTB isolates revealed a total of 39 polymorphic sites, of which 7 were shared amongst two or more isolates (Table S2 in Text S1). One of these shared polymorphisms was derived from *M. bovis* and *M. tuberculosis* (H37Rv). Using the known sampling dates, we estimated the evolutionary rate was 3.40 (CI: 0.87–5.93) substitutions per site per year (equivalent to 0.147 substitutions per genome per year), about an order of magnitude lower than within-host mutation rates recently estimated for *M. tuberculosis* [3].

The inferred phylogeny revealed that most outbreaks involved genetically distinct isolates (Figure 2) demonstrating the suitability of WGS to track *M. bovis* spread at the herd level. The more extensively sequenced outbreaks were dominated by a single common genotype (see below). Repeated outbreaks within the same herd tended to involve closely related isolates falling into the same genetic lineage (e.g. Herd 1: 1999/2004; Herd 5: 2008/2010). The exception was Herd 3, which fit this pattern from 1999 to 2004 and from 2007 to 2010 but with distinct lineages causing...
outbreaks during those two periods. Though few badger isolates were available, their genetic distances from cattle isolates were small and comparable to those observed among cattle isolates; two of the five \textit{M. bovis} genomes sequenced from badgers were genetically indistinguishable from those seen in cattle during the same year. No \textit{M. bovis} sample taken from a badger was more than four mutational steps away from the nearest cattle isolate. Interestingly, the two isolates from different tissues of the same badger were separated by five mutational steps, as great a genetic distance as found across different cattle in serial outbreaks within the same cattle herd. This suggests either multiple infections of the same animal or a long-term infection that had allowed for within-host divergence to evolve. In contrast, isolates collected during a series of four outbreaks in Herds 3 and 5 between 2007 and 2010 showed little divergence, with 9 out of the 20 genomes being indistinguishable based on our data, despite the considerable timeframe involved.

Outbreaks could be epidemiologically linked through local transmission, which may for example involve badgers, unobserved infections in cattle, environmental contamination or contiguous contact between farms. Alternatively, transmission may be due to the network of livestock movements that connects the five farms through shared links to other farms. We compared the number of mutations separating isolates to two proxy measures of epidemiological distance, 1) the Cartesian distance between the main holding locations of cattle herds, and 2) one minus the relative probability that herds had been connected by movement of infected cattle, called here a “network separation” (see Materials and Methods for our definition of this).

As there is a strong genetic and spatial autocorrelation for samples obtained from single herds, we compared only across herds, while noting that this does not completely eliminate the dependence problem. We found a positive relationship between the pair-wise Cartesian distance among farms and the smallest genetic distance among their bTB isolates (Figure 3A), consistent with at least one of the possible local transmission mechanisms being important. As would be expected, where different lineages are considered as separate outbreaks (e.g. the two lineages

\textbf{Figure 1. Main holdings associated with herds in the dataset and badger locations by year.} Herd locations indicate centroids of main holdings, and do not include isolated land parcels or rented land. doi:10.1371/journal.ppat.1003008.g001
associated with Herd 3), there was a much poorer correlation \( R^2 = 0.444 \) for the earlier outbreaks compared to the later ones.

To construct the cattle network, life history data were extracted for all cattle in the database from 01/01/1990 to 31/12/2010 as well as all farms in NI recording a breakdown due to VNTR type 10 between 1999 and 2010, amounting to 58 herds, 3321 cattle and 14258 recorded individual tuberculin test results from 29/07/1993 to 09/12/2010, plus records of post-mortem examination for tuberculous lesions and subsequent confirmatory tests including histology and culture. The movement network included movements amongst all herds in this group of 58 farms. The resulting network was highly connected, including our five farm cluster (Figure S3). The only two farms directly connected through recorded cattle movements were Herds 3 and 5 but this reflected a single animal and, in the absence of further indirect contact, corresponded to a high network separation. In contrast, other farm pairs were much more strongly connected through indirect contact, generally involving one or two intermediary farms. However, these reduced network separations correlated only weakly with genetic distance (Figure 3B).

Because of the indeterminate and potentially long timescales for \( M. \) bovis transmission, the relationship between dates of sampling and dates of transmission are uncertain. Fitting a mathematical model of transmission to the observed life history data allowed us to compare likely epidemiological processes to the observed tree structure, in order to analyse dynamics at the within-herd, individual animal level. The long period from 2007/8 to 2010 during which no reactors were identified in Herds 3 and 5, combined with the large scale depopulation that occurs following identification of a breakdown (all test reactors are slaughtered and herds are tested every two months until two consecutive clear whole herd tests are achieved), would suggest multiple introductions of infection into these herds. In contrast, the ‘flat’ genetic structure for the same outbreak, with no obvious chains, would suggest one outbreak from a single source. We therefore asked whether the recorded cattle life history and movement data, interpreted in the context of a simple nonlinear model, are consistent with the observed genetic signature. We constructed a compartmental model of transmission, assuming all animals were in one of four states: susceptible to disease (S), exposed (E), potentially test positive but not yet infectious (T) or infectious (I) [18]. Exploiting the explicit animal life histories in all five herds (Figure 4), plus import/export data from the remaining 53 herds in the group, state probability distributions for each animal were generated based on the known dates of bTB detection [6,19].

