Point source introduction of *Mycobacterium bovis* at the wildlife-livestock interface can lead to clonal expansion of the disease in a single ecosystem

Supplemental Data

Detailed Methodology:

Sample selection

The isolates sequenced for the current study were collected previously, were characterised with traditional mycobacterial genotyping methods and were described in Michel et al. (2009) and Mostowy et al. (2005) (Michel et al., 2009; Mostowy et al., 2005). A representative selection of seventeen *M. bovis* isolates from various locations in South Africa from an in-house culture bank, with diverse IS*6110* RFLP fingerprints (as described by Michel et al. (2009)) were selected for whole genome sequence analysis.

Alignment and variant detection:

Quality assessment of the sequencing data (in FASTQ format) was done using FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc), followed by trimming of adapters and low-quality bases with a Phred quality score of less than 20 and filtering for a minimum read length of 36 using Trimmomatic. A minimum read length of 36 base pairs was used for subsequent mappingReads shorter than 15 bases were not used in consequent analysis steps. A multi-software approach was followed to align the high quality reads to the reference genome (*Mycobacterium tuberculosis* H37Rv, GenBank NC000962.2 and *M. bovis*, GenBank NC002945.3) and to detect variants. Reads were mapped to the reference genome with three mapping algorithms; Novoalign (Novocraft), Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009), and SMALT (Ponstingl and Ning, 2010). The Genome Analysis Tool Kit (GATK) was used as recommended in the user documentation to call single nucleotide

polymorphisms (SNPs) and small insertions and deletions (Indels) in all the alignment files from the different mapping algorithms used (McKenna et al., 2010). Specifically, RealignerTargetCreator, IndelRealigner, and UnifiedGenotyper were used according to the GATK best practices for bacterial genome analyses.

Phylogenomic analysis

Concatenated sequences containing high-confidence variable sites (coding and non-coding SNPs) with respect to the *M. tuberculosis* H37Rv reference genome, were written to a multi-fasta file (one entry for each isolate included) and were used as input in Modelgenerator to determine the optimal substitution model that fits the data structure (Felsenstein, 1989; Keane et al., 2006). The general time reversal (GTR) model scored the lowest in the hierarchical likelihood ratio test; Bayesian information criterion (BIC), and thus described the substitution pattern occurring in the dataset most accurately. The GTR model of substitution was subsequently applied to construct a maximum likelihood phylogeny of the isolates included in this analysis with Randomized Axelerated Maximum Likelihood (RaxML) with 1000 bootstrap pseudo-replicates (Stamatakis, 2006, 2014). Positions containing gaps or missing data were not considered for the analysis.