Genomic epidemiology of *Mycobacterium bovis* infection in sympatric badger and cattle populations in Northern Ireland.

Assel Akhmetova[1], Jimena Guerrero[2], Paul McAdam[3], Liliana C.M. Salvador[4], Joseph Crispell[5], John Lavery[6], Eleanor Presho[6], Rowland R. Kao[7], Roman Biek[1], Fraser Menzies[8], Nigel Trimble[8], Roland Harwood[8], P. Theo Pepler[1], Katarina Oravcova[1], Jordon Graham[6], Robin Skuce[6], Louis du Plessis [9], Suzan Thompson[6], Lorraine Wright[6], Andrew Byrne[10], Adrian R. Allen[6].

1 - University of Glasgow, Glasgow, UK.

2 - Centro de Investigacion en Alimentacion y Desarrollo A.C., Hermosillo, Sonora, Mexico

3 – Fios Genomics, Edinburgh, UK.

4 – Department of Infectious Diseases, College of Veterinary Medicine, University of Georgia, Athens, GA, USA.

5 – Foreign, Commonwealth and Development Office, Glasgow, UK.

6 - Agri-Food and Biosciences Institute, AFBI Stormont, Belfast, UK.

7 – University of Edinburgh, Roslin Institute, Edinburgh, UK

8 – Department of Agriculture, Environment and Rural Affairs (DAERA), Belfast, UK.

9 – Department of Zoology, University of Oxford, Oxford, UK

10 – Department of Agriculture Food and the Marine (DAFM), Dublin, Ireland.

Abstract

Background

Bovine tuberculosis (bTB) is a costly, epidemiologically complex, multi-host, endemic disease. Pathogen whole genome sequencing can improve the resolution of epidemiological tracing. We genome sequenced an exceptional data set of 619 *Mycobacterium bovis* isolates from badgers and cattle in a 100km^2 bTB 'hotspot'. Historical molecular subtyping data permitted the targeting of an endemic pathogen lineage, whose long-term persistence provided an opportunity to study genome epidemiology in detail. To assess whether badger population genetic structure was associated with the spatial distribution of pathogen genetic diversity, we microsatellite genotyped hair samples from 769 badgers trapped in this area.

Results

Eight lineages of *M. bovis* were circulating in the study area, seven of which were likely non-endemic, and imported by animal movement. The endemic lineage exhibited low genetic diversity with an average inter-isolate genetic distance of 7.6 SNPs (s.d. ± 4.0), consistent with contemporary transmission. Bayesian phylogenetic methods determined an evolutionary rate of 0.30 substitutions per genome per year for this lineage, estimating its emergence 40-50 years before present, while Bayesian Skyline analysis identified significant population expansion of the endemic lineage in the 1990s and again in 2011-2012. The phylogeny revealed distinct sub-lineages, all of which contained isolates from both cattle and badger hosts, indicative of the sharing of closely related strains and inter-species transmission. However, the presence of significant badger population genetic structure was not associated with the spatial distribution of *M. bovis* genetic diversity.

Conclusions

Our data provided unparalleled detail on the evolutionary history of an endemic *M. bovis* lineage. Findings are consistent with ongoing interspecies transmission in the study area but suggest that badger intra-species transmission may not be a major driver of persistence in this area. In addition, the data collected permitted the tracking of incursions of novel pathogen lineages into the study area and means to determine if they were involved in disease transmission.

1. Introduction.

1.1 Background

Mycobacterium bovis infection in cattle (*Bos taurus*) and badgers (*Meles meles*) is a persistent and costly problem for the farming industries and governments of the United Kingdom and Republic of Ireland (Allen et al. 2018). In Northern Ireland alone, the bovine tuberculosis (bTB) eradication scheme cost £44 million in 2017/2018 (Northern Ireland Audit Office, 2018). The complex epidemiology of the disease is well recognised, with the role of wildlife in transmitting infection to cattle acknowledged as an impediment to eradication (Godfray et al. 2018).

Multi-host zoonotic infections of slowly-evolving pathogens, such as the members of the *M. tuberculosis* complex (MTBC), present significant challenges to researchers who wish to use molecular epidemiological methods to understand disease transmission dynamics (Biek et al. 2015). Previously, multi-locus variable number of tandem repeats analysis (MLVA) and spoligotyping were used to characterise spatio-temporal patterns in bTB epidemiology (Kamerbeek et al. 1997; Skuce et al. 2010; Milne et al. 2019; Skuce et al. 2020). These methods have demonstrated how *M. bovis* infections typically present as a series of geographically localised micro-epidemics (Skuce et al. 2010; Skuce et al. 2020). However, MLVA and spoligotype loci, whilst extremely useful in defining the home ranges of endemic infections (Trewby, 2016a; Milne et al. 2018), evolve at rates considerably slower than the inter-host transmission rate, thereby limiting their utility for contemporary disease outbreak investigations (Meehan et al. 2018).

Whole Genome Sequencing (WGS) technologies and associated phylogenetic analytical frameworks, have helped to reveal sources of infection and to improve surveillance and control for various pathogens (Harris et al. 2013; Mellman et al. 2011; Walker et al. 2013). These phylodynamic methods have been most effectively applied to fast-evolving viral pathogens, whose mutation rates can, with dense sampling, permit inference of fine scale disease dynamics, over short time intervals (Volz et al. 2009; Biek et al. 2015).

Biek et al (2012) were the first to apply WGS and Bayesian phylogenetics to the *M. bovis* epi-system, focusing on an emerging endemic strain found in the east of Northern Ireland. Since then, a variety of phylodynamic studies using ancestral state reconstruction and structured coalescent methods, have been applied to the *M.* bovis epi-system and have helped to characterise transmission dynamics involving livestock and wildlife (Trewby et al. 2016b; Crispell et al. 2017; Salvador et al. 2019; Crispell et al. 2019, Crispell et al. 2020; Rossi et al. 2020).

Pathogen spread across landscapes is recognised to be an inherently spatial process (Biek and Real, 2010), leading to distinct patterns in pathogen genetic structure. Similarly, free ranging wildlife hosts, exhibit partitioning of their own genetic variation across landscapes (Guerrero et al. 2018), for example, isolation by distance (IBD) (Wright, 1943). An appreciation of how these types of landscape-genetic phenomena intersect can help to inform epidemiological investigations in wildlife populations (Biek and Real, 2010). A key question within localised bTB micro-epidemics is whether significant badger population structure is observed over smaller geographic scales in Ireland, and if so, whether it explains any of the partitioning of *M. bovis* genetic variation in the landscape. If badgers are playing a key role in maintaining disease in a landscape by sustaining infection among themselves and passing infection on to cattle, then one may expect that their philopatric population structure will affect the population structure of *M. bovis* – more closely related badgers, should harbour more genetically similar sequence types of *M. bovis*.

1.2 Objectives

In this study we sought to better understand the molecular epidemiology of this important veterinary pathogen using the highest resolution WGS data available, in a region of high pathogen prevalence. Compared to some previous studies, we undertook a systematic, sympatric sampling of both cattle and badgers. The study was conducted in a small 100 km^2 area of Northern Ireland (Figure 1A), using data from a recently completed (2014-2018) wildlife intervention. The 'test and vaccinate or remove' (TVR) selective culling protocol applied to the local badger population (Arnold et al. 2021; Menzies et al. 2021) provided an important opportunity to systematically sample sympatric cattle and badger populations for *M. bovis*, and to apply WGS, phylogenetic and population genetic methods to attempt to determine the roles that both cattle and badgers play in the local disease dynamics.

Specifically, we sought to define the evolutionary history of an endemic pathogen lineage, its genetic diversity and the relatedness of isolates derived from cattle and badgers to broadly inform on transmission dynamics. Within this endemic lineage, we also sought to determine the effect of badger population structure on the spatial partitioning of pathogen diversity. More generally, we also sought to assess the utility of WGS in detecting the incursion of non-endemic lineages and whether they established persistent foci of infection.

2. Methods.

2.1 Sampling of cattle and badgers.

The TVR zone (Figure 1A), was chosen because it has a high prevalence of cattle TB (24% confirmed bTB herd prevalence over two years [2011/2012] – wider Northern Ireland average was ~6-7%), and is embedded within County Down, which also has one of the highest average badger densities (3.88 badgers per km^2) in Northern Ireland – (Reid et al. 2012).

An initial survey was conducted in 2014 to establish sett locations (DAERA, 2018). Overnight cage trapping was used across all years between the months of May and October 2014-2018. Trapped badgers were anaesthetised, trap-side, hair sampled and tested for *M. bovis* infection using the dual path platform (DPP) serology test based on the StatPak method (Chambers et al. 2009). Tracheal aspirate was taken from all trapped badgers, whether DPP-positive or -negative and sent for bacteriological culture as described below. From 2015-2018, DPP positive badgers were humanely euthanized and DPP-negative badgers were vaccinated using injectable Bacillus Calmette Guerin (BCG), and released (Menzies et al. 2021). Culled badgers underwent post-mortem examination according to a standardised protocol (Courcier et al. 2020) and specified tissues were submitted for bacteriological culture.