Figure 2. Maximum likelihood network of \( M. \) bovis genomes with tips arranged according to sampling date. Position of other nodes is simply shown for convenience and does not reflect known branch times. Black circles represent single nucleotide polymorphisms separating sequences, dashed lines indicates branches without mutational events. The size of the circle proportional to the number of isolates sharing the same sequence.

doi:10.1371/journal.ppat.1003008.g002
infection due to latent or hidden infections, where each non-reactor is assumed to contribute to the overall force of infection in proportion to the calculated probability that itself had been infected. The overall force of infection was calculated assuming homogeneous, density dependent mixing within each herd and with contacts between herds via recorded cattle movements. A best-fit model was determined using a likelihood-based Markov Chain Monte Carlo approach (further details in the Materials and Methods section below and in Figures S4 to S6, including posterior distributions for all parameters) [20].

**Figure 3. Genetic versus spatial and network distances.** On both axes, all values are scaled to the maximum value for herd-to-herd interactions. On the x-axes, the minimum number of SNPs differentiating isolates from the two herds (X-Y). In panel (A) above, spatial distance versus genetic distance between herds (black squares). On the y-axis, the cartesian distance between the main holdings of two herds (X-Y) showing a high level of correlation with genetic distance \( R^2 = 0.720 \). For reference, the equivalent data for the badger isolates (not fitted) are shown as unfilled circles and diamonds for badger-badger and badger-herd relationships, respectively. Panel (B) below, network separation versus genetic distance between herds. On the y-axis, the network separation defined as \( (1 - p_{ij}) \), where \( p_{ij} \) is the probability that herds i and j are linked via cattle movements through the network, considering all possible pathways through any herd from which the same genotype of M. bovis has been isolated, and \( p_{\text{animal}} \) is the per animal probability of contact that best explains the genetic distance data. The best fit value \( (p_{\text{animal}} = 1.35 \times 10^{-3}) \) shows a poor correlation with genetic distance \( R^2 = 0.094 \).

doi:10.1371/journal.ppat.1003008.g003
The considerable overlap in cattle life histories (Figure 4) could have potentially allowed for long within-herd epidemics. However, the probabilities of a transmission chain linking cattle-based isolates suggest that while cattle-to-cattle transmission can likely explain some maintenance of individual genetic lineages (at its peak, the cattle to cattle force of infection was calculated to be roughly 50× that from hidden sources), new outbreaks are usually better explained by an unspecified ‘reservoir’ (Figure 5). In particular, despite similar overlaps in reactor life histories across the entire recorded history of Herd 3, and similar observed genetic distances, the earlier outbreak is epidemiologically distinct from the later outbreak in 2007 to 2010, with separate introductions suggested for 1999 and 2004 unless at least one particular individual had an uncharacteristically long infectious period. In contrast, the 2007 to 2010 outbreaks in Herds 3 and 5 are well supported by cattle-to-cattle transmission alone. Herds 3 and 5 are the only ones recorded as having directly traded with each other, however our herd-to-herd analysis suggests that this link was

Figure 4. Life histories of all cattle with from which *Mycobacterium bovis* samples of VNTR type 10 were obtained. Showing all individuals residing within the five study herds at some time from 1994 to 2010. Cattle residence times indicated by the length of the horizontal bars (each bar representing a single animal), in black, all cattle from which sequences are derived (herd indicated by surrounding type). Test dates on which one herd received a whole herd test are indicated by vertical dashed lines. Herd colours correspond to colours in Figure 1 (1 – pink, 2 – purple, 3 – blue, 4 – orange, 5 – red).

doi:10.1371/journal.ppat.1003008.g004
unlikely to have connected the outbreaks in the two herds (Figure 3). All other herds appear only linked by the external force of infection, which was of the same order of magnitude as the force of infection due to a single infectious animal (see parameter estimates in Figure S5). Correlations between genetic distances and network distances are poor (Spearman rank correlation coefficient $r = -0.326$) with any correlation due to between- rather than within-herd differences, indicating that the WGS data are not of sufficient resolution to make inferences on transmission chains at the within-herd scale.

