

1 **Who infects Whom? - Reconstructing infection chains of *Mycobacterium***
2 ***avium* ssp. *paratuberculosis* in an endemically infected dairy herd by use of**
3 **genomic data**

4 Annette Nigsch^{a,*}, Suelee Robbe-Austerman^b, Tod P. Stuber^b, Paulina D. Pavinski Bitar^c, Yrjö Gröhn^c
5 and Ynte H. Schukken^{a,d}

6 ^a Department of Animal Sciences, Wageningen University, Wageningen, The Netherlands

7 ^b USDA APHIS National Veterinary Services Laboratories, Ames, Iowa, United States of America

8 ^c Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine,
9 Cornell University, Ithaca, New York, United States of America

10 ^d GD Animal Health, Deventer, The Netherlands

11 * Corresponding author. E-mail: annette.nigsch@wur.nl (AN).

12

13 Abstract

14 Recent evidence of circulation of multiple strains within herds and mixed infections of cows marks the
15 beginning of a rethink of our knowledge on *Mycobacterium avium* ssp. *paratuberculosis* (MAP)
16 epidemiology. Strain typing opens new ways to investigate MAP transmission. This work presents a
17 method for reconstructing infection chains in a setting of endemic Johne's disease on a well-managed
18 dairy farm. By linking genomic data with demographic field data, strain-specific differences in
19 spreading patterns could be quantified for a densely sampled dairy herd. Mixed infections of dairy
20 cows with MAP are common, and some strains spread more successfully. Infected cows remain
21 susceptible for co-infections with other MAP genotypes. The model suggested that cows acquired
22 infection from 1–4 other cows and spread infection to 0–17 individuals. Reconstructed infection chains
23 supported the hypothesis that high shedding animals that started to shed at an early age and showed
24 a progressive infection pattern represented a greater risk for spreading MAP. Transmission of more
25 than one genotype between animals was recorded. In this farm with a good MAP control management
26 program, adult-to-adult contact was proposed as the most important transmission route to explain the
27 reconstructed networks. For each isolate, at least one more likely ancestor could be inferred. Our
28 study results help to capture underlying transmission processes and to understand the challenges of
29 tracing MAP spread within a herd. Only the combination of precise longitudinal field data and bacterial
30 strain type information made it possible to trace infection in such detail.

31

32 **Keywords**

33 *Mycobacterium avium* ssp. *paratuberculosis*, reconstruction of transmission trees, endemic, genomic
34 diversity, longitudinal data, dairy herd

35

36 Introduction

37 *Mycobacterium avium* ssp. *paratuberculosis* (MAP) is the causative agent of Johne's disease,
38 or paratuberculosis, a chronic, slowly progressing disease of ruminants associated with high
39 economic losses, especially in dairy herds. Challenges in the surveillance and control of MAP are a
40 long incubation period of 1–15 years (1), and inefficient diagnostic tests, which lead to limited success
41 of control programmes. The role of MAP in the pathogenesis of Crohn's disease in humans is still
42 controversial (2).

43 The primary route of MAP infection is faecal-oral by direct or indirect contact with the
44 pathogen. Calves are highly susceptible during the first weeks after birth, and resistance to infection
45 increases until one year of age (3). Calves become infected either horizontally or vertically (*in utero*).
46 Transmission from dams to calves at an early age is currently regarded as the most important route of
47 infection and is therefore the focus of many control programmes. In adults, ingestion of MAP does not
48 necessarily lead to infection, but repeated uptake of high doses of bacilli may result in adult infection
49 (4,5). Adult-to-adult, calf-to-calf and heifer-to-heifer infections have been shown to exist (4,6–9).
50 These routes typically receive little attention in MAP control programmes.

51 Large differences in MAP shedding patterns can be observed. Intermittent shedders, low
52 shedders (≤ 50 colony-forming units per gram (cfu/g) faecal matter), high shedders ($> 50–10^4$ cfu/g
53 faecal matter), and super-shedders ($> 10^4$ cfu/g of faecal matter) are known shedding categories for
54 individual animals. The majority of cows will never develop high shedding levels, since many cows
55 never reach advanced enough age (10). Schukken et al. found two distinct infection patterns, so
56 called progressors and non-progressors (8)(8). Progressors are characterised by continuous and
57 progressive shedding of high MAP loads and high antibody production. Non-progressors present
58 intermittent and low shedding of MAP bacteria and a virtual absence of a humoral immune response,
59 suggesting that they have the infection process under control. Building on these findings, Mitchell et
60 al. distinguished between two categories of progressors, linked to immune control and the age at
61 onset of shedding: cows that start shedding at a younger age partially control the infection, but
62 eventually become high shedders (slow progressive infection), while cows that start shedding

63 persistently at an older age progress rapidly with shedding and lack effective control of infection (10).
64 Obviously, super-shedders represent the greatest risk for spreading MAP among herd mates (11).
65 However, removing high-shedding animals (which are easily detected) has shown to be insufficient to
66 address long-term persistence of MAP (12,13). Simulation models have given further support to the
67 hypothesis that intermittent, low and transiently shedding animals play an important role in
68 maintaining low prevalent infections in dairy herds (14). Quantitative estimates of the importance of
69 transmission routes at all ages of the host and of the role of animals presenting these different
70 shedding patterns are essential to decide on relevant control procedures.

71 The MAP genome is extremely stable with an estimated mutation rate μ of the core genome
72 of one mutation per 2–7 years (15). Earlier literature assumed clonal infections of herds with a single
73 strain of MAP bacteria (16,17). However, several studies now have shown that multiple strains of
74 MAP may be simultaneously present in a herd (7,18), suggesting that several concurrent infection
75 cycles within a single population are possible. A more recent study even demonstrated the incidence
76 of a mixed, simultaneous infection by three genotypically diverse MAP isolates in a single dairy cow
77 (19). Within the individual host, the MAP population is initially thought to be genomically
78 homogeneous, but will diversify over time due to mutations. These processes of within-host evolution
79 of MAP and mixed genotype infection of hosts with multiple MAP strains need to be considered in
80 further studies to draw valid conclusions about the complexities of MAP transmission (20,21). For
81 such studies sequencing a single isolate from each case was suggested to be inadequate in the
82 presence of within-host diversity, but frequent sampling will improve accuracy (22). For MAP it is
83 currently not known whether the low mutation rate will allow detailed analyses of infection chains.
84 With whole genome sequencing (WGS) data the highest possible degree of discrimination between
85 pairs of isolates can be achieved. Nevertheless, integration of non-WGS data into analysis of
86 transmission pathways is suggested to lead to considerable refinement in our understanding of the
87 epidemiology of mycobacterial disease (20).