All cattle in Northern Ireland are TB tested annually using the standardised (European Council, 1964) single intradermal comparable cervical tuberculin (SICCT) test (Abernethy et al. 2006). Specified tissue from all SICCT positive bTB reactor cattle in the TVR region from 2014-2017 were harvested at the time of slaughter and submitted for bacteriological culture.

To provide extra temporal depth, a selection of historical isolates (n=243) from both hosts in the TVR zone were re-cultured from the AFBI *M. bovis* Northern Ireland wide, strain archive. The temporal window for all isolates runs from 1986 to 2017. All historical badger isolates were derived from road

157 traffic accident (RTA) post-mortems in a surveillance scheme run by the Department of Agriculture,
158 Environment and Rural Affairs in Northern Ireland (DAERA-NI) (Courcier et al. 2018). In addition, an
159 area directly neighbouring the study zone has a distinct lineage of *M. bovis* present, four isolates of
160 which were detected in the study zone. A random sample of historical isolates from this
161 neighbouring area was collected to provide additional phylogenetic context.

162 **2.2 *M. bovis* culture and genomic DNA extraction.**

163 *M. bovis* isolates were initially cultured in the liquid BD BACTEC MGIT system, solid Stonebrinks and
164 Löwenstein-Jensen media and single colonies were selected for sub-culture (Skuce et al. 2005).
165 Isolates were heat-killed in a water bath at 80°C for 30 min. DNA was extracted using standard high
166 salt/cationic detergent cetyl hexadeycl trimethyl ammonium bromide (CTAB) and solvent extraction
167 protocols (Parish and Stoker, 2001; van Soolingen et al. 2002).

168 **2.3 Spoligotyping and MLVA analysis.**

169 *M. bovis* isolates were genotyped by spoligotyping (Kamerbeek et al., 1997) and 8 locus MLVA using
170 previously described methods (Skuce et al., 2010). Authoritative names for spoligotype patterns
171 were obtained from www.mbovis.org (Smith and Upton, 2012). MLVA profiles were named using a
172 laboratory nomenclature (Skuce et al. 2005).

173 **2.4 Genome sequencing and bioinformatic analyses.**

174 Sequencing libraries were prepared using the Illumina Nextera XT method to produce inserts of
175 approximately 500-600bp. One hundred samples were sequenced at AFBI using an Illumina MiSeq
176 platform with Illumina V2 chemistry, producing paired-end reads of 250 bp. A further 100 samples
177 were sequenced at the Glasgow Polyomics facility using an Illumina Miseq producing 2×300 bp
178 paired end reads. All remaining samples were sequenced by Eurofins Scientific using an Illumina
179 HiSeq producing 2 x250 bp paired end reads. For quality assurance (QA) purposes, comparison of
180 sequencing performance across all three sites was undertaken at AFBI. Specifically, we re-
181 sequenced 15 randomly selected isolates from those sent to other institutes and compared these
182 duplicates to the initial sequencing data.

183 Reads for each sample were mapped to the recently updated/annotated (Malone et al. 2017)
184 reference genome for *M. bovis* strain AF2122/97 (GenBank accession LT708304.1) using the
185 mapping-based phylogenomics pipeline, RedDog V1beta.10.3 (Edwards et al. 2016) to identify SNP
186 variants across all isolates. Alignment and mapping were carried out in Bowtie2 v2.2.9 (Langmead et
187 al. 2012) using the sensitive local mapping setting. SNP calling was undertaken using SAMtools and
188 BCFtools (Li et al. 2009) using the consensus caller setting. A minimum depth of 10x was set for SNP
189 calling. The average coverage failure filter, average depth filter and average mapping failure filters
190 were set at 98%, 10x, and 75% respectively. Transposable elements, repeat regions and the PE/PPE
191 regions as defined in the Genbank annotation, were excluded from SNP calling using the parseSNP
192 tool in the RedDog pipeline (Edwards et al. 2016).

193 Some isolates were sequenced by Glasgow Polyomics, others were sequenced by Eurofins
194 laboratories. 15 randomly selected isolates were re-sequenced by AFBI. No SNP distance was
195 observed between duplicates and the previously sequenced isolates.

196 **2.5 Badger microsatellite genotyping.**

197 Nuclear DNA extracted from all badger hair samples was genotyped at 14 microsatellite loci using
198 methods previously described (Guerrero et al. 2018). We re-profiled 5% of DNA samples as a quality
199 assurance measure.

200 **2.6 Testing for IBD in badger population.**

201 Using the package 'PopGenReport' (Adamack and Gruber, 2014) in R (R Development Core Team,
202 2020), we constructed a microsatellite distance matrix (Smouse and Peakall method) for all unique
203 badgers captured in the study zone, and for the subset of badgers that produced *M. bovis* cultures
204 from the endemic lineage (see below), to ensure we had sufficient power to detect IBD in sub-
205 populations. For both datasets, we then constructed inter-animal Euclidean distance matrices using
206 the R package, 'Geosphere' (Hijmans et al. 2019). We then performed a Mantel test, with 10,000
207 repetitions, for IBD using the package 'ade4' (Dray and Dufour, 2007) in R. For the larger meta-
208 population, Mantel tests and linear regressions were carried out for each capture year, with an
209 analysis of covariance carried out to compare the slopes of each genetic distance vs euclidean
210 distance relationship.

211 **2.7 Phylogenetic analyses.**

212 The most appropriate nucleotide substitution model for our phylogenetic analyses of the FASTA
213 alignments of informative SNPs was assessed using the 'modelTest' function of the package
214 'Phangorn' (Schliep, 2011) in R. Specifically, the fit of the General Time Reversible (GTR), Jukes
215 Cantor, and Hasegawa Kishino Yano (HKY) models to the data were assessed. The nucleotide
216 substitution model with the lowest AIC (GTR – AIC: 26860.39) was used to build a maximum
217 likelihood phylogeny using RAxML v8 (Stamatkis, 2014). The rapid bootstrapping search method in
218 RAxML was selected and stopped after 400 replicates. The phylogeny was visualized and assessed in
219 FigTree v1.4.4 (Rambaut, 2018) and the ape package (Paradis and Schliep, 2018) in R. Final figures of
220 maximum likelihood trees were produced using ggtree (Yu, 2020). The presence/absence of a
221 temporal signal in the phylogenetic data of the endemic clade was assessed using the program
222 TempEst v1.5.1 (Rambaut *et al.* 2016). After selecting the reference strain (AF2122/97) as the best-
223 fitting, outgroup root isolate, which maximises the temporal signal, the root to tip divergence model
224 was fitted using the residual mean squared method.

225 To test the significance of the temporal signal in our dataset, we randomised the tip dates for the
226 302 chosen isolates from the endemic clade, in ten replicate analyses, as *per* Firth et al (2010). Tip
227 dates were randomised using the 'Tipdatingbeast' package (Rieux and Khatchikian, 2017) in R. The
228 original dataset and each replicate were subjected to a simplified BEAST 2 (Bayesian Evolutionary
229 Analysis by Sampling Trees - Bouckaert et al. 2014), constant population, coalescent analysis, using a
230 relaxed clock model, the GTR nucleotide substitution model. A chain length of 100,000,000 MCMC
231 steps was used for the three non-randomised tip date runs, with a longer 300,000,000 MCMC steps
232 used for all ten randomised runs. 10% of MCMC steps were discarded as burn-in. After checking
233 convergence in Tracer 1.7.1 (Rambaut et al. 2018) – (ESS > 200 for all parameters), the median
234 substitution rate for each replicate was compared to those of the logcombined non-randomised data
235 set.

236 To determine the past demographics of the endemic *M. bovis* lineage in the study region, we also
237 performed a Bayesian Skyline analysis in BEAST 2 (Drummond et al. 2005). As above, we used a
238 relaxed lognormal clock and the GTR nucleotide substitution model. We set the model to assess
239 historical effective population size (N_e) over 4 dimensions. A chain length of 200,000,000 MCMC

240 steps was used for three replicates. 10% of steps were discarded as burn in and replicate log and
241 tree files were then combined in logcombiner and analysed in Tracer 1.7.1.

242 Also within the endemic *M. bovis* clade, we ran a Mantel test to determine if closely related
243 sequence types exhibited any geographic localization / clustering within the study area. We
244 constructed an inter-isolate SNP distance matrix using the 'Biostrings' package (Pagés et al. 2021),
245 and an inter isolate Euclidean distance matrix in 'Geosphere' (Hijmans et al. 2019) in R. The Mantel
246 test was run for 10,000 iterations in the 'ade4' package in R as described above.

247

248 **2.8 Assessing effect of badger population structure on *M. bovis* spatial partitioning.**

249 We sought to determine if *M. bovis* inter-isolate SNP distance is associated with pairwise Euclidean
250 distance between trapped badger locations, pairwise difference in time of *M. bovis* isolation and
251 pairwise host genetic distance.For *M. bovis* genome sequences derived from badgers infected with
252 the endemic lineage, we constructed two inter-isolate distance matrices: (i) SNP distance using the R
253 package 'ape' (Paradis and Schliep, 2019); (ii) time of isolation difference using the 'dist' function in
254 R.