**Discussion**

Measuring genetic variation at the whole genome scale enabled us to genetically distinguish most isolates of *M. bovis*. This is particularly notable given the small spatial extent of the study cluster, with no two farms being more than 5 km apart. Compared to traditional typing methods, for which the same genotype may be distributed over many hundred square kilometres and only broad associations can be rigorously defined [21], WGS affords a level of resolution for epidemiological studies previously limited to rapidly evolving RNA viruses [6–8].

In addition to most isolates and outbreaks being genetically unique we found that subsequent outbreaks on the same farm tended to involve the same genetic lineage previously detected in that location. This indicates that lineages are commonly able to persist locally within the direct environment of a farm, although the mechanisms for this are not yet clear. Results of our mathematical models based on individual cattle histories indicate that persistence on a farm is overall poorly explained by ongoing
transmission within herds. In the cases of Herds 1 and 3 for example, the model identified independent introductions for subsequent outbreaks (Fig 5), despite the fact that these serial outbreaks involved the same genetic lineage (Fig 2). Based on these findings, the detected infections (including unsequenced reactors) are insufficient in explaining local persistence on farms, instead suggesting a number of possible alternative mechanisms, such as re-introduction of the same lineage from neighbouring herds, environmental persistence, or alternative hosts. In contrast, and despite the relative simplicity of the modelling approach used, the persistence of the outbreaks in Herds 3 and 5 from 2002 to 2010 are consistent with lineage persistence resulting from extensive within-herd transmission, despite the multiple clear whole herd tests that would have occurred in between dates for reactors. While forward simulations were used to corroborate the robustness of our modelling approach, any extrapolation of our results for bTB epidemiology in general must be viewed with caution, both because of the small size of the dataset and because some of the modelling assumptions (in particular the assumption of explicitly time dependent generation times, see supplementary information) were not intended to be mechanistic descriptions of the underlying transmission process. Nevertheless, the fact that such a parsimonious model identifies a cattle-only contact structure largely consistent with the observed phylogeny generates confidence in our results.

In addition to local persistence, we also found evidence for the introduction of new genetic lineages onto farms and our analyses allow us to partially resolve how these introductions occur. Though cattle movements are a known risk for between-herd spread of bTB in Britain [22,23], they do not appear to be important for the events observed here, as the probabilities of transmission amongst herds involved in the extensive network of all observed VNTR10 outbreaks only poorly predict the genetic divergence amongst isolates. In contrast, we found Cartesian distance to be a good predictor of genetic distance among isolates at a very fine scale. Though the small sample size means that inferences regarding between-herd contacts should be viewed with caution, these results are consistent with the relatively low importance accorded to movements compared to local risk factors in bTB endemic areas that was previously observed at a national scale in GB [23], and suggests that a more extensive analysis of the balance between local spatial and network processes would be merited. As it stands, the most parsimonious explanation for these outbreaks involves a local transmission process that could be due to a number of causes. A non-exhaustive list of these includes both cross-boundary contact or unrecorded local movements between herds and transmission from a common badger reservoir (where the interaction is spatially localised, consistent with the badger’s stable social structure and strong territory [24]), or a combination of these factors. While our sample size for badgers is low, the badger-derived sequences are remarkably similar to those in cattle, demonstrating very recent cross-over events between the two populations, or alternatively recent infections from a common source, such as the environment [25]. The demonstration of a high M. bovis diversity in a single badger suggests either a lengthy infection in that badger, or multiple exposures to different sources of infection.

Although our current estimate for the rate at which M. bovis genomes evolve must be considered preliminary, it is considerably slower than the rate observed in M. tuberculosis [3]. Should it be confirmed, this has obvious implications for the level of temporal resolution that WGS can provide for unravelling epidemiological dynamics for bTB. In our current data, we were unable to genetically resolve relationships among multiple isolates stemming from the same outbreak for example and saw serial outbreaks commonly involving identical genotypes. This is corroborated by the poor correlation between the genetic distances and our estimates of the within-herd contact structure. However, apart from limiting opportunities for molecular epidemiological inference, these observations may also hold clues with respect to M. bovis biology and transmission. A recent study conducting experimental infections with M. tuberculosis in primates, found mutation rates to be equivalent during latent and active infections and proposed oxidative damage as a potential mechanism [9]. If this is relevant to M. bovis, one could hypothesise that the slower rates of evolution seen here at the population level, could be caused by the bacterium spending extended periods outside the host, in the environment. Future studies and analyses are needed to obtain more accurate estimates for the genomic rate of evolution in M. bovis and to test for potential rate heterogeneity and its underlying factors.