88 With falling costs of large-scale genome sequencing and advances of biostatistical tools,
89 population genomic studies are increasingly used to study pathogen spread within populations.
90 Traditionally, network inference models were used to identify transmission chains in early stages of
91 disease outbreaks. In endemic settings, network inference faces multiple challenges, such as: (a)

92 non-sampled early generations of cases and thus uncertainty about which of the sampled strains is
93 genomically closest to the originally introduced strain and can thus be considered as the most recent
94 common ancestor; (b) multiple introductions of genomically diverse strains over time, resulting in a
95 polyphyletic sample; and (c) as a consequence of (b), exposure of hosts to multiple strains which may
96 lead to mixed genotype infections.

97 This study aims to identify individual animal-to-animal infection chains (“who infects whom”),
98 in order to better understand the infection dynamics of MAP in endemically infected dairy herds.
99 Transmission trees will be constructed by using WGS data in combination with detailed longitudinal
100 epidemiological data. Support of the reconstructed infection chains for the current prevailing
101 hypotheses on transmission routes will be evaluated, and the role of individual animals in infection
102 spread will be investigated. In addition, within-host and within-herd diversity of MAP will be
103 characterised to provide fundamental input to all advanced analyses. To conclude, it shall be
104 discussed whether observational field data are precise enough to perform relevant analyses to inform
105 future research.

106 **Methods**

107 **Study population**

108 **Data collection**

109 Longitudinal data from an endemically infected MAP dairy herd in New York State in the
110 northeast United States were collected over eight years. The dairy herd consisted of approximately
111 330 cows. Johne's disease status of individual cows was determined *ante mortem* through biannual
112 faecal and quarterly serum sampling. Sampling of cows started at first calving. An additional 170 cull
113 cows could be tracked to the abattoir, where four gastrointestinal tissues and a faecal sample were
114 collected from each cow *post mortem*. The harvested tissues included two lymph nodes located at the
115 ileocecal junction and two pieces of ileum, one taken from 20 cm proximal to the ileocecal valve and
116 the other taken from very near the ileocecal valve. *Ante mortem* sampling commenced in February

117 2004 and continued until October 2010, and the last abattoir samples were taken in 2011. In total,
118 2.7% (114/4,158) faecal samples, 24.0% (149/621) tissues and 1.5% (89/5,937) serological samples
119 from 1,056 individual cows were MAP positive. The farm environment was sampled in approximately
120 20 locations on a biannual basis, resulting in 14.8% (34/230) positive bacterial cultures. In addition,
121 precise demographic data – including birth date, birth pen location, calving dates, fertility data, animal
122 pen locations, dry-off dates and eventually culling information and cull dates – were collected during
123 1988–2012. For a more complete description of sampling, see Pradhan et al. (7)(7) and Schukken et
124 al. (8)(8). Ethical approval was not required, as all samplings took place as part of an ongoing MAP
125 herd control programme.

126 The selection of the study population for this research was done retrospectively and was
127 based on the availability of sequenced MAP isolates. Accordingly, the study population consisted of
128 all MAP positive cows (n = 66) and all MAP positive environmental samples (n = 22) from which MAP
129 isolates (n = 150) could be successfully sequenced.

130 **Farm management and MAP control**

131 The farm participated in a MAP control programme, was well managed and had a good
132 hygiene status. It was a closed farm for years before the start of the study and did not purchase
133 animals during the study. Apparent herd-prevalence based on bacterial culture (faecal matter and
134 tissues) and serology was as high as 7.6% in 2004 and decreased to 0–2.4% in 2010 (23).
135 Throughout the study, the farm owner was informed about all test results and advised on optimal
136 management practices to reduce MAP prevalence. In terms of management groups, youngstock and
137 cows were transferred among 14 different locations: individual calf hutches, six calf and heifer rearing
138 pens, and seven freestall pens for cow groups in high and low lactation, including separated maternity
139 pens and sick pens. In 10-30% of calvings, maternity pens were used for more than one cow. Calves
140 were separated immediately after birth and were not allowed to nurse cows. Colostrum fed to calves
141 was from MAP-negative tested cows. Youngstock were not kept near adults, but indirect contact was
142 possible via employees. Bred heifers were housed on another nearby facility. Animal pen location
143 data were kept accurately so that animal location was reliably available on a daily basis. Based on
144 these pen location data a social network with number of days with direct pen contact between pairs of

145 cows was established for the subset of 66 cows. Pairs in this subset of MAP-shedding cows had on
146 average 142 contact days during their lives (median: 84; max: 1167). For 27% of pairs of cows no
147 direct pen contact was recorded. These were in particular cows born in different years and cows that
148 calved around 6 months apart from each other. The dataset contained sequences from four dam-
149 daughter pairs. Nine cows were super-shedders and seven cows were progressors; these cows were
150 culled between 6–29 months after their first positive MAP test. The decision when to cull was taken by
151 the farm owner based on economic reasons.

152 **Laboratory analysis, strain sequencing and genotyping**

153 Faecal samples, tissues and environmental samples were cultured in Herrold's egg yolk
154 media (HEYM) for up to 16 weeks at 37°C and shedding levels (cfu of MAP/tube) were determined.
155 Each culture with colony growth was sub-cultured. DNA was extracted from single bacterial colonies
156 sub-streaked on HEYM slants. The analytical protocol for bacterial culture was described in detail by
157 Pradhan et al. (7)(7)(7). For DNA extraction, strains were grown for 12 weeks at 37°C in Middlebrow
158 7H9 broth with 10% Middlebrook OADC, 0.05% tween 20, 1ug/ml micobactin J, and 0.01%
159 cyclohexamide from multiple colonies picked of HEYM slants. Epicentre's MasterPure Gram Positive
160 DNA extraction kit was used with the addition of a 20min 80C incubation prior to lysis. DNA was
161 prepared for sequencing with the Nextera XT DNA library kit and sequenced using Illumina HiSEQ
162 2500 2x100 paired end rapid run. Analysis of WGS sequences was performed using vSNP, National
163 Veterinary Services Laboratory's in-house single nucleotide polymorphisms (SNP) detection pipeline
164 (24). Briefly, the Illumina sequence reads for each isolate were mapped to the reference genome
165 MAP K-10 using the Burrows Wheeler Aligner (25) and Genome Analysis Toolkit (GATK) (26–28);
166 according to GATK best practices. Integrated Genomics Viewer was used to visually validate SNPs.
167 The final SNP alignment contained 150 sequences of 1,472 SNPs of the core genome, with collection
168 dates ranging from 17th February 2004 to 5th March 2008 (data published in S1 File). For detailed
169 genomic statistics see Richards et. al (29)(29).