255 Using these *M. bovis* distance matrices and the two already produced to assess IBD in the infected
256 badgers (see above), we performed a multiple regression on distance matrices (MRM) analysis using
257 the R package 'ecodist' (Goslee and Urban, 2007) with 10000 repetitions.

258 ## 3. Results

259 **3.1 Sampling of cattle and badgers.**

260 A total of 642 *M. bovis* isolates were used in this study, of which 611 were collected from badgers
261 and cattle in the TVR zone, and 31 from a neighbouring region. Of the 642 isolates, 15 were
262 sequencing QA duplicate controls as discussed in section 2.4. In addition, we re-sequenced the *M.*
263 *bovis* reference sample, AF2122/97 as an internal control. Of the 642 survey isolates, 399 (282 cattle;
264 117 badger) were sampled contemporaneously in the TVR zone during the project (2014-2017). A
265 further 242 historical isolates (232 cattle; 10 badger=10) from 1986-2013 were available from
266 archived cultures from the zone and its environs. Cattle isolates across all years were single isolates
267 per animal from 185 herds. Multiple isolates (n=86) were cultured from 24 individual badgers, with
268 single isolates derived from a further 36 badgers. In total, between 1986 and 2017, we collected *M.*
269 *bovis* isolates from 60 unique badgers and 483 unique cattle (Figure 1B). Full details of sample
270 locations, year of isolation and species of origin are given in Supplementary Data 1.

271 **3.2 Spoligotyping and MLVA analysis.**

272 22 MLVA types and 6 spoligotypes were observed for the 642 isolates. From prior analyses (Skuce et
273 al. 2010), the spoligotype and MLVA genotypes could be grouped into eight related 'strain families'.
274 Each is dominated by a probable founder genotype (source of the family name) with related
275 daughter strains varying by spoligotype and / or MLVA polymorphism. Numbers of isolates per strain
276 family are shown in Table 1A. The MLVA 6, spoligotype 263 family (6.263) was considered to be
277 endemic in the region and accounted for most of the observed isolates (Skuce et al. 2010; Skuce et
278 al. 2020). The remaining seven strain families (1.140, 2.142, 3.140, 4.140, 5.140, 19.140 and 20.131)
279 were not likely to be endemic in the TVR cull zone as each has a home range elsewhere in Northern
280 Ireland (Skuce et al. 2010; Trewby, 2016a). Of the 36 isolates from strain family 20.131, 32 were

281 collected from a region neighbouring the TVR zone, 4 badger isolates were found within the zone.
282 Full details of isolate MLVA genotypes and spoligotypes are supplied in Supplementary Data 1.

283 **3.3 Sequencing, bioinformatic analyses and Quality Assurance.**

284 The RedDog pipeline was used to process the isolates. 24 (22 cattle and 2 badgers) failed the
285 sequencing QA filters (98% coverage filter for the reference genome) and were excluded. The
286 remaining 618 survey isolates plus AF2122/97 control passed all QA filters. Detailed QA meta-data
287 for all 619 isolates are given in Supplementary Data 2. Forward and reverse reads for all QA passing
288 isolates are deposited at the National Centre for Biotechnology Information (NCBI) Sequence Read
289 Archive (SRA), bio-project XXXXXXXX. Accession numbers for all reads are given in Supplementary
290 Data 2.

291 From the 619 isolates with good quality sequence reads, 1562 SNPs passed QA calling rules and were
292 used to conduct phylogenetic analyses. Details of all SNPs passing QA, and their location in the
293 reference sequence are given in Supplementary Data 3.

294 The AF2122/97 control exhibited a 3 SNP distance from the reference sequence (Genbank
295 LT708304.1) likely due to accrual of mutations after culture passages at AFBI.

296 **3.4 Badger microsatellite genotyping and IBD analyses.**

297 769 unique badgers were captured, location recorded, sampled and successfully genotyped between
298 2014 and 2018. Random 5% re-genotyping for QA purposes produced identical allele calls.
299 Microsatellite profiles, capture locations and date of capture can be found in Supplementary Data 4.
300 Summary population genetic statistics for all animals, per year are collated in Supplementary Table
301 S1.

302 Samples from 45 badgers produced positive *M. bovis* cultures. Spoligotyping and MLVA placed them
303 in the major endemic lineage in the study zone. Capture locations for the 45 endemic strain positive
304 badgers are illustrated in Supplementary Figure S1A.

305 Across all years, the badger population exhibited consistent levels of IBD, as indicated by significant
306 Mantel tests (r=0.11-0.17, $p<0.05$) (Table 2). The slopes of the relationships between genetic
307 distance and Euclidean distance were very similar. Small, significant differences were observed in
308 the ANCOVA (Supplementary Figure S2 and Table 2), mainly due to the increased slope of the IBD
309 relationships years 2015, 2016, and 2018, consistent with badger genetic differentiation being
310 observed over shorter distances.

311 The 45 *M. bovis* culture positive badgers also exhibited significant IBD (Table 2) similar to that of the
312 larger study population.

313 **3.5 Preliminary phylogenetic analyses.**

314 **3.5.1. All isolates.**

315 The Maximum Likelihood tree constructed in RAxML for all 619 sequenced isolates is shown in Figure
316 2. The phylogeny was rooted using the 20.131 strain family as an out-group, as these isolates are
317 known to derive from an older common ancestor than other extant strains (Allen et al. 2013). Eight
318 major lineages, each with high bootstrap support, were observed in the phylogeny. The eight strain
319 families defined previously by MLVA and spoligotyping were monophyletic and in perfect
320 concordance with the SNP based tree topology. We determined the within-lineage diversity, as
321 defined by total number of SNPs recorded, for each of the eight major lineages observed (Table 1B).

322 Additionally, from distance matrices generated during phylogenetic analyses, pairwise, inter-isolate
323 SNP distance, summary statistics were computed within all eight major lineages (Table 1C).

324 **3.5.2 Endemic lineage – 6.263.**

325 The endemic major lineage of 6.263 presented the best opportunity to investigate *M. bovis*
326 transmission dynamics among cattle and badgers in the TVR zone. 6.263 has been consistently
327 associated with the study area for over two decades in local cattle and badgers, unlike lineages
328 whose home range is elsewhere in Northern Ireland (Skuce et al. 2020). A higher resolution SNP
329 phylogeny of 6.263 is shown in Figure 3, using the subset of 302 isolates described above. Cattle and
330 badger isolates were observed in all sub-lineages of the endemic lineage, with no sub-lineages made
331 up of isolates exclusively from a single host species. Major sub-lineages had good bootstrap support
332 (>90). A bar-plot showing frequency of isolates per host species taken from the endemic clade over
333 the period 1986-2017, is presented in Supplementary Figure S3. A smaller maximum likelihood
334 phylogeny of the 45 endemic clade badgers is presented in Supplementary Figure S1B.

335 Badger isolates were predominantly found in the endemic 6.263 lineage. No badger isolates were
336 found in the 1.140, 2.142, 3.140, 5.140 and 19.140 lineages. Four isolates, from two unique badgers,
337 sampled in the TVR zone, were found in the 20.131 lineage, whilst five isolates from four unique
338 badgers were found in the 4.140 lineage (Supplementary Figures S4 and S5). A full breakdown of
339 number of isolates from each species in each lineage is shown in Table 1A.

340 The n=302 endemic lineage phylogeny analysed in TempEst, rooted against the AF2122/97 reference
341 genome, exhibited a positive correlation between genetic divergence (root to tip distance) and
342 sampling time, with moderate evidence of temporal signal (R^2=0.25; p<0.001), and a conservative
343 clock rate of approximately 0.22 substitutions per genome, per year - see Supplementary Figure S6A.
344 All ten tip date randomised replicate datasets run using the simple coalescent model, exhibited
345 similar substitution rates, all of which were considerably lower than the substitution rate inferred on
346 the non-randomised data set (0.30 substitutions per genome, per year 95% HPD: 0.24-0.37) and
347 exhibited no overlap in 95% highest posterior density (HPD) (Supplementary figure S6B).

348 The Bayesian Skyline analysis indicated a substitution rate that was very similar to that observed in
349 the non-randomised simple coalescent model described above (0.35 substitutions per genome, per
350 year 95% HPD:0.28-0.41- Supplementary Figure S6B). It also indicated that after the origin of the
351 6.263 clade in the 1970s-1980s, the effective population size substantially expanded in the early to
352 mid 1990s, and again in 2011-2012 prior to the TVR trial beginning (Figure 4).

353 The Mantel test conducted for the endemic clade indicated that there was no significant spatial
354 autocorrelation for sequence types (correlation coefficient = 0.04, p=0.157) – consistent with there
355 being no significant clustering of closely related sequence types within the study area – see
356 Supplementary Figure S7. An illustration of the spatial distribution of sub-clades within the endemic
357 6.263 lineage, and their substantial overlap, is provided in Supplementary Figure S8.