While cattle movements and long-term, hidden persistence within herds have both been shown to contribute significantly to herd breakdowns [22,23,26], these previous analyses were aimed at the identification of statistical associations; here we have shown that WGS data are able to identify local interactions as the principle culprit in specific herds. This makes WGS both a valuable tool for forensic epidemiology, and an aid in the construction of improved mathematical and statistical models of disease dynamics. In contrast, the poor correlation between network and genetic distance at the within-herd level suggests important limits to the resolution that WGS can provide for this system. The local effects identified here may be due to the local infected badger population, but are also consistent with local herd-to-herd spread. Our simplified modelling approach was chosen to maximise the use of available epidemiological contact data, but at the expense of a more detailed exploration of the possible hypotheses regarding the sources of transmission. However, it is likely that WGS based on more extensive sampling will allow for more sophisticated approaches, that could be used to directly estimate the role of badgers in the maintenance of bTB in British and Irish cattle. While insights into particular disease problems will depend on many factors we cannot consider here, our study supports the proposition that WGS data alone can provide insight into the impacts of unobserved populations on observed epidemics even in the absence of detailed transmission chain information, for M. bovis, other members of the M. tuberculosis complex, and other pathogens involving reservoir hosts.

Materials and Methods

VNTR typing

M. bovis is a member of the closely related Mycobacterium tuberculosis complex, which consists of several species with a shared ancestry [27] but which have evolved marked but not absolute host preferences [28,29]. On a global scale, the M. tuberculosis complex can now be subdivided into discrete lineages, which show strong phylogeographical localisation to regions [30,31]. VNTR profiling is a genotyping technique based on determining the copy number of a series of short, simple DNA repeats, originally identified by genome analysis [32]. However, while mutations in VNTR type have been observed within the timescale of observation at the regional level, in most cases, probable transmission events are associated with the sameVNTR-type, therefore requiring finer resolution typing to order the members of these groups.

Herd-level M. bovis genotyping has been performed by the AgriFood and Biosciences Institute, Belfast, UK since 2003, as follows.
The first (disclosing) *M. bovis* isolate from all bovine TB herd incidents was subjected to genotyping (eight-VNTRs and spoligotyping convention). Heat-inactivated *M. bovis* cell lysates were used directly as PCR-ready templates. VNTR profiling, spoligotyping, nomenclature, reference strains and quality control were as described [32]. The inferred tandem repeat copy number at each VNTR locus was used to produce a concatenated multi-locus VNTR profile (a string of integers), which was then simplified to a number indicating the prevalence of that profile. Genotype 001 (SB0140), with a spoligotype of SB0140, was the most prevalent in Northern Ireland when surveyed in 1999 to 2003 [32]. Spoligotypes were named according to an agreed international convention (www.mbovis.org).

**Datasets**

Anonymised records of cattle tuberculin tests, farm locations and cattle movements among them were drawn from the Animal and Public Health Information System (APHIS), a database containing details of all cattle in Northern Ireland that is administered by the Department of Agriculture and Rural Development [33]. The locations (main farm building) of 58 herds that had a breakdown with VNTR type 10 between 1999 and 2010 were extracted from APHIS. A cluster of five of these herds was selected based on their spatial clustering and the proximity of badger-derived isolates of type 10. The life histories of all 3299 cattle that passed through them since 1995 were compiled, comprising birth and death dates, and the dates of movements into and out of the cluster herds. These animals underwent routine bTB skin testing every year and the lifetime test history of each animal was extracted, containing the dates and results of all tuberculin tests (a total of 14258 individual test results). Results of post-mortem examination for tuberculous lesions and any subsequent confirmatory tests (laboratory based histology, culture and VNTR typing) were also incorporated.

Cattle movement was investigated by extracting a broader dataset, comprising all herds that animals passing through the 58 VNTR 10 herds had also visited (a total of 14 096 herds, excluding livestock markets). All cattle movements among the expanded set between 1992 and 2010 were extracted, a total of 5,875,510 individual animal movements.