170 **Analysis of strain diversity**

171 Strain diversity was estimated at cow-level (within-host strain diversity) and at herd-level
172 (within-herd strain diversity). Isolates sampled from the same cow were compared to isolates shed by
173 other cows. As a measure of genomic diversity, pairwise distance was calculated based on the
174 number of SNPs between each pair of isolates. Isolates with zero SNP differences are referred to as
175 isolates with identical genotype; isolates that differ by at least 1 SNP are referred to as different
176 genotypes. MEGA7 was used to estimate the maximum likelihood phylogeny (30).

177 **Reconstruction of transmission trees**

178 To reconstruct within-herd transmission trees, a phylogenetic network analysis was performed
179 with an algorithm called SeqTrack (31), implemented in the *adegenet* package (32) in R (33). This
180 algorithm is a graph-based approach recovering maximum parsimony phylogeny to identify the most
181 likely ancestries from aligned core SNPs in pathogen genomes. Jombart et al. based their method on
182 three observations: 1) each sampled isolate will only have one unique ancestor (in the absence of
183 recombination and reverse mutations), 2) descendants will always follow their ancestors in time, and
184 3) among all possible ancestries of a particular isolate, some are more likely than others, and this
185 likelihood of ancestry can be estimated from the genomic distance between sampled isolates (31). In
186 situations where several potential ancestors may exist for a given isolate, additional rules are needed:
187 the best ancestor will be selected by adding proximity information in the form of weighting matrices.
188 This rule is particularly relevant for slowly evolving pathogens where even a long-term, endemic
189 setting may result in low genomic diversity. These are indeed characteristics of endemic MAP
190 infections in dairy herds. In addition, SeqTrack's flexibility to incorporate various types of
191 epidemiological information with a number of different weighting matrices was judged to be beneficial
192 for the analysis of our detailed longitudinal cohort data as it allowed investigation of the additional
193 value of epidemiological data in the reconstruction of transmission trees.

194 Inference of ancestry follows a strict hierarchy: 1) temporal order: isolates with the earliest
195 dates are at the root, and those with the latest dates are at the tips of the reconstructed tree: we used

196 two different dates: inferred start of shedding and birth date of cow; see “Scenarios”, 2) genomic
197 distance: based on the number of SNP differences between pairs of isolates (entered as distance
198 matrix), 3) epidemiological weight: see “scenarios”, and 4) probability p of observing a given number
199 of mutations between an isolate and its ancestor: p was computed based on the mutation rate μ of the
200 pathogen (0.25 substitutions/core genome/year), time interval between each pair of isolates, and
201 length of partial nucleotide sequences (1,472 SNPs), using maximum likelihood. As the genomic
202 distance between two isolates a and b increases, p decreases that a is the direct ancestor of b .
203 SeqTrack only relies on epidemiological weights and p to resolve ties in the choice of ancestry: if an
204 isolate has more than one potential ancestor in identical genomic distance, the ancestor with the
205 higher weight is assigned. If two potential ancestors have the same weight, the ancestor with the
206 higher p value is assigned. The analysis is thus largely insensitive to μ .

207 **Extension of SeqTrack to endemic infection**

208 A number of extensions were made in this work to take SeqTrack a step forward to derive
209 individual infection chains for endemic infection with characteristics of MAP.

210 **Epidemiological unit**

211 SeqTrack was designed for epidemics where one single isolate is sampled from each case
212 and cannot capture within-host pathogen genomic diversity. To distinguish between mixed genotype
213 infections and within-host evolution, the transmission tree was built at isolate level (= epidemiological
214 unit), instead of case (cow) level. Most supporting epidemiological data were collected at cow level.
215 However, isolate specific parameters were sampling date, pen contacts of the cow at sampling day,
216 and duration of exposure to other MAP-shedding cows.

217 Transmission events at isolate level are referred to as ancestries between an ancestor and its
218 descendant; at cow level, the terms source of infection and recipient are used. A source or recipient
219 could be either another cow or an environmental sample.

220 **Duplicate genotypes**

221 If the identical genotype could be isolated multiple times from the same cow, duplicate
222 sequences ($n = 22$) were discarded. Continuous shedding from the earliest to the last sampled isolate
223 of this genotype was assumed and the infectious period was set as described in the next section. A
224 total of 128 isolates were included in the reconstruction of transmission trees.

225 **Scenarios**

226 Temporal order of sampling can be misleading owing to variable delays between exposure
227 and sampling, even for MAP with such a slow rate of mutation. The challenge of determining the
228 probable time window when MAP-positive cows became infected and infectious was addressed by
229 comparing six infection scenarios. We hypothesize that if several scenarios resulted in the same
230 choice of ancestor for a given isolate, this would add support to the accuracy of reconstructed
231 transmission chains.

232 **Scenario [Basic] - Basic transmission tree:** based only on genomic distance (without any
233 weighting).

234 **Scenario [E] - Weighting by exposure time:** [Basic] plus weighting matrix with number of
235 days cow X spent in the same pen with any other cow Y during cow Y's infectious period before cow
236 X started to shed (Fig 1). The number of days of exposure time [E] was calculated for each pair of
237 isolates and entered in a matrix with 128 rows x 128 columns (one column and row for each isolate).
238 The value of [E] was then used as the weight, with highest weights for longest exposure. For cows
239 with several MAP genotypes, the duration of the infectious period and [E] were calculated for each
240 genotype separately. The genotype-specific infectious period was defined as starting at the mid-day
241 between last negative and first positive sampling date and ending at the mid-day between last positive
242 and consecutive negative sampling date. The infectious period for abattoir samples ended the day
243 before culling. The inferred mean infectious period was 95 days (min: 1 day for cows that tested
244 negative the day before slaughter, median: 71, max: 779 days for cows shedding the same genotype
245 serially at several sampling dates). Accordingly, the mean duration of [E] for pairs of isolates sampled

246 directly from cows was 8 days (min–median–max: 0–0–419 days), with 82% of pairs of isolates with 0
247 days of [E].

248 **Fig 1. Example of contacts between two cows (X and Y) over time.** The exposure time [E] is the
249 time cows X and Y have pen contact and cow Y sheds a certain MAP isolate, whereas cow X does
250 not yet shed. The vertical arrows indicate the shedding starts of both cows. The shedding start of cow
251 X corresponds to the end of [E]. Overall pen contact days: contact during and outside [E].