358 **3.6 Effect of badger population structure on *M. bovis* spatial partitioning.**

359 From the full model in the MRM analysis, modeling *M. bovis* genetic distance (SNP-based) as a
360 function of inter badger genetic distance, inter badger Euclidean distance and inter *M. bovis* time of
361 isolation difference, we observed that only inter-badger Euclidean distance was significantly
362 associated with *M. bovis* genetic distance (p = 0.04). However, the overall fit of the model was non-
363 significant (F-test p >0.05). Badger microsatellite derived genetic relatedness was therefore not
364 associated with *M. bovis* SNP-derived genetic differentiation. A full summary of MRM findings is
365 presented in Supplementary Table S2.

366

367 # 4. Discussion

368 **4.1 General observations on WGS for enhanced molecular epidemiology.**

369 Our data are consistent with the findings that WGS provides unparalleled resolution for
370 epidemiological investigations of zoonotic disease (Kao et al. 2014; Kao et al. 2016) – indexing
371 additional pathogen variation to stratify isolates that are homogeneous according to the classical
372 tuberculosis molecular epidemiological tools of spoligotyping and MLVA (Skuce et al. 2010; Skuce et
373 al. 2020).

374 Regarding the latter tools, we find perfect congruence between spoligotype and MLVA data and the
375 basal nodes defining major lineages of the phylogeny for all isolates from the TVR region (Figure 2).
376 This concordance is testament to the clonality of *M. bovis* (Smith et al. 2006; Allen et al. 2013) and is
377 indicative that existing databases of classical molecular markers, can be used to target 'hotspots' of
378 persistent, endemic infection for closer investigation using WGS.

379 Contemporary transmission of slowly evolving pathogens, such as members of the *Mycobacterium*
380 *tuberculosis complex* (MTBC), is typically characterised by little within clade diversity and resulting
381 reduced SNP distances / phylogenetic branch lengths between isolates (Meehan et al. 2018).
382 Previous studies have suggested various minimal SNP distance thresholds, for the definition of
383 contemporaneous, epidemiologically linked isolates. Five and twelve SNP distances have been
384 proposed to be consistent with such transmission clusters in *M. tuberculosis* outbreaks in the United
385 Kingdom (Walker et al. 2013; Walker et al. 2014), whilst ten SNP thresholds have been proposed in
386 other studies with *M. tuberculosis* and *M. bovis* (Bryant et al. 2013; Roetzer et al. 2013; Yang et al.
387 2017; Jajou et al., 2018; Crispell et al. 2019). Meehan et al (2018) have observed that in *M.*
388 *tuberculosis*, SNP distances between one and five can represent transmission events up to 10 years
389 apart. Given that MTBC evolution appears to consistently involve relaxed molecular clock like
390 behaviour across lineages (Crispell et al. 2017; Crispell et al. 2019; Menardo et al. 2019), it is perhaps
391 not surprising that selecting a definitive threshold for contemporary transmission is difficult, and as a
392 result, those set can appear arbitrary. In addition to the issues with relative clock like behaviour of
393 different *M. bovis* lineages, intensity of sample collection and representation of multiple time points
394 can be crucial to establishing robust substitution rates (Kao et al. 2016). The general rule of thumb
395 remains however - the shorter the SNP distance between isolates, the more likely they are more
396 closely epidemiologically linked. In this study, the lineage we know to be endemic in the study area
397 from years of MLVA surveillance, exhibits the shortest average pairwise SNP distance between
398 isolates (7.6 s.d. ±4.0 – see Table 1), which is likely indicative of contemporary transmission in the
399 region, and compares favourably to the thresholds discussed above.

400 The advantage that WGS will provide in disease tracing compared to historical molecular
401 epidemiology methods in the *M. bovis* epi-system, is that outbreaks can be traced back to higher
402 resolution, WGS defined lineages and sequence types, found in more precise locations than those
403 defined by genetically homogeneous, MLVA home ranges that can cover substantial geographical
404 areas (Skuce et al. 2010; Skuce et al 2020). A recent and pertinent example of this, is an outbreak
405 from Cumbria in northwest England, which genome sequencing revealed was linked to the outbreak
406 area under study here (Rossi et al. 2020).

407

408 **4.2 Endemic 6.263 lineage**

409 Our Bayesian phylogenetic analysis suggests that the endemic 6.263 lineage has been present in the
410 TVR zone since the 1960s-1980s. The rate of molecular evolution of 0.30-0.35 substitutions per
411 genome per year is consistent with previous *M. bovis* phylodynamic studies – see Supplementary
412 Table S3. The presence of bacterial isolates from both cattle and badgers throughout all sub-lineages
413 of the endemic 6.263 lineage, and their close genetic relatedness, indicates likely bi-directional
414 transmission between both species. The Bayesian Skyline analysis of the endemic clade was
415 indicative that the population size of the pathogen expanded in the early to mid 1990s, and again
416 substantially prior to the beginning of the sett survey and TVR protocol described here from 2014-
417 2017 – a finding congruent with the increase in herd prevalence observed in Northern Ireland over
418 the same time-period (Robinson, 2015), and more particularly with the higher than average 24%
419 herd prevalence observed in the study zone from 2011/2012. The magnitude of the demographic
420 expansion observed violates the assumptions of constant population size that structured coalescent
421 models require to accurately infer transition rates between different states (DeMaio et al. 2015;
422 Müller et al. 2018), therefore it was not possible to undertake robust analyses with this dataset.
423 Exploratory analyses using different structured coalescent methods (data not shown) yielded
424 contradictory outcomes dependent on assumptions made about pathogen effective population sizes
425 within hosts. In cases wherein substantial changes to population size have occurred, structured
426 coalescent methods can then attempt to explain variation in terms of population structure, resulting
427 in some within state effective population sizes to be artificially low, and some transition rates to be
428 artificially high (du Plessis et al 2020).

429 More generally, the clonal expansion of *M. bovis* in this period, after initial success of the test and
430 slaughter protocol of the 1950s and 1960s in other regions across the UK has been noted before,
431 and suggested to potentially result from the adaptive spread of a favourable mutation or invasion of
432 a new territory (Smith et al. 2003; Smith et al. 2006). Future work should seek to determine if similar
433 demographic expansions are observed in other lineages and to search for candidate functional
434 polymorphisms and signatures of selection that could underpin any potential pathogen phenotype
435 of interest to control schemes and the development of diagnostic and vaccine tools.

436 The lack of spatial clustering of closely related sequence types from the entire endemic clade is an
437 intriguing finding given that a general feature of *M. bovis* molecular epidemiology to date in the
438 Northern Irish landscape has been association of molecular types with specific regions (Skuce et al.
439 2020). It may be however that while the genomic data is sufficiently high resolution to detect such
440 an association, the location data for cattle may not be so useful. The data used are geolocations of
441 the major farm holdings and may not necessarily be representative of the locations at which cattle
442 have been exposed to *M. bovis,* owing to factors of farm fragmentation and the common practice of
443 land rental distant from farm buildings (concacre), which is a feature of Northern Irish agriculture
444 (Allen et al. 2015). This deficiency in the recording of cattle data is in contrast to the endemic *M.*
445 *bovis* positive badger data which have an exact capture location, and which exhibited some evidence
446 of spatial clustering as per the MRM analysis described above. Future use of WGS data for precision
447 molecular epidemiology should aim to use better quality geolocation data for cattle derived isolates.
448 Weighted centroids of pasture and farm buildings may be a more useful metric.

449 Badger genetic population structure, whilst remaining stable over the intervention period, had no
450 association with how *M. bovis* genetic diversity was spatially distributed. It is possible that this lack
451 of association is due to factors other than reduced badger to badger intraspecies transmission
452 dynamics. The endemic lineage, if it were a relatively recent incursion, may have had little time to
453 establish foci of persistent infection in badgers, and diffuse across the landscape through philopatric
454 contact networks. This lineage has, however, been present in the region for ~33-48 years, providing

455 ample time for establishment to occur. Alternatively, perturbation of the badger population, and
456 associated dispersal arising through the application of culling, even at a small-scale (Bielby et al.
457 2014), may have served to obscure any association between pathogen and host population
458 structures. The relative stability of the IBD relationship we observe, and the fact that during the
459 culling period of 2015-2018, genetic differentiation occurs over shorter distances than the survey-
460 only year of 2014, are consistent with no perturbation signal. GPS collar data from badgers from this
461 same region have also suggested badger ranging did not increase significantly in cull years compared
462 to survey (O'Hagan et al. 2021). Given these observations, we postulate that our badger genetic
463 data may be indicative that intra-species badger transmission of *M. bovis* may not be a major driver
464 of endemic disease persistence in the study area.