**DNA extraction and sequence analysis**

*M. bovis* was isolated and confirmed from suspect bovine granulomatous tissue using standard protocols. Confirmed cultures were grown to single colonies on LJ slopes and single colonies were amplified for DNA extraction using the standard CTAB and solvent extraction protocol [34]. Extracted DNA was sequenced at the Sir Henry Wellcome Functional Genomics Facility at the University of Glasgow using an Illumina Genome Analyser IIX. Pair-end reads of 70 bp in length, separated by an average of about 350 bp, were trimmed from both ends based on phred quality scores so as to result in an error rate of 0.001 or less for each base call in the remaining sequence. Reads were mapped to a published UK reference genome (AF212297) [35] using the Geneious assembler under the “medium-low sensitivity” option, allowing for a maximum of 10% gaps and mismatches per read [36]. The reference sequence belongs to the same spoligotype (SB0140) as VNTR type 10 and shares identical repeat numbers with it for four out of eight loci used for typing. Mapping resulted to an error probability of 1 in 1,000,000) was applied to each position to ensure that accuracy of the final consensus sequence was dependent on both quality and read depth, rather than read depth alone. Below this threshold, the consensus base call was scored as unknown (‘N’). Alignment of consensus sequences was carried out using Mauve [37], as implemented within Geneious, assuming collinear genomes and with automatic calculation of seed weight and of the minimum Locally Collinear Blocks (LCB) score. Regions that were difficult to align or which contained $\geq$3 consecutive columns of unknown bases or gaps were removed from the final alignment. Similarly, sites that were polymorphic solely due to one or more sequences having ambiguity base calls were removed; this was the only context in which ambiguities were observed. The final alignment, which still represented 99.2% of the reference genome, thus only contained dimorphic single site polymorphisms situated within otherwise invariable regions.

After stripping identical sites, a total of 39 SNPs were identified (38 substitutions, 1 deletion, Table S2 in Text S1), of which seven were shared between two or more sequences. All SNPs were examined to confirm their validity before further analysis. Of particular concern was the potential inclusion of spurious SNPs associated with repeat regions in the genome for which mapping may be unreliable. While four of the SNPs were found to fall either in or close to potentially problematic regions, the reliability of the mapping and SNP calling could be confirmed in all four cases (see Supplementary Materials). All SNP calls were supported by at least 38x coverage, with high consistency among reads (usually>95%). The only exception to this was a SNP in position 221927 (G $\rightarrow$ A), for which consensus calling was ambiguous in one of the four isolates in which it occurred (Herd3_E_2010, 92x, A: 64%, G: 36%, Table S2 in Text S1). Preliminary analyses further showed that the phylogenetic information provided by this site was in conflict with that of other informative positions (which were in complete agreement). Because these observations raised doubts about the reliability of scoring this SNP as well as about the information it provided, the site was removed from the data set. The final data set was used to generate a maximum likelihood tree using phylm [38] under a Jukes-Cantor model using a heuristic search and the reference genome for outgroup rooting.

All sequence data generated for this project are available from the European Nucleotide Archive (http://www.ebi.ac.uk/ena/) under accession number ERP001418.

**Modelling of transmission chains**

Using the anonymised life history data of the herds containing genotyped isolates, we adapted a method previously used to study foot and mouth disease virus epidemiology and transmission phylodynamics at the between-farm scale [6,19]. All cattle were assumed to be in one of four states that reflect the critical points in the cattle infection process [18,39–41]: susceptible (S), exposed (E), where an animal is infected but neither tests positive nor infects others, potentially test positive (T), where the animal can test positive but is not yet infectious and infectious (I), with a constant transition rate from the E to T states and from the T to I states. States T or I can be truncated by a positive test. Knowing the dates each reactor (an animal that tests positive) was detected, we use the transition rates to determine the set of distribution functions for each animal describing the probability of being in each of the four infection states at any time in its life. We used these distributions to determine the likelihood of the observed pattern of reactors in the cluster of herds.

This key simplification (conditioning on the observed test data) is less rigorous than an approach where the likelihood is integrated forward across all possible dates of infection and the observed data...
is treated as a random variable. However the latter approach would have been computationally prohibitive for our system. The simplification implies that mean generation times are consistent over the observed timeframe, even though they would naturally be expected to contract over the course of an epidemic [42]. The assumption is expected to be reasonable if the disease is in an endemic state, or an endemic state is rapidly reached after the initial introduction, where ‘rapid’ is relative to the observed timescale – i.e. the distribution across states is ‘quasi-stationary’, with E and T proportional to the reciprocal of their respective transition rates, i.e. 1/σ and 1/γ respectively. Though observation of our incidence data suggested this to be likely, to test this assumption, we allowed the proportion in state I at the time of detection to vary over the course of the epidemic, with probability \( q_I(t) \) that it is infectious (i.e. in the I state) and \((1 - q_I(t))\) for being in the T state, where \( t \) is the time since the genotype was initially detected. Here, we have assumed that \( q_I(t) \) has the form \( q_I(t) = a t + b \), where \( a \) and \( b \) are fitted constants and \( q_I(t) \) is a probability with range \([0…1]\).

We assumed that all cattle are equally infectious if in state ‘T’ but that an infected animal does not contribute to further infection after it is detected; most reactor cattle are removed from the herd within days of being identified. All infection is caused by either infectious cattle or by a reservoir (either undetected infected cattle within the herd or external, probably local, factors).

At this population scale, we assumed that the probability of a false positive is negligible. To account for untested and undetected contributions to the epidemic, we assumed all test-negative/untested cattle to have a fitted probability of being infectious.