252 Environmental samples were assumed to represent (potentially non-detected) infectious
253 cows. Their “infectious period” of spill-back was therefore defined in the same manner as for cows,
254 and its start was used as the date for both [birth] and [shed] scenarios (explained below). The
255 environment served as a potential source of infection for all cows that were, during, the spill-back
256 period, at the location where the environmental isolate was sampled. Average duration of spill-back of
257 the 22 environmental samples was calculated to be 91 days (min–median–max: 42–91–116), and
258 their mean duration of [E] was 26 days (min–median–max: 0–0–116 days), with 69% of pairs of
259 environmental isolates and isolates sampled directly from a cow with 0 days of [E].

260 **Scenario [S] - Weighting by susceptibility:** [Basic] plus [E] plus additional weighting matrix
261 to reflect the decreasing susceptibility of cows over time. Seven social network patterns (weights from
262 6 to 0) were used to weight potential transmissions based on age of the susceptible animal at contact
263 and duration of its exposure. The weighting order was based on accepted knowledge of MAP
264 epidemiology knowledge. Each pair of isolates was assigned one weight that reflected their
265 epidemiological link:

- 266 6. **Cow-to-calf contact:** (direct or indirect) cow-to-calf contact within the first days of life of a
267 newborn calf (maximum weight for isolates from own dam or any other cow present in the
268 maternity pen that calved \pm 15 days around birth date of the cow),
- 269 5. **Calf-to-calf contact:** direct contact in the first year of life of a cow within the same age cohort
270 (other cows born \pm 30 days around birth date of the cow),
- 271 4. **Adult-to-adult contact during the infectious period:** pen contact ($[E] \geq 1$) during adulthood,
- 272 3. **Longer direct contact:** long pen contact during adulthood with cows outside their infectious
273 period (≥ 100 pen contact days, but $[E] \geq 0$),

274 2. **Limited direct contact:** limited pen contact during adulthood with cows outside their infectious
275 period (1–99 pen contact days, but $[E] \geq 0$),
276 1. **Indirect contact:** pairs of cows which lived on the farm during the same period, but with no
277 recorded pen contact days, and
278 0. **No contact:** one cow was culled/sold before the other cow was born.

279 [Basic], [E] and [S] scenarios were each calculated with two different dates to account for the
280 uncertainty of the temporal sequence of exposure times: [birth]: birth date of cow, and [shed]:
281 potential start of MAP shedding and thus of the (genotype-specific) infectious period (for calculation of
282 [shed] see [E] scenarios). These two dates were selected to investigate how ancestries change if
283 susceptibility is put as the focus of infection dynamics (with scenarios using [birth]) versus
284 infectiousness (with scenarios using [shed]). As infection spread is driven by both infection states, it
285 was expected that [birth] and [shed] scenarios with [E] and [S] weights would result in more similar
286 trees than [Basic] scenarios. In total, six scenarios were then calculated, namely [birth_Basic],
287 [birth_E], [birth_S], [shed_Basic], [shed_E] and [shed_S].

288 SeqTrack will define the best fitting transmission tree based on maximum parsimony. The
289 basic model without weights as described in these scenarios will logically provide the maximum
290 parsimony model. The basic model is only based on genomic distances and does not take
291 epidemiological limitations into account. A genomic connection between two isolates from two cows
292 that in real life were never on the farm at the same time is acceptable in the basic model, but will
293 receive a very low weight in the model expanded with epidemiological information. Therefore,
294 maximum parsimony should only be compared within the same scenario or between scenarios when
295 no conflicting epidemiological information is present.

296 **Within-herd circulation of genetically diverse strains**

297 SeqTrack assumes monophyletic genomic data (infection caused by a single external source)
298 and will add all isolates into one single transmission tree, independent of how distant (and thus less
299 likely) reconstructed ancestries may be. With the low mutation rate of MAP, strain diversification within
300 the study period through evolution was expected to be limited. A genomic distance threshold of 6

301 SNPs was defined based on the overall herd-level genomic diversity (Fig 2). Pairs of isolates
302 exceeding this threshold were considered not to have arisen from directly linked cases. If no ancestor
303 within this threshold could be found in the sample for a particular isolate, it was set as the root of a
304 separate transmission tree, indicating that the true ancestor had not been sampled. Generally, the
305 lower a threshold of number of SNPs is chosen, the more transmission trees with multiple generations
306 will be broken up into smaller individual trees or unconnected singleton isolates (resulting trees of a
307 sensitivity analysis with different threshold values are not shown).

308 **Fig 2. Distribution of pairwise genomic distances (n = 150 isolates with 1,472 SNPs).** The
309 dashed line marks the genomic distance threshold of 6 SNPs.

310 **Censored data**

311 Non-sampled early generations of cases (before study start) would lead to overestimation of
312 the number of descendants for isolates at the root of the transmission tree. In the absence of their
313 true ancestor in the sample, SeqTrack assigns more descendants to the earliest sampled isolates. In
314 addition, data were right censored, and the number of descendants were underestimated, particularly
315 for isolates sampled in the late phase of the study. Whereas the transmission trees were
316 reconstructed with all 128 isolates, the following conservative assumptions were made for the
317 estimation of genotype-specific and cow-specific reproduction ratios to account for temporality in the
318 data structure: For the earliest 10% of isolates only one third of the assigned descendants were
319 assumed to be their true descendants. For the latest 10% of isolates the number of descendants per
320 isolate were not calculated as these descendants were not yet fully sampled within the study period.
321 Consequently, the role of individual cows in infection spread was analysed for the remaining 84
322 isolates, sampled from 57 cows.

323 **Lack of genomic resolution**

324 If isolates with identical genotype can be sampled over years from generations of cows, a
325 range of alternative infection chains may exist (Fig 3). In addition to the number of recipients
326 according to the one, optimal tree of SeqTrack, the “potential number of recipients” was calculated for

327 each cow, assuming that all isolates with identical genotype that fulfilled certain criteria could be
328 descendants of the same cow. The criteria were those for [S] weights 4–6, as described under
329 “Scenarios”.

330 **Fig 3. Reconstructed transmission trees.** Transmissions are depicted by edges and isolates by
331 vertices in a directed network. Edge labels and edge colour indicate number of SNPs of differences
332 between ancestor and descendant. **(A)** and **(B)** two alternatives of potential infection chains of five
333 isolates with identical genotype (dark grey). More alternatives exist.