465

466 **4.3 Monitoring of non-endemic lineages**

467 Sequence data are useful for identifying probable incursions of non-endemic disease lineages into
468 new areas, and for potentially determining if the incursion results in contemporary transmission and
469 persistence. Intra-lineage, inter-isolate SNP distances greater than that observed for the endemic
470 6.263 lineage, are possibly indicative of a lack of contemporary transmission, likely associated with
471 lineages which are non-endemic in the study region as per previous observations from Northern
472 Ireland wide molecular epidemiological surveillance (Skuce et al. 2010; Skuce et al. 2020). This
473 certainly appears to hold true for lineages 1.140, 2.142 and 3.140, which exhibit average pairwise
474 inter-isolate distances of between 17.6 and 21.6 SNPs. Lineages 5.140 and 19.140 are only
475 represented by two isolates each, but for both lineages, inter-isolate distances are again observed to
476 be larger than the endemic lineage at 79 SNPs and 13 SNPs respectively. Given this observation, and
477 the fact that we know these five lineages are outside of their MLVA defined home ranges (Skuce et
478 al. 2020), it seems probable that they could have arrived in the study area through multiple, long
479 distance, cattle movements. It is noteworthy, but anecdotal, that these lineages are comprised
480 solely of isolates from cattle, suggestive perhaps that the lack of contemporary transmission for
481 these incursive strains has resulted in no infection reaching the wildlife population. However, with
482 deficiencies in sampling, badger 'trappability' and TB test diagnostics as discussed in the
483 supplementary materials, one cannot be definitive that a non-endemic, visiting pathogen lineage has
484 not established a focus of infection in the study area. Only continued surveillance over a wider
485 temporal window could assess that. The establishment of long-term genome-based surveillance
486 systems could in the future help to inform on successful incursions (Gardy and Loman, 2018).

487 The 20.131 lineage exhibits mean inter-isolate SNP distances which are comparable to that of the
488 endemic 6.263 lineage (Table 1). However, only four isolates of this lineage, from two badgers (See
489 Supplementary Figure S4), were found within the study area, with the majority sampled from a
490 neighbouring region in which this strain has a focus of infection. The short inter-isolate distances
491 observed are therefore more likely to be consistent with contemporary transmission in the
492 neighbouring region. It is noteworthy however that the two badgers sampled for this lineage and
493 found in the study zone were found in isolation, with no associated, contemporary, study zone cattle
494 isolates. It could be that badgers have dispersed from the neighbouring region into the study zone,
495 carrying infection with them, which has yet to appear in the cattle population. However, it is also
496 possible that an undetected reservoir of cattle may have entered the study zone and transmitted
497 infection to local badgers. Alternatively, an undetected reservoir not picked up by sampling could be
498 residing within the zone. With so few isolates of the 20.131 lineage from within the study area, and
499 the previously mentioned biases in sampling, it is impossible to be definitive. Again, detailed,

longitudinal, genome facilitated surveillance would perhaps be able to inform more fully on this incursive lineage in this region.

Interestingly, one of the historically non-endemic lineages (4.140) (Skuce et al. 2020), does appear to be persisting in the TVR zone, with multiple badger-sourced isolates observed to exhibit shorter inter-isolate SNP distances (0-1) between each-other and cattle from the area (Supplementary Figure S5), consistent with more contemporary transmission events. The latter observation does highlight the usefulness of this WGS based approach for on-going surveillance, and for detecting incursions that establish new foci in new regions. It seems most probable from this case, that new foci of non-endemic lineages are introduced by cattle movements, with subsequent spill-over to badgers.

Conclusions

Our data serve to further illustrate the utility of WGS for bovine tuberculosis molecular epidemiology, indicating ongoing transmission of infection in an endemically affected region. Our novel use of badger population genetic data alongside pathogen genomic data has also helped to demonstrate that in this particular region, this philopatric species' stable population structure is not associated with *M. bovis* population structure, which may be indicative that badgers are not playing a major role in disease persistence. The latter may not be the case in other regions however and will require empirical assessment.

Our findings also illustrate the importance of assessing the past and recent demography of endemic lineages. Significant expansion and contraction of pathogen population size can limit the utilisation of phylodynamic tools to more closely inform on inter host transition dynamics.

More generally, the high-resolution data presented herein, serve to shrink the home range over which disease tracing investigations need to take place in comparison to MLVA based methods, thereby improving precision and inferences on likely infection sources. Developing national databases for other common, endemic lineages associated with other locations will substantially modernise field epidemiology and control schemes, enabling detection of incursions of novel sequence types into regions that are not their traditional home range (Rossi et al. 2020), and determining if such incursions have resulted in ongoing transmission between cattle and wildlife, thereby establishing novel foci of infection. A key requirement to make best use of data in the ways described will be improving the precision with which cattle geo-location data are collected.

Finally, in addition to the molecular epidemiological information that a national database of WGS data would provide, the sequence data collected would also be a useful resource for the study of evolutionary questions pertaining to possible adaptive phenotypes of *M. bovis* that may be influencing the success of the eradication scheme across Britain and Ireland.

Acknowledgments

Assel Akhmetova is supported by a Bolashak International Scholarship. This work was funded by the Department of Agriculture, Environment and Rural Affairs for Northern Ireland (DAERA-NI) through its 'Evidence and Innovation' programme – project no. 15/3/07. Additional funding was provided by the UK's Biotechnology and Biological Sciences Research Council (BBSRC) – grant numbers BB/P0105598 and BB/M01262X. The funders had no role in study design, analysis or decision to

publish. The authors wish to recognise Shane Collins and Carl McCormick and their team for supervising all field work and AFBI Disease Surveillance and Bacteriology teams who provided excellent post-mortem examinations and mycobacteriology support. Thanks to Dr Dez Delahay of the Animal and Plant Health Agency (APHA) in GB for advice on surveying local badger populations and trapping.

Research Ethics

All badger field work was carried out under licences issued by the Northern Ireland Environment Agency. All scientific procedures performed on badgers were conducted according to the guidelines of the Animals Scientific Procedures Act (ASPA - Licence 2767) overseen by the Department of Health for Northern Ireland.

Conflict of interest

The authors declare no conflict of interest in the production of this work.

Supplementary Files

R scripts and BEAST xml run files used in the performance of this work are curated at the following GitHub repository:

https://github.com/AdrianAllen1977/Genome-epidemiology-of-Mycobacterium-bovis-infection-in-contemporaneous-sympatric-badger-and-cattle

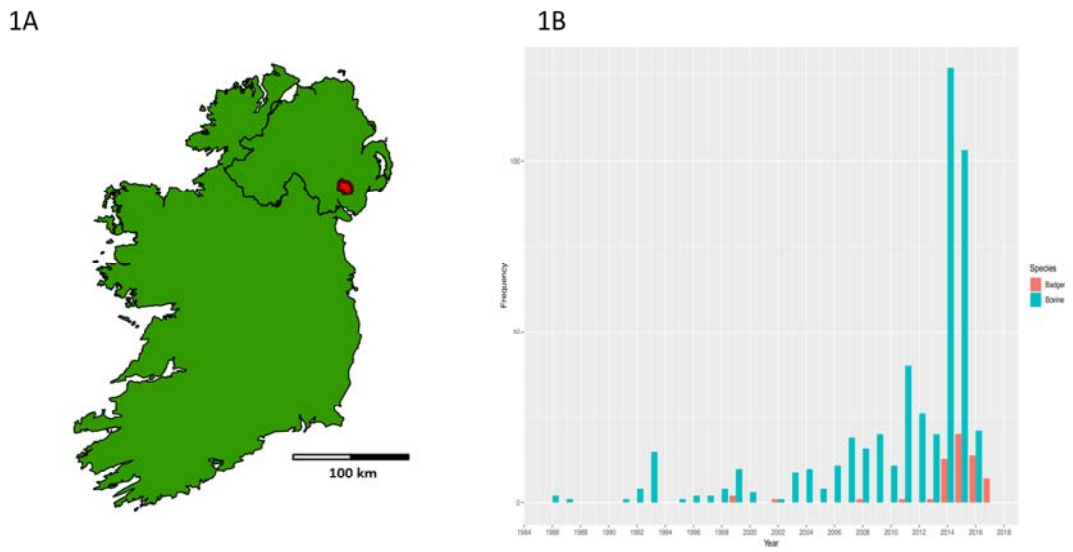

Figure 1 – A: Map of Ireland with location of 100 km^2 TVR cull zone highlighted in red. B: Frequency bar-plot of *Mycobacterium bovis* isolates sampled annually, from unique badgers and cattle, in the TVR zone prior to (1986-2013) and during the intervention (2014-2017).