The date(s) of an animal “\( t^+ \)” having a positive test is denoted as \( t^+ \), at which it may be infectious with a probability \( q_I(t^+ - \text{timebreak}) \), where \( \text{timebreak} \) is the time the outbreak was initially detected in the herd and \( t \) is equal to \( t^+ \) at the point of evaluation. We assumed that a reactor in the infectious state became infectious (i.e. moved from the test sensitive state) within a maximum of \( T^+ \) days prior to the positive test. \( T^+ \) is the maximum infectious lifetime of the animal prior to detection. Because a uniform test is applied to all cattle over 6 months of age on an annual basis, the date of first test after infection would fall in the range of 0 to 1 year; therefore we assumed that the duration of the effective infectious period is uniform. This has the (conservative) effect of minimising the potential duration of cattle-only transmission chains since it does not allow for long infectious periods. The probability that a reactor was infectious at any time prior to the date it tested positive is:

\[
P_I^d(t) = \begin{cases} q_I(t^+ - \text{timebreak}) \frac{(t-t^+)^{t^+}}{t^+}, & t > t^+ - T^+ \\ 0, & \text{otherwise} \end{cases}
\]

All other model states were assumed to be exponentially distributed. The probability that an individual is test sensitive at a given time \( t \) is therefore

\[
P_T^d(t) = (1 - q_I(t^+ - \text{timebreak}) e^{-\gamma(t^+ - \gamma)} + \]

\[ q_I(t^+ - \text{timebreak}) \int_{t^+}^{t} \delta P_T^d(t') e^{-\gamma(t - t') dt'}, \]

where \( \gamma \) is the transition rate from the T to the I states in the forward SETI model. The parameter \( \gamma \) determines the duration of the state T and therefore the time when the backwards T to E transition occurs, conditional on the backwards transition from I to T and the time of detection \( t^+ \). The first term accounts for the animal being in the potentially test positive state at \( t^+ \) and the second term accounts for the animal being infectious at \( t^+ \). Here, \( \delta P_T^d(t) \) is the increase in \( P_T^d(t) \) that occurs due to transition from \( P_T^d(t) \) at time \( t \). Similarly, the probabilities of being in the exposed and susceptible states are

\[
P_T^e(t) = \left(1 - \int_{t}^{t^+} P_T^d(t') e^{-\gamma(t' - t)} dt'\right) e^{-\gamma(t - t^+)} + \]

\[ q_I(t^+ - \text{timebreak}) \int_{t^+}^{t} \delta P_T^e(t') e^{-\gamma(t' - t)} dt', \]

where \( \sigma \) is the transition rate from the E to the T states in the forward-in-time model. The forward-in-time transition from the S to the E state is determined by \( \beta \) and is estimated via the calculated force of infection based on the probabilities of being in the I state. It is incorporated directly into the calculation of the likelihood as described below.

A similar approach was used for animals identified by a post mortem test in which case \( t^+ \) is the date the animal was post mortem examined. We assumed further that post mortem testing only identifies the same infection classes as the standard tuberculin test (i.e. only T or I class individuals are detected).

We calculated the probability that an animal became infected by considering the forces of infection on it during its lifetime. The life history of each animal that resided in the cluster between January 1990 and 31 December 2010 was converted into a lookup table of the herds in which it resided in fortnightly time steps. If an animal moved between herds in a time step it was treated as belonging to both herds since it can contribute to an outbreak in both herds during that step. Animals born into a herd are added to the lookup table for that time step and conversely death results in an animal being removed from the lookup table – movements back to the herd cluster via other herds are also retained. Using this lookup table we were able to calculate the probability an animal became infected in each period.

We calculate the probability that an animal would have become infected at any time \( t \), \( P_{\text{inf}}(t) \), given that is was a known reactor based on the force of infection calculated from the model. If an animal was a reactor, then the incremental probability \( dP_{\text{inf}}(t) \) that an animal, \( i \), was infected by all sources over the time interval \( t \) is then given as

\[
dP_{\text{inf}}(t) = P_{\text{inf}}(t) \left(x + \beta \sum_{i \neq i} P_{\text{inf}}^i(t)\right) dt,
\]

(1) calculated using the explicit infection histories. If the animal never tested positive then

\[
dP_{\text{inf}}(t) = \left(1 - \int_0^t dP_{\text{inf}}(t') dt'\right) \left(x + \beta \sum_{i \neq i} P_{\text{inf}}^i(t)\right) dt,
\]