334 **Network analysis**

335 **Analysis at isolate level**

336 A MAP genotype-specific effective reproduction ratio R_{GT} was calculated by aggregating the
337 number of descendants per genotype at herd level. Reconstructed transmission trees resulting from
338 all six scenarios were compared and changes in the branching of the trees due to epidemiological
339 weightings were assessed. Ancestries of pairs of isolates were identified that were identical in two or
340 more scenarios. By comparing the maximum likelihood p of all individual ancestries across scenarios,
341 overall statistical support for each scenario was assessed.

342 **Analysis at cow level**

343 Number of recipients produced during lifetime (animal-specific effective reproduction ratio R_A),
344 “potential” recipients and sources of infection per cow were quantified by summing up ancestors and
345 descendants of isolates. Support of the reconstructed infection chains for the current prevailing
346 hypotheses on transmission routes was evaluated by quantifying epidemiological links of ancestries
347 across scenarios: all reconstructed ancestries were retrospectively matched with the seven social
348 network patterns defined under [S].

349 Results on the number of recipients produced were validated against the literature on risk
350 factors for MAP spread. For this purpose, associations between four *ante mortem* detectable Johne’s

351 disease phenotypes of the cows in the study population and their number of recipients were tested for
352 all six scenarios. The four phenotypes were: **shedding level** (three levels: [0] faecal culture negative,
353 [1] low (1–50 cfu/tube of faecal matter), [2] high (>50 cfu/tube of faecal matter)), **age at first**
354 **shedding** (three levels based on Mitchell et al. (10)(10): [0] first positive faecal sample collected
355 before the age of 3 years, [1] first positive faecal sample >3 years, [2] cow was *ante mortem* never
356 positive), **infection progress** (Inclusion criteria: individual cows with at least four MAP culture results,
357 and at least one faecal sample taken after a positive MAP culture. Three levels: [0] faecal culture
358 negative, [1] non-progressor, [2] progressor), and **serostatus** (two levels: [0] no ELISA-positive test,
359 [1] at least one ELISA-positive test). For shedding level, infection progress and age at first shedding
360 correlation was calculated with Spearman's rank correlation. For the binary variable serostatus, the
361 difference between mean number of recipients was calculated with Welch's two sample *t* test.

362 The R code for this analysis is published in the Supporting Information S2 File.

363 **Results**

364 **Strain diversity within the host over lifetime (at animal- 365 level)**

366 Up to 8 MAP isolates could be sequenced per cow. Cows had up to 5 non-clonal MAP
367 genotypes. For 43 cows with only one MAP isolate sampled, strain diversity could not be assessed.
368 Out of 23 cows with 2–8 isolates, only two (8%) shed the identical genotype in series at different
369 sampling days. Only from three (13%) cows could an identical genotype be isolated both *ante mortem*
370 from faecal matter and *post mortem* from tissue; all three were super-shedders and progressors. MAP
371 positive tissue – confirming true infection – contained 1–3 different genotypes per cow. For eight
372 (35%) cows, all isolates differed by a maximum of 6 SNPs. The remaining 15 cows had at least one
373 isolate with 7–234 SNPs of difference, indicating mixed infection.

374 Strain diversity within-herd (at herd-level)

375 A total of 94 different genotypes were recovered from all 150 sequenced isolates (Fig 4); of
376 these 84 (89%) were only detected once. The most prevalent genotype was isolated 32 times from
377 twelve cows and four environmental samples. Several MAP strains with genomic distances of more
378 than 100 and 200 SNPs between strains were recorded (Fig 2). With an estimated mutation rate of
379 one substitution per 2–7 years, a genomic distance of 100 SNPs indicates multiple introductions of
380 MAP strains rather than within-herd evolution from a common ancestor. Two dominant strains (D1
381 and D2) of MAP could be detected during the whole study period. These dominant strains were
382 responsible for 19% (D1) and 35% (D2) of MAP infections. Dominant strains consisted of clusters of
383 24 (D1) and 21 (D2) genotypes that differed by maximally 4 and 7 SNPs, respectively. Each dominant
384 strain had one “super-spreading” genotype which resulted in 27 (D1) and 31–33 (D2) descendants,
385 depending on the scenario. Averaged over the whole cluster of genotypes, the R_{GT} of the dominant
386 strains were 1.2 (D1) and 1.7–1.8 (D2). Genotypes that did not belong to the dominant strains had an
387 average R_{GT} of 0.4.

388 **Fig 4. Maximum likelihood phylogeny of sequenced MAP isolates.** Phylogenetic tree based on
389 SNP data from 128 sequenced MAP isolates with 1472 nucleotide positions, using the Tamura-Nei
390 model. A total of 94 different genotypes were recovered. The tree with the highest log likelihood (-
391 7449.75) is shown. Branch lengths is measured in the number of substitutions per site. D1 and D2
392 indicate isolates belonging to the two dominant strains.

393 Reconstructed transmission trees

394 Transmission trees showed similar features in all six scenarios: two main trees, a range of
395 small trees with 2–4 generations and singular, unconnected isolates. These unconnected isolates
396 were not (closely) related to any other isolate sampled on this farm and there was no indication that
397 they spread during the study period within the herd. Main trees were formed by a cluster of genotypes
398 of one of the dominant strains. Remarkable features of main trees were long branches of infection
399 chains with isolates of identical genotype, and 2–4 super-spreaders (each with 10–20 descendants),

400 plus variable numbers of isolates with 2–6 descendants. Characteristics of these features differed
401 considerably between scenarios (Fig 5; for [shed] scenario trees see S1 Fig): two long branches with
402 8 and 15 generations of isolates with identical genotype and super-spreaders at or close to the root
403 were particularly prominent in [Basic] scenarios. These long infection chains had short time-intervals
404 between generations of descendants of a few weeks to months, untypical for MAP. [E] and [S]
405 scenarios resulted in more branched trees, less prominent super-spreaders, and more isolates
406 serving as ancestors for up to 6 descending isolates. Infection chains in [E] and [S] scenarios had a
407 maximum of 7–8 generations, indicating longer time-intervals between two transmission events.
408 Individual transmission trees were composed of exactly the same isolates in [birth] and [shed]
409 scenarios, but in different orders.

410 **Fig 5. Reconstructed transmission tree of three scenarios (n = 128 isolates). (A)** [birth_Basic],
411 **(B)** [birth_E], **(C)** [birth_S]. Isolates sampled from the same cow are labelled with successive numbers
412 and are shown in vertices of the same colour and outline. White vertices represent cows with only one
413 isolate. Dark green vertices (labelled 99–120) represent environmental samples. Edge labels and
414 edge colour indicate number of SNPs of difference between ancestor and descendant.