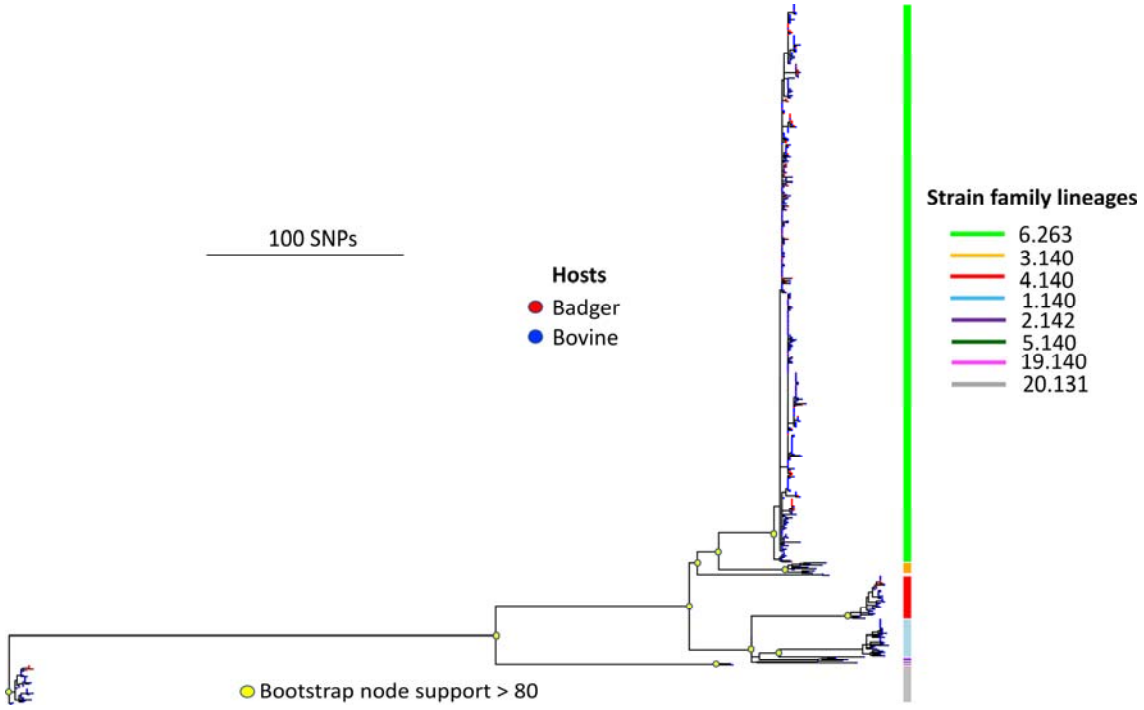

Figure 2 – 1562 SNP Maximum Likelihood phylogeny of all 619 isolates that passed sequencing QA.

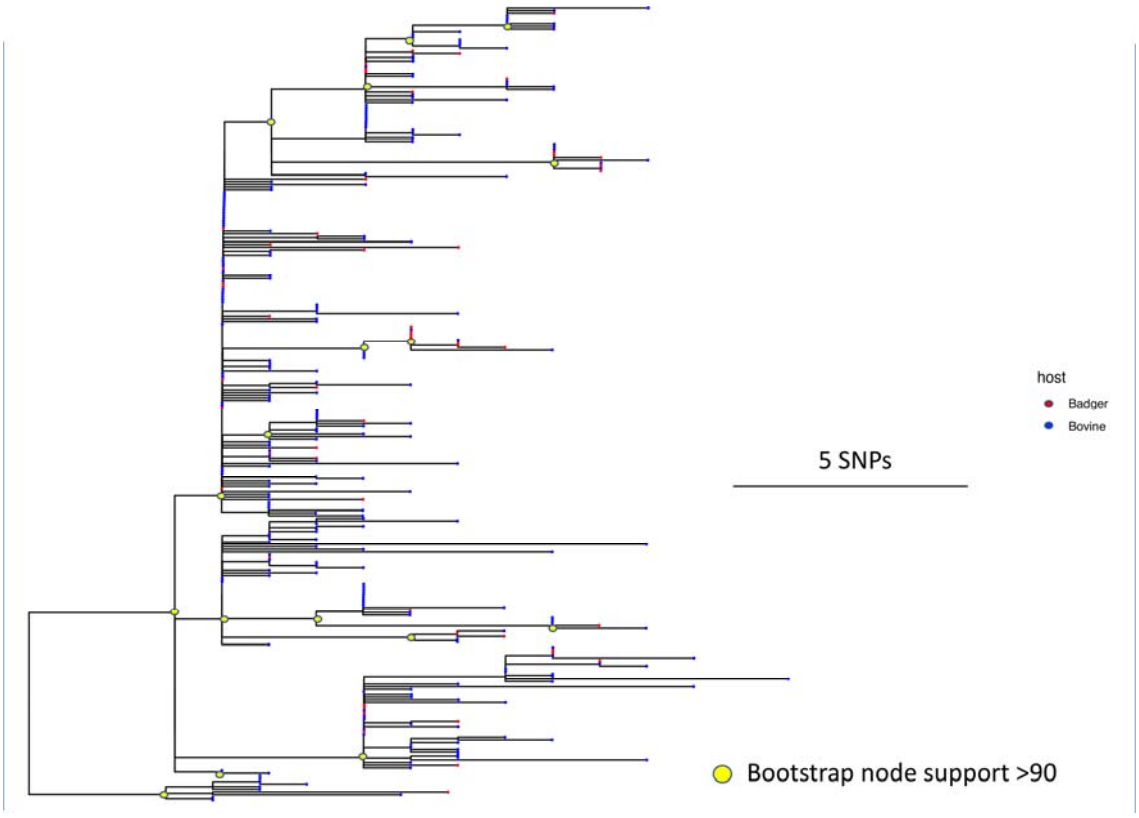

Figure 3 – 377 SNP Maximum Likelihood phylogeny of 302 isolates from the 6.263 endemic lineage in the TVR zone. Tree rooted to AF2122/97 reference, but reference removed for aid of visualisation.

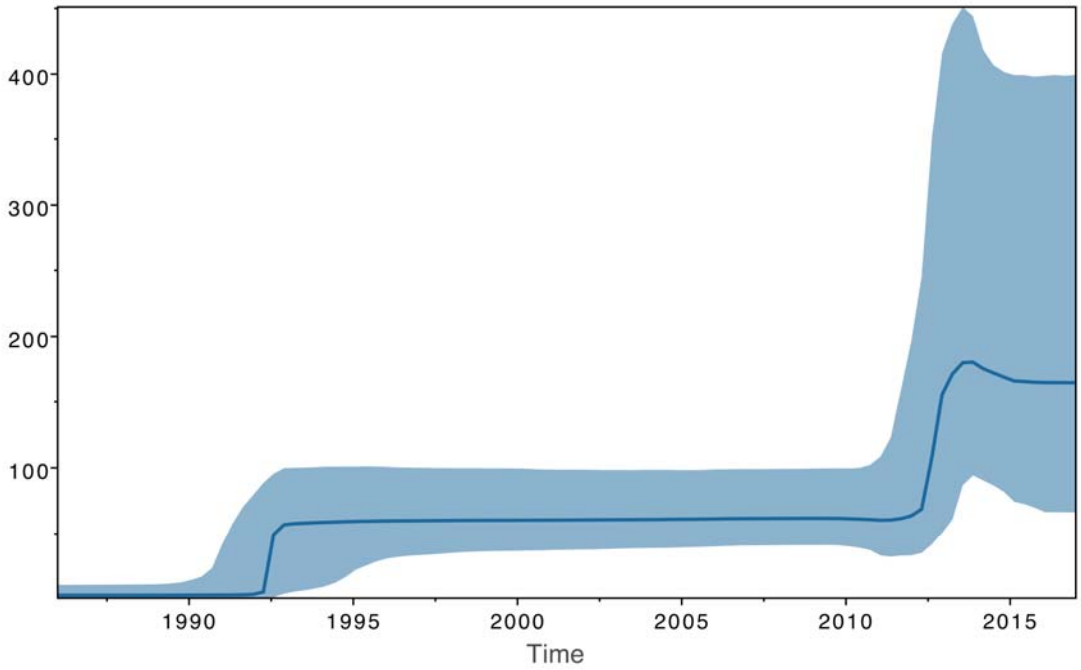

Figure 4 – Bayesian Skyline plot of the 6.263 endemic lineage.

	MLVA Strain families	1.140	2.142	3.140	4.140	5.140	6.263	19.140	20.131
A	No. of isolates	33	4	10	38	2	478	2	36
	No. cattle isolates	33	4	10	33	2	367	2	32
	No. badger isolates	0	0	0	5	0	111	0	4
B	No. SNPs in clade	186	38	84	92	82	377	13	53
C	Min	0.0	8.0	8.0	0.0	N/A	0.0	N/A	0.0
	1st Quartile	10.0	2.0	16.0	8.0	N/A	5.0	N/A	6.0
	Median	15.0	20.5	21.0	12.0	N/A	7.0	N/A	8.0
	Mean	17.6	19.5	21.6	11.7	N/A	7.6	N/A	8.0
	3rd Quartile	19.0	21.0	27.0	15.0	N/A	10.0	N/A	11.0
	Max	96.0	27.0	34.0	28.0	79.0	25.0	13.0	18.0
	St. Dev	20.0	6.2	7.3	6.3	N/A	4.0	N/A	3.8

Table 1 – A. number of isolates per strain family sampled in the study area with breakdown of number of isolates per host species. B. number of SNPs detected in strain family clades, and C. pairwise SNP distance statistics for each of the eight major lineages of *M. bovis* found in the TVR zone.