(2) i.e. we consider the probability that an animal was infected despite never testing positive. Young calves destined for slaughter at a young age are test exempt and are considered to be a negligible risk due to their short lifespan; they are therefore excluded from the analysis. The parameter \( \beta \) is the transmission rate from a contact with an infected animal in the same herd at the same time (the summation is over animals in the same herd at \( t \)). The reservoir term, \( x \), can be written as \( x = \lambda_{\text{ext}} + P_{\text{inf}}^\text{ext}(N(t) \text{ i.e. a combination of any external forces of infection, } \lambda_{\text{ext}}, \text{e.g. infected wildlife}) \) that is represented by a fixed rate and \( N(t) \), the fraction of
We calculated the probability each animal was infected during its lifetime by integrating $dP_i^{inf}(t)$ over the animals’ entire life (approximated by a summation over fortnightly time steps) and thus calculate the likelihood for the model as:

$$L = \prod_{\text{reactors}} P_i^{inf} \prod_{\text{non-reactors}} \left(1 - \Omega (1 - P_i^{inf}) \right).$$

Here, $\Omega$ is the fitted sensitivity of the test for bTB infection. Consistent with the low number of reactors in officially TB free countries such as Scotland, the test specificity is assumed to be effectively 100%. Using $L$, we used the Metropolis-Hastings algorithm to derive the posterior distributions over the parameter space defined by $(\beta, \sigma, \gamma, \zeta_{ext}, P_{latent}, a, b, T^i)$. The length of burn-in for the MCMC simulations was checked using multiple MCMC chains with dispersed starting points, perturbing a subset of the parameters at each step in the chain (further descriptions in the supplementary information, and Figures S4 and S7 for the evolution of the MCMC chains and the posterior distribution of the parameters). Prior distributions were chosen to be uniform, with ranges $\beta = [1 \times 10^{-5}, 0.01]$, $\sigma = [\text{6 hours, 4 months}]$, $\gamma = [4 \text{ months, 1 year}]$, $\zeta_{ext} = [1 \times 10^{-5}, 2 \times 10^{-2}]$, $P_{latent} = [1 \times 10^{-5}, 2 \times 10^{-2}]$, $a = [-0.1, 0.1]$, $b = [0.2, 0.8]$ and $T^i = [50 \text{ days, 1.5 years}]$, based on the parameters estimates previously reported in various field and experimental studies.

A running sum of the probability of a transmission event ($\beta P_i^{inf}(t) P_j^{sus}(t)$) was retained, ignoring potential transmissions that were incompatible with the phylogenetic tree (for example between animals with isolates in different clades of the tree, excluding these animals from being linked by transmission). We used a modified Dijkstra’s algorithm [46] to identify all possible infection paths between cattle for which we have isolates. Each link “i” in a defined chain has a probability $p_i$ of being associated with transmission, so that the probability of that transmission chain occurring is given by the product of the $p_i$’s. Then the probability that a path will exist between two individuals “A” and “B” is simply the probability that at least one of the identified possible chains will connect the two, and therefore

$$P_{AB} = 1 - \prod_{\text{all chains}} \left(1 - \prod_{\text{its chain}} p_i\right).$$

The calculation included all chains that do not pass through another sequenced isolate, and for computational convenience, we have limited the calculation to chains of less than three different intermediate cattle. In order to test our assumptions, we ran forward simulations of a SEI model, based on the most likely parameter values of the posterior distributions from the outbreak data analysis. Applying our estimation approach to the simulated outbreak data confirmed that all the input parameter values fell within the 95% credible intervals of their corresponding posterior distributions derived from the original analysis of the data (not shown).

**Identifying the best-fit network distance between herds**

Similar to the approach described above, all paths amongst Herds 1 to 5 that do not pass through the remaining herds were calculated. The link strength probability between Herds C and D was taken to be

$$P_{CD} = 1 - \prod_{\text{all chains}} \left(1 - (1-p_{animal})^{n_{link}}\right),$$

where the product considers any direct path (only between Herds 3 and 5), all chains of path length 2 (i.e. with one intermediate), or where these do not exist, chains of path length 3. The parameter $n_{link}$ is the number of cattle moved between pairs of premises in the chains.

The ‘network separation’, defined as $[1-P_{CD}]$, was used as a measure of within-herd network distance, and this was plotted against the minimum genetic distance between the two herds. The value of $P_{animal}$ (i.e. the relative weighting assigned to each animal moved, where $0 < P_{animal} < 1$) was set to optimise the R² value (Figure 3B).

**Supporting Information**

Figure S1 Maximum likelihood tree of 31 *M. bovis* genomes based on 38 concatenated SNPs. Tree search was conducted in phylm under a Juke-Cantor model and reference sequence AF2122/97 was used as an outgroup (removed for clarity) to establish the root node.