415 Thirty-eight (30%) isolates were assigned the identical ancestor in all six scenarios, and for 6
416 (5%), 18 (14%) and 49 (38%) isolates the algorithm returned five, four or three times with an identical
417 ancestor, respectively. No isolate was assigned a different ancestor in every scenario, and for each
418 isolate, at least one ancestral isolate could be identified with more statistical and/or epidemiological
419 support. The three [birth] scenarios and the three [shed] scenarios showed more identical ancestries
420 amongst themselves (64–80%) than [birth] compared to [shed] scenarios (45–61% identical
421 ancestries). [birth_E] and [shed_E] shared 61%, [birth_S] and [shed_S] shared 60%, and [birth_Basic]
422 and [shed_Basic] shared 48% of ancestries. Scenarios with weighting for exposure time and
423 susceptibility were thus closer to each other than [Basic] scenarios without any epidemiology
424 incorporated. Overall statistical support was numerically highest for [birth_Basic], followed by
425 [birth_E], [birth_S], [shed_Basic], [shed_S] and [shed_E]. [Basic] scenarios will, by definition, have the
426 highest p values, as they represent the most optimal genomic tree, but they lack epidemiological
427 support. Epidemiological weighting incorporated in [E] and [S] can only deviate from this optimal tree,
428 but outweighs loss of numerical credibility by epidemiological reliability. Differences in statistical

429 support were small (all scenarios had 34–35 ancestries with $p > 0.95$; and the average p of ancestries
430 ranged from 0.34–0.35 across all six scenarios), indicating that epidemiology-informed scenarios had
431 similar statistical support as [Basic] scenarios (data published in S1 Table).

432 Adult-to-adult contact during the infectious period was in all scenarios except [birth_Basic] the
433 most frequent social network pattern leading to infection, followed by direct contact during adulthood
434 outside the infectious period and indirect contact (Table 1). Remarkably, cow-to-calf contact was the
435 least important transmission route (0–4% or 0–4 ancestries in each scenario) and was even less
436 frequent than ancestries with no epidemiological link at all. Genomic distances of the four dam-
437 daughter pairs in the sample were 0, 25, 26 and 97 SNPs. For the only dam-daughter pair with
438 isolates of identical genotype, only in [birth_S] was the dam assigned as the direct ancestor to her
439 daughter. In all other scenarios dam and daughter were in the same tree, but either separated by 1 or
440 5 generations, or dam and daughter were in the same generation and assigned to a common
441 ancestor. Inferred ancestries with no epidemiological link indicate the presence of non-sampled or
442 incompletely sampled cows.

443 **Table 1. Ranking of transmission routes.** Ranking of transmission routes by the proportion of
444 inferred ancestries based on social network patterns. Column percentages add up to 100%. The
445 overall rank was inferred from the sum of ranks of all scenarios.

Weight	Description	Scenarios						Rank
		[birth	[birth	[birth	[shed	[shed	[shed	
		_Basic]	_E]	_S]	_Basic]	_E]	_S]	
6	Cow-to-calf	0.0%	0.0%	4.0%	1.0%	0.0%	3.0%	7
5	Calf-to-calf	5.9%	6.9%	11.9%	3.0%	4.0%	11.1%	5
4	Adult-to-adult contact during the infectious period	24.8%	49.5%	42.6%	33.3%	69.7%	59.6%	1
3	Longer direct contact	29.7%	20.8%	24.8%	19.2%	7.1%	9.1%	2
2	Limited direct contact	14.9%	9.9%	12.9%	17.2%	9.1%	11.1%	3
1	Indirect contact	16.8%	7.9%	3.0%	19.2%	7.1%	5.1%	4
0	No contact	7.9%	5.0%	1.0%	7.1%	3.0%	1.0%	6
n	Number of isolates for which ancestries could be inferred ^a	101	101	101	99	99	99	

446 ^a For 128 – 101 = 27 isolates (in [birth] scenarios) and for 128 – 99 = 31 isolates (in [shed] scenarios),
447 no ancestor could be found within the genomic distance threshold of 6 SNPs and, thus, no
448 transmission route could be inferred. Each of these 27 and 31 isolates was the root of a separate
449 transmission tree.

450 **Role of individuals in MAP spread**

451 Seventy percent of cows had only one source of infection; 22%, 7% and 1% had two, three
452 and four different sources, respectively. In all scenarios, transmission of 2–3 genotypes between pairs
453 of cows was recorded, indicative of mixed infections. For nine (39%) of a total of 23 cows with multiple
454 isolates, the algorithm indicated within-host evolution: three cows had 3, and one cow even 4 directly
455 linked isolates. Ancestries within the collection of isolates sampled from the same cow were more
456 frequent in [E] and [S] than in [Basic] scenarios. In [birth] scenarios, each cow infected on average
457 1.2–1.3 cows, and 46% of animals spread infection to at least one cow, with a maximum of 9–17
458 recipients. In [shed] scenarios, mean R_A (0.9–1.0), percentage of spreaders (33–42%) and maximum
459 number of recipients (5–7) were lower.

460 With respect to alternative infection chains due to lack of genomic resolution, the potential
461 number of recipients per cow was on average 1.5 [birth] / 1.3–1.4 [shed], and super-spreaders had up
462 to 10–20 [birth] / 7–10 [shed] potential recipients, depending on the scenario.

463 Regarding the environment, the same genotype was only twice isolated from the environment
464 and from a cow on the same day. This cow may thus have been the most likely contaminating source.
465 The source of the remaining 20 samples from the farm environment remained undetected. For 11
466 (50%) isolates from the environment at least one scenario resulted in spill back (or spread by the cow
467 that originally excreted that particular isolate) to 1–6 cows. One environmental sample was even a
468 super-spreader in [shed_E] and [shed_S] scenarios.

469 **Correlation between reconstructed number of recipients 470 and disease phenotypes**

471 For all scenarios, some correlation between number of recipients per cow and a cow's MAP
472 shedding level, age at first shedding and pattern of infection progress was observed (Table 2).
473 Figures with numbers of recipients for all investigated phenotypes are published in S2 Fig. There was
474 good to weak evidence (p of 0.02–0.08) that high shedding animals produced more new infections

475 compared to low shedders or cows that were *ante mortem* never tested faecal culture positive. Age at
476 onset of shedding was negatively correlated with number of recipients produced during lifetime: the
477 earlier a cow started to shed, the greater risk for spreading she posed. This correlation was significant
478 ($p = 0.01\text{--}0.04$) for all scenarios except [birth_Basic] ($p = 0.14$). Regarding infection progress, in all
479 scenarios progressors had the highest mean number of recipients. However, only [E] scenarios
480 showed good evidence for a correlation between pattern of infection progress and number of
481 recipients (p of 0.02 and 0.05). Serostatus was the only investigated phenotype that did not appear to
482 be associated with a risk for spreading. Cows with antibody response, had at least in [birth] scenarios
483 higher numbers of recipients produced compared to animals with only ELISA-negative serum
484 samples. However, no scenario resulted in strong evidence for a more important role of cows with
485 measurable immune response for MAP spread (p of 0.07–0.95). To summarize across all six
486 scenarios, p values were generally smallest for [birth_E] and [shed_E] scenarios, which resulted in
487 significant correlations for three of four investigated phenotypes.