Cohort genotyped	No. animals	Mantel test Pearson coefficient r	Mantel p value	Linear model beta	Linear model p value	R^2	1 unit differentiation per x km distance
2014	273	0.11	0.0001	$1.6x10^{-4}$	$<2x10^{-16}$	0.012	6.25 km
2015	152	0.16	0.0001	$2.1x10^{-4}$	$<2x10^{-16*}$	0.024	4.74 km
2016	97	0.17	0.0001	$2.6x10^{-4}$	$<2x10^{-16**}$	0.030	3.84 km
2017	113	0.11	0.0001	$1.5x10^{-4}$	$<2x10^{-16}$	0.011	6.67 km
2018	134	0.13	0.0004	$1.7x10^{-4}$	$<2x10^{-16***}$	0.017	5.88 km
TB +ve	45	0.16	0.0024	$3.1x10^{-4}$	$1.1x10^{-7}$	0.030	3.22 km

Table 2 – Badger meta population isolation by distance (IBD) relationship for all sampling years. * =2015 significantly different than slopes for years 2014, 2016 and 2017. ** = 2016 slope significantly different than slopes for years 2014 and 2015. ***=2018 slope significantly different than slopes for years 2014 and 2017.

	Substitution rate	Time to most recent common ancestor (tMRCA) - yrs
Simple Bayesian coalescent	0.30 (0.24-0.37)	48.87 (33.88-82.56)
Bayesian Skyline	0.35 (0.28-0.41)	33.22 (31.00-37.81)

Table 3 – Summary statistics for log-combined simple Bayesian coalescent and Bayesian Skyline models. All estimates are median and 95% highest posterior density (HPD). Substitution rate is expressed as sites per genome, per year. Time to most recent common ancestor (tMRCA) for the endemic 6.263 clade is expressed in years prior to the last sampling date, 2017.

716

717

718

719

720 ## References

721 Abernethy DA, Denny GO, Menzies FD, McGuckian P, Honhold N, Roberts AR (2006) The Northern
722 Ireland programme for the control and eradication of *Mycobacterium bovis*. Vet Microbiol 112(2-4):
723 231-237.

724 Adamack AT, Gruber B (2014) PopGenReport : simplifying basic population genetic analyses in R.
725 Methods Ecol Evol 5(4): 384-387.

726 Allen AR, Dale J, McCormick C, Mallon TR, Costello E, Gordon SV et al (2013) The phylogeny and
727 population structure of *Mycobacterium bovis* in the British Isles. Infect Genet Evol 20: 8-15.

728 Allen, A, Breadon E, Byrne A, Mallon T, Skuce R, Groussaud P, Dainty A, Graham J, Jones K, Pollock L,
729 Whatmore A (2015) Molecular epidemiology of *Brucella abortus* in Northern Ireland – 1991-2012.
730 PLoS One 10(9): e0136721.

731 Allen AR, Skuce RA, Byrne AW (2018) Bovine Tuberculosis in Britain and Ireland - A Perfect Storm?
732 the Confluence of Potential Ecological and Epidemiological Impediments to Controlling a Chronic
733 Infectious Disease. Front Vet Sci 5: 109.

734 Arnold ME, Courcier EA, Stringer LA, McCormick CM, Pascual-Linaza AV, Collins SF et al (2021) A
735 Bayesian analysis of a Test and Vaccinate or Remove study to control bovine tuberculosis in badgers
736 (*Meles meles*). PLoS One 16(1): e0246141.

737 Biek R, O'Hare A, Wright D, Mallon T, McCormick C, Orton RJ et al (2012) Whole Genome Sequencing
738 Reveals Local Transmission Patterns of *Mycobacterium bovis* in Sympatric Cattle and Badger
739 Populations. PLoS Pathog 8(11): e1003008.

740 Biek R, Pybus OG, Lloyd-Smith JO, Didelot X (2015) Measurably evolving pathogens in the genomic
741 era. Trends Ecol Evol 30(6): 306-313.

742 Biek R, Real LA (2010) The landscape genetics of infectious disease emergence and spread. Mol Ecol
743 19(17): 3515-3531.

744 Bielby J, Donnelly CA, Pope LC, Burke T, Woodroffe R (2014) Badger responses to small-scale culling
745 may compromise targeted control of bovine tuberculosis. Proc Nat Acad Sci USA 111(25): 9193-9198.

746 Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D et al (2014) BEAST 2: a software platform
747 for Bayesian evolutionary analysis. PLoS Comp Biol 10(4): e1003537.

748 Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V et al (2013) Inferring
749 patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing
750 data. BMC Infect Dis 13:110.

751 Chambers MA, Waterhouse S, Lyashchenko K, Delahay R, Sayers R, Hewinson RG (2009)
752 Performance of TB immunodiagnostic tests in Eurasian badgers (*Meles meles*) of different ages and
753 the influence of duration of infection on serological sensitivity. BMC Vet Res 5: 42.

754 Courcier EA, Menzies FD, Strain SAJ, Skuce RA, Robinson PA, Patterson IAP et al (2018) Monitoring
755 *Mycobacterium bovis* in Eurasian badgers (*Meles meles*) killed by vehicles in Northern Ireland
756 between 1998 and 2011. Vet Rec 182(9): 259.

757 Courcier EA, Pascual-Linaza AV, Arnold ME, McCormick CM, Corbett DM, O'Hagan MJH et al (2020)
758 Evaluating the application of the dual path platform VetTB test for badgers (*Meles meles*) in the test
759 and vaccinate or remove (TVR) wildlife research intervention project in Northern Ireland. Res Vet Sci
760 130: 170-178.

761 Crispell J, Benton CH, Balaz D, De Maio N, Ahkmetova A, Allen A et al (2019) Combining genomics
762 and epidemiology to analyse bi-directional transmission of *Mycobacterium bovis* in a multi-host
763 system. Elife 8.

764 Crispell J, Cassidy S, Kenny K, McGrath G, Warde S, Cameron H et al (2020) *Mycobacterium bovis*
765 genomics reveals transmission of infection between cattle and deer in Ireland. Microb Genom 6(8).

766 Crispell J, Zadoks RN, Harris SR, Paterson B, Collins DM, de-Lisle GW et al (2017) Using whole genome
767 sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand.
768 BMC Genomics 18(1): 180.

769 DAERA. (2018) Test and vaccinate or remove (TVR) wildlife intervention research.
770 https://www.daera-ni.gov.uk/articles/test-and-vaccinate-or-remove-tvr-wildlife-intervention-
771 research

772 De Maio N, Wu CH, O'Reilly KM, Wilson D (2015) New Routes to Phylogeography: A Bayesian
773 Structured Coalescent Approximation. PLoS Genet 11(8): e1005421.

774 Dray S, Dufour A (2007) The ade4 Package: Implementing the Duality Diagram for Ecologists. J Stat
775 Software 22: 1-20.

776 Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G. (2005) Bayesian coalescent inference of past
777 population dynamics from molecular sequences. Mol Biol Evol 22(5): 1185-1192.

778 du Plessis L, Vaughan TG, Müller N (2020) MASCOT v2.1.2 Taming the BEAST Tutorial.
779 https://taming-the-beast.org/tutorials/Mascot-Tutorial/

780 Edwards D, Pope B, Holt K (2016) RedDog mapping based phylogenomics pipeline.
781 https://github.com/katholt/RedDog

782 European Council. (1964) "Council Directive 64/432/EEC". Off J Eur Comm.

783 Firth C, Kitchen A, Shapiro B, Suchard MA, Holmes EC, Rambaut A (2010) Using time-structured data
784 to estimate evolutionary rates of double-stranded DNA viruses. Mol Biol Evol 27(9): 2038-2051.

785 Gardy JL, Loman NJ (2018) Towards a genomics-informed, real-time, global pathogen surveillance
786 system. Nat. Rev. Genet. 19, 9-20.

787 Godfray C, Donnelly C, Hewinson RG, Winter M, Wood J (2018) Bovine TB strategy review. DEFRA
788 https://www.gov.uk/government/publications/a-strategy-for-achieving-bovine-tuberculosis-free-
789 status-for-england-2018-review

790 Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. J
791 Stat Software 22(7): 1-19.

792 Guerrero J, Byrne AW, Lavery J, Presho E, Kelly G, Courcier EA et al (2018) The population and
793 landscape genetics of the European badger (*Meles meles*) in Ireland. Ecol Evol 8(20): 10233-10246.

794 Harris SR, Cartwright EJ, Torok ME, Holden MT, Brown NM, Ogilvy-Stuart AL et al (2013) Whole-
795 genome sequencing for analysis of an outbreak of meticillin-resistant Staphylococcus aureus: a
796 descriptive study. Lancet Infect Dis 13(2): 130-136.

797 Hijmans RJ, Williams, E., Vennes C. (2019) Geosphere: Spherical trigonometry for geographic
798 applications. https://cran.r-project.org/web/packages/geosphere/index.html

799 Jajou R, de Neeling A, van Hunen R, de Vries G, Schimmel H, Mulder A et al (2018) Epidemiological
800 links between tuberculosis cases identified twice as efficiently by whole genome sequencing than
801 conventional molecular typing: a population-based study. PLOS One 13:e0195413.
802
803 Kamerbeek J, Schouls L, Kolk A, van Agterveld M, van Soolingen D, Kuijper S et al (1997)
804 Simultaneous detection and strain differentiation of Mycobacterium tuberculosis for diagnosis and
805 epidemiology. J Clin Microbiol 35(4): 907-914.