(EPS)

Figure S2 Accumulation of genetic changes in *M. bovis* genomes through time. Genetic divergence of 31 cattle and badger isolates from the root node (measured in substitutions per genome) increases with sampling date, consistent with clock-like evolution. The estimated slope provides an evolutionary rate of 3.40 (CI: 0.87–5.93) x 10⁻¹ substitutions per site per year. To improve visibility, points were jittered randomly.

(EPS)

Figure S3 Network of contact via cattle movements amongst all cattle herds in Northern Ireland where breakdowns due to VNTR type 10 have been identified. Herds from which sequenced isolates were derived are indicated in red and number as in Figure 1. All other herds in yellow.

(TIF)

Figure S4 Evolution of the Gelman-Rubin shrink factor for the Markov Chain Monte Carlo chains (burn-in period removed). Each chain (6 in total) was started at a different point in the parameter space of the model. At each step in the chain we perturbed the set of parameters to make the next step different from the previous (i.e. the relative weighting assigned to each animal moved, where $0 < P_{animal} < 1$) was set to optimise the R² value (Figure 3B).

(TIF)

Figure S5 Posterior kernel density estimates. Illustrated are the distributions obtained from the MCMC chains after the burn-in period was removed. Convergence towards optimum values in all parameters is observed, with considerable mixing of the parameters. Here $\beta, \sigma, \gamma$ are the transition rates from the susceptible to exposed, exposed to test sensitive and test sensitive to infectious states respectively; $\zeta_{ext}, P_{latent}$ are the external and internal [latently infected animals within the herd] reservoir terms respectively, $T^f$ is the length of the infectious stage and $a$ and $b$ inform the probability $p_i$ that a reactor animal was infectious (rather than test sensitive) at the time of a positive test, according to the form $p_i = at + b$. Mean posterior values with 95% credible intervals ($\beta = 3.52 \pm [2.27, 4.77] \times 10^{-3}$ fortnight⁻¹, $\sigma = 0.387 \pm [0.392, 0.392]$ fortnight⁻¹, $\gamma = 0.266 \pm [0.262, 0.270]$ fortnight⁻¹, $\zeta_{ext} = 0.633 \pm [0.631, 0.635]$ fortnight⁻¹, $P_{latent} = 1.010 \pm [1.000, 1.430] \times 10^{-4}$ fortnight⁻¹, $a = 0$, $b = 0.0860 \pm [0.0837, 0.0884]$, $T^f = 71.6 \pm [68.9, 74.3]$ days. See Figure S6 for the distribution of all the points sampled.

(PS)

Figure S6 Distribution of the samples taken in the Markov Chain. The lower panel shows all the sampled points of the Markov Chain and the upper is colour coded with the lighter colours denoting those samples corresponding to a higher
likelihood. The clumping that is observed in the sampling regime for some parameters is due to the convergence of the chains. Here $p_L$, $p_T$, and $p_{QI}$ are the transition rates from the susceptible to exposed, exposed to test sensitive and test sensitive to infectious states respectively; $a_{ext}$ and $a_{inter}$ are the external and internal (latent infected animals) reservoir terms respectively, $T_L$ is the length of the infectious stage $a$ and $b$ inform the probability $p_{SI}$ that a reactor animal was infectious (rather than test sensitive) at the time of a positive test, according to the form $p_{SI} = at + b$. The priors used can be seen from the limits of the sampled points, in all cases we used uniform priors over these limits. The parameter $L$ is the maximum length of the infectious period (in days). The priors used for $\sigma$, $\gamma$, and $\Omega$ correspond to stages with lengths 1–110 days and 120–280 days respectively, and for $\Omega$, 0.50–0.85.

Figure S7 Trace of the parameters of the model. Illustrated are the traces of the parameters used in the model. Convergence towards posterior values in all parameters is observed from the dispersed starting points. Here $p_L$, $p_T$, and $p_{QI}$ are the transition rates from the susceptible to exposed, exposed to test sensitive and test sensitive to infectious states respectively, $a_{ext}$ and $a_{inter}$ are the external and internal (latent infected animals) reservoir terms respectively, $T_L$ is the length of the infectious period (in days), $\alpha$ is the sensitivity of the routine herd test applied to each animal.

References


Author Contributions

Conceived and designed the experiments: RB AO TM SM RAS RRK. Performed the experiments: SM CM. Analyzed the data: RB AO DW RJ H T RRK. Wrote the paper: RB AOH RRK.

Acknowledgments

The authors gratefully acknowledge access to anonymised animal-level data on APHIS, and for the isolates obtained from RTA survey badgers, both from DARD. Thanks are given to Profs. N.H. Smith and D.T. Haydon for helpful comments on the manuscript.

Table S1 Supporting Material and Methods: Examining the reliability of SNP calls among the 31 VNTR-10 isolates.

(DOC)