488 **Table 2. Association between number of recipients of an individual cow and her disease**

489 **phenotype, by scenario.** For phenotypes “shedding level”, “age at first shedding” and “infection
490 progress”, Spearman’s rank correlation coefficients and (*p* values) are presented. For “serostatus”,
491 mean number of recipients for both phenotype levels and (*p* values) of Welch’s two sample *t* test are
492 presented (n = 57 cows).

Disease phenotype	Scenarios					
	[birth_Basic]	[birth_E]	[birth_S]	[shed_Basic]	[shed_E]	[shed_S]
Shedding level^a	0.28 (0.04*)	0.32 (0.02*)	0.28 (0.04*)	0.25 (0.06)	0.27 (0.04*)	0.24 (0.08)
Age at first shedding^b	-0.20 (0.14)	-0.33 (0.01*)	-0.29 (0.03*)	-0.30 (0.02*)	-0.29 (0.03*)	-0.28 (0.04*)
Infection progress^c	0.32 (0.06)	0.38 (0.02*)	0.28 (0.09)	0.33 (0.06)	0.34 (0.05*)	0.26 (0.13)
Serostatus^d	1.0 / 1.7 (0.53)	1.0 / 2.3 (0.25)	1.1 / 2.1 (0.24)	1.1 / 0.5 (0.07)	0.9 / 0.9 (0.95)	1.0 / 1.0 (0.93)

493 *Good evidence against the null hypothesis of Spearman’s rank correlation coefficient = 0 at the
494 significance level of *p* < 0.05.

495 ^a Sheding level, levels: 0 – always faecal culture negative, 1 - low, 2 – high.

496 ^b Age at first shedding, levels: 0 - ≤3 years, 1 - >3 years, 2 - *ante mortem* negative.

497 ^c Infection progress, levels: 0 - *ante mortem* negative, 1 - non-progressor, 2 – progressor.

498 ^d Serostatus, levels: 0 - ELISA-negative, 1 - ELISA-positive.

499 Discussion

500 Key findings

501 This work presents a method to reconstruct who-infected-whom in an endemic Johne's
502 disease setting, by joining WGS data with explicit longitudinal data. Up to 5 different MAP genotypes
503 could be isolated from individual cows. Genomic distances between isolates were far beyond that
504 expected within-herd and within-host with evolution over time, providing a strong indication for
505 multiple introductions of MAP strains into herds and mixed genotype infections between cows.
506 Reconstruction of transmission trees led to consistent results: cows acquired infection from 1–4
507 different sources and spread infection to 0–17 recipients, suggesting repeated exposure to shedding
508 animals at different points in time or mixed shedding of one source which led to infection with a
509 heterogeneous inoculum. In the light of low test sensitivities and undetected MAP cases, these
510 numbers should be considered as conservative estimates. For each isolate at least one more likely
511 ancestor could be inferred. For 49% of isolates, even 4 out of 6 infection scenarios resulted in the
512 same choice of ancestor, adding support to the accuracy of reconstructed transmission chains.

513 Based on comparison of transmission trees and relevant epidemiological constraints, the
514 model that resulted in the best 'who infects whom' answer was the [birth_E] scenario, which included
515 time ordering of isolates based on birthdate of the host and a weighting matrix with number of days
516 that a given susceptible cow spent in the same pen with another cow during her infectious period (Fig
517 5B). The [birth_E] scenario had not only the highest statistical support for the reconstructed
518 transmission tree among all scenarios with incorporated epidemiology, it also showed the strongest
519 evidence for associations between disease phenotypes of cows and their number of infected
520 recipients. On this farm with a well implemented MAP control program this transmission model
521 favouring horizontal transmission was preferred over other possible models of MAP transmission
522 where weighting preference was given to increased susceptibility in young animals.

523 Concurrent circulation of dominant strains could be recorded over several years, indicating
524 that some strains were more successful in terms of transmission and infection progression (7,8).
525 Other important features of transmission trees were some minor strains that could only be recorded

526 over 2–4 generations of transmissions, and singleton isolates not related to any other isolate.
527 Transmission studies assuming spread of a monophyletic strain will certainly underestimate the
528 complexity of multiple infection chains occurring in parallel.

529 A particular challenge was inconclusive ancestries for clusters of isolates with identical
530 genotype within the dominant strains. In situations with simultaneous exposure of a susceptible
531 individual to multiple shedders of the same genotype, neither genomic distance nor contact data can
532 be used to resolve ties in ancestries. However, the critical question is: is it relevant to know whether
533 cow X or cow Y infected cow Z, given that large parts of the operation were perpetually contaminated
534 with one strain? Management-wise, a holistic control strategy would be required, as removing single
535 known shedders could result in limited success in interrupting the infection cycle.

536 **Mixed infections**

537 Mixed infections are common. This finding adds complexity to the estimation of standard
538 measures in epidemiology, such as the effective reproduction ratio R . Despite the decreasing
539 prevalence during the study period, several scenarios resulted in an animal-specific $R_A > 1$. An
540 explanation for this contradiction is that R_A does not consider that some cows acquire infection more
541 than once whereas the proportion of cows that apparently remained MAP negative gradually
542 increased over time. Dependent on this clustering of co-infections with different MAP genotypes, R_A
543 will be considerably larger than the mean number of newly infected recipients generated in the herd
544 (R_H) from one generation to the next in the infection chain. This leads to two conclusions: First,
545 infected cows remain susceptible to co-infections with other genotypes. This finding will be relevant
546 for mathematical models: transition between infection states might not be as clear-cut as often
547 assumed. Second, $R_A > R_H$ indicates that some cows have a lower risk of becoming infected than
548 others. How much of these differences can be explained by varying exposure intensity, susceptibility
549 or an alternative explanation was beyond the scope of this study and remains to be investigated.

550 **Association between disease phenotypes and infection**

551 **spread**

552 For cows with high levels of shedding, early age at shedding onset and progressive infection,
553 the algorithm returned higher numbers of recipients. These phenotypes measure two concepts: high
554 shedding and exposure over longer time periods as risk factors for MAP transmission. This finding is
555 in line with the literature: super-shedders represent a greater risk for spreading, and close contact
556 with a case and repeated uptake of high doses of bacilli may lead to adult-to-adult infection (4,5,11).
557 Statistical support for this correlation was consistently strong for [E] scenarios weighted by exposure
558 time. This could indicate that duration of exposure to a shedder was a decisive factor for MAP
559 transmission on the investigated farm. Of note, serostatus was the only investigated disease
560 phenotype that appeared not to be associated with risk for spreading. An explanation of this result
561 could be the limited antibody response measured in this herd: only 16% of cows with sequenced MAP
562 isolates were also ELISA-test positive. False-negative test results could have potentially led to
563 misclassification in relation to a cow's true immune response to MAP infection.

564 The results linked to disease phenotypes presented here should be considered as
565 exploratory; they aimed to show the possibilities of this novel approach to study MAP transmission. A
566 more detailed follow-up is needed to take account of the effects of a changing environment over time,
567 such as reduced infection pressure due to adaptations of the farm management and the decreasing
568 prevalence. Relevant questions in this respect are: what happens if super-shedders are removed: can
569 we see - or even quantify - an effect in the transmission dynamic of the remaining genotypes in
570 response to the control intervention? Will another cow take over the role of a super-spreader or does
571 removing high shedding animals indeed reduce R ?

572 **Transmission routes**

573 Adult-to-adult contact during the infectious period of the source cow was the most important
574 transmission route to explain the reconstructed networks. This finding is consistent with Schukken et
575 al. (8), who showed that cows that were infected with a particular MAP strain were significantly more

576 exposed as adults to cows shedding the same strain compared to cows that were culture negative for
577 MAP at slaughter. Cow-to-calf and calf-to-calf contacts during early life accounted for less than 1 in 12
578 transmissions. These analyses thus did not support the hypothesis that dam-daughter infections were
579 the principal transmission route on the farm under investigation. This farm had implemented rigid
580 interventions for MAP control, daughters were separated from dams as quickly as possible and
581 contact between age groups was limited. This could explain the minor role of dam-newborn calf and
582 calf-to-calf transmissions and indicate success of the implemented interventions. However, MAP
583 exposure during adulthood appeared to be sufficient to maintain infection over years. To confirm this
584 hypothesis, the analysis needs to be repeated for farms with different calf management protocols to
585 investigate correlation between interventions and contribution of transmission routes for long-term
586 persistence of MAP.

587 **Strengths of this study**

588 In the presence of within-host diversity, analysis of a single isolate will miss clonal, closely
589 related or distant strains shared between the source and its recipients, resulting in inaccurate
590 conclusions about transmission. Particularly with pathogens with slow divergence, uncertainty will
591 remain even if all genotypes are observed, as individual transmission routes cannot be resolved by
592 sequence data alone (22). The conclusions of this study were based on a unique dataset that allowed
593 the study of individual transmission networks of genotypes and diversification of the MAP population
594 within-host and within-herd over time and in many facets. Only the combination of precise longitudinal
595 data on infection status, detailed demographic data and bacterial strain type information of a densely
596 sampled population made it possible to trace infection in such detail.

597 **Choice of method**

598 A large number of algorithms to infer transmission chains using genomic data have been
599 developed in the past. The majority of these methods limit their use of epidemiological field data to
600 one or two date variables, such as a) sampling, collection or observation date of cases to infer
601 infection date or start of the infectious period, and b) removal or culling date for studies in the

602 veterinary field as indicators for the end of the infectious period (22,34–40). A smaller number of
603 methods include an additional spatial component, such as location or geographical distance between
604 cases (41–46). Another limitation of most of these methods is, that they were primarily designed for
605 pathogens that accumulate enough new mutations during the sampling period. The limited new
606 genomic variation observed within the 8-year sampling period of this MAP study might preclude
607 inference of transmission chains predominately based on sequencing data.

608 Only recently, have methods been published which allow the integration of additional
609 epidemiological information (47–49). These authors highlight the need to expand outbreak
610 reconstruction tools to utilize other types of epidemiological data due to limited information value of
611 sequence data (48), or even conclude for certain outbreaks that contact data is equally or more
612 informative than sequencing data (47).

613 Our study objectives required an algorithm that supported incorporation of a much broader
614 collection of detailed epidemiological field data. SeqTrack proved to be a suitable backbone, versatile
615 enough to be expanded for investigations of endemic disease. The algorithm has limitations regarding
616 inference of the underlying transmission process and timing of exposure. However, SeqTrack
617 supports incorporation of all kinds of data to describe proximity in the form of weighting matrices, a
618 criterium we prioritized in the assessment of available methods to apply in our study as strain
619 diversification was expected to be limited. Weighting by exposure time [E] was based solely on
620 explicit data and weighing by susceptibility [S] was based on accepted knowledge of MAP
621 epidemiology. Of note, scenarios informed by data and knowledge resulted in more consistent
622 ancestries compared to scenarios based solely on genomic data. On purpose, no vague assumptions
623 about transmission routes were made, to enable independent evaluation of reconstructed networks
624 for transmission routes. To allow for ancestries with short time-intervals between two transmission
625 events, no minimum duration of pre-infectious period was entered. Hence, transmission routes for
626 scenarios in which time between MAP uptake and shedding is assumed to be short, such as calf-to-
627 calf transmission or pass-through shedding of adults, could be investigated. Testing for correlation
628 between phenotypes and number of recipients in reconstructed networks demonstrated that the
629 results produced with this method are consistent with the literature. Data on a disease phenotype
630 were thus utilisable for validation and gave further support to the accuracy of the inferred ancestries.

631 The authors are confident that this method will also be valuable for estimation of strain-specific
632 transmission parameters.

633 An apparent limitation of SeqTrack is its strong dependence on temporal ordering of isolates.
634 The algorithm does not clearly account for the potentially long delay between unobserved exposure
635 and observed sampling events. Due to the long incubation period of MAP, the order of ancestry
636 cannot be inferred with certainty. Other approaches such as Bayesian algorithms will be superior in
637 the inference of the underlying transmission process and timing of exposure. Nevertheless, limiting
638 the problem to more exact inference of time of exposure of recipients would clearly underestimate the
639 complexity of MAP epidemiology. Even if we knew the time point of infection, we would still face the
640 challenge of the long incubation period, lesion