806 Kao RR, Haydon DT, Lycett SJ, Murcia PR (2014) Supersize me: how whole genome sequencing and
807 big data are transforming epidemiology. Trends in Microbiology 22 (5): 282-291.
808
809 Kao RR, Price-Carter M, Robbe-Austerman S (2016) Use of genomics to track bovine tuberculosis
810 transmission. Revue Sci Tech 35(1): 241-258.

811 Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9(4):
812 357-359.

813 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The Sequence Alignment/Map
814 format and SAMtools. Bioinformatics 25(16): 2078-2079.

815 Malone KM, Farrell D, Stuber TP, Schubert OT, Aebersold R, Robbe-Austerman S et al (2017)
816 Updated Reference Genome Sequence and Annotation of *Mycobacterium bovis* AF2122/97. Genome
817 Announc 5(14).

818 Meehan CJ, Moris P, Kohl TA, Pečerska J, Akter S, Merker M et al (2018) The relationship between
819 transmission time and clustering methods in Mycobacterium tuberculosis epidemiology.
820 EBioMedicine 37: 410-416.

821 Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A et al (2011) Prospective
822 Genomic Characterization of the German Enterohemorrhagic Escherichia coli O104:H4 Outbreak by
823 Rapid Next Generation Sequencing Technology. PLoS ONE 6(7): e22751.

824 Menardo F, Duchêne S, Brites D, Gagneux S (2019) The molecular clock of Mycobacterium
825 tuberculosis. PLoS Pathog 15(9): e1008067.

826 Menzies FD, McCormick, C.M., O'Hagan, M.J.H., Collins, S.F., McEwan, J., McGeown, C.F.et al (2021)
827 Test and vaccinate or remove: methodology and preliminary results from a badger intervention
828 research project. Vet Rec - Accepted.

829 Milne GM, Graham J, Allen A, Lahuerta-Marin A, McCormick C, Presho E et al (2019) Spatiotemporal
830 analysis of prolonged and recurrent bovine tuberculosis breakdowns in Northern Irish cattle herds
831 reveals a new infection hotspot. Spat Spatiotemporal Epidemiol 28: 33-42.

832 Milne MG, Graham J, Allen A, McCormick C, Presho E, Skuce R et al (2019) Variation in
833 *Mycobacterium bovis* genetic richness suggests that inwards cattle movements are a more
834 important source of infection in beef herds than in dairy herds. BMC Microbiol 19(1): 154.

835 Müller NF, Rasmussen D, Stadler T (2018) MASCOT: parameter and state inference under the
836 marginal structured coalescent approximation. Bioinformatics 34(22): 3843-3848.

837 Northern Ireland Audit Office. (2018) Bovine Tuberculosis Report. http://www.niauditoffice.gov.uk

838 O'Hagan M, Gordon, A., McCormick, C., Collins, S., Trimble, N., McGeown, C. et al (2021) Effect of
839 selective removal of badgers (*Meles meles*) on ranging behaviour during a 'Test and Vaccinate or
840 Remove' intervention in Northern Ireland. Epidemiol Infect 149: e125.

841 Pagés H, Aboyoun P, Gentleman R, DebRoy, S (2021) Biostrings: Efficient manipulation of biological
842 strings. https://bioconductor.org/packages/Biostrings

843 Paradis E, Schliep K (2019) ape 5.0: an environment for modern phylogenetics and evolutionary
844 analyses in R. Bioinformatics 35(3): 526-528.

845 Parish T, Stoker, N.G. (2001) Mycobacterium tuberculosis protocols. Humana Press.

846 R Development Core Team. (2020) R Foundation for Statistical Computing,Vienna, Austria.

847 Rambaut A (2018) Figtree v1.4.4. https://github.com/rambaut/figtree/releases

848 Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA (2018) Posterior Summarization in Bayesian
849 Phylogenetics Using Tracer 1.7. Syst Biol 67(5): 901-904.

850 Rambaut A, Lam TT, Max Carvalho L, Pybus OG (2016) Exploring the temporal structure of
851 heterochronous sequences using TempEst (formerly Path-O-Gen). Virus evol 2(1): vew007.

852 Reid N, Etherington TR, Wilson GJ, Montgomery WI, McDonald RA (2012) Monitoring and population
853 estimation of the European badger *Meles meles* in Northern Ireland. Wildlife Biol 18(1): 46-57.

854 Rieux A, Khatchikian CE (2017) tipdatingbeast: an r package to assist the implementation of
855 phylogenetic tip-dating tests using beast. Mol Ecol Res 17(4): 608-613.

856 Robinson PA (2015) A history of bovine tuberculosis eradication policy in Northern Ireland.
857 Epidemiol Infect 143: 3182-3195.

858 Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J et al (2013) Whole genome sequencing versus
859 traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal
860 molecular epidemiological study. PLOS Med 10:e1001387.
861
862 Rossi G, Crispell, J., Brough, T., Lycett, S.J., White, P.C.L., Allen, A. et al (2020) Phylodynamic analysis
863 of an emergent *Mycobacterium bovis* outbreak in an area with no previously known wildlife
864 infections. bioRxiv. 2020.11.12.379297; doi: https://doi.org/10.1101/2020.11.12.379297

865 Salvador LCM, O'Brien DJ, Cosgrove MK, Stuber TP, Schooley AM, Crispell J et al (2019) Disease
866 management at the wildlife-livestock interface: Using whole-genome sequencing to study the role of
867 elk in *Mycobacterium bovis* transmission in Michigan, USA. Mol Ecol 28(9): 2192-2205.

868 Schliep KP (2011) phangorn: phylogenetic analysis in R. Bioinformatics 27(4): 592-593.

869 Skuce R, Breadon E, Allen A, Milne G, McCormick C, Hughes C et al (2020) Longitudinal dynamics of
870 herd-level *Mycobacterium bovis* MLVA type surveillance in cattle in Northern Ireland 2003-2016.
871 Infect Genet Evol 79: 104131.

872 Skuce RA, Mallon TR, McCormick CM, McBride SH, Clarke G, Thompson A et al (2010)
873 *Mycobacterium bovis* genotypes in Northern Ireland: herd level surveillance (2003 to 2008). Vet Rec
874 167(18): 684-689.
875
876 Skuce RA, McDowell SW, Mallon TR, Luke B, Breadon EL, Lagan PL et al (2005).Discrimination of
877 isolates of *Mycobacterium bovis* in Northern Ireland on the basis of variable numbers of tandem
878 repeats (VNTRs). Vet Rec 157(17): 501-504.

879 Smith NH, Dale J, Inwald J, Palmer S, Gordon SV, Hewinson RG (2003) The population structure of
880 *Mycobacterium bovis* in Great Britain: Clonal Expansion. Proc Natl Acad Sci USA 100(25): 15271-
881 15275.

882 Smith NH, Gordon SG, de la Rua Domenech R, Clifton-Hadley RS, Hewinson RG (2006) Bottlenecks
883 and broomsticks: the molecular evolution of *Mycobacterium bovis*. Nat Rev Microbiol 4:670-682.

884 Smith NH, Upton P (2012) Naming spoligotype patterns for the RD9-deleted lineage of the
885 Mycobacterium tuberculosis complex; www.Mbovis.org. Infect Genet Evol 12(4): 873-876.

886 Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
887 phylogenies. Bioinformatics 30(9): 1312-1313.

888 Trewby H (2016a) The genetic and spatial epidemiology of bovine tuberculosis in the UK: from
889 molecular typing to bacterial whole genome sequencing. PhD thesis, University of Glasgow.

890 Trewby H, Wright D, Breadon EL, Lycett SJ, Mallon TR, McCormick C et al (2016b) Use of bacterial
891 whole-genome sequencing to investigate local persistence and spread in bovine tuberculosis.
892 Epidemics 14: 26-35.

893 van Soolingen D, de Haas PE, Kremer K (2001) Restriction fragment length polymorphism typing of
894 mycobacteria. Methods Mol Med 54: 165-203.

895 Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW (2009) Phylodynamics of
896 Infectious Disease Epidemics. Genetics 183(4): 1421-1430.

897 Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ et al (2013) Whole-genome
898 sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study.
899 Lancet Infect Dis 13(2): 137-146.

900 Walker TM, Lalor MK, Broda A, Saldana Ortega L, Morgan M, Parker L et al (2014) Assessment of
901 Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007–12, with whole pathogen
902 genome sequences: an observational study. Lancet Respir. Med. 2(4): 285-292.
903
904 Wright S (1943) Isolation by Distance. Genetics 28(2): 114-138.

905 Yang C, Luo T, Shen X, Wu J, Gan M, Xu P et al (2017) Transmission of multidrug-resistant
906 Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-
907 genome sequencing and epidemiological investigation. Lancet Infect. Dis. 17:275–284.
908

Yu G. (2020) ggtree: Elegant Graphics for Phylogenetic Tree Visualization and Annotation. https://guangchuangyu.github.io/ggtree-book/short-introduction-to-r.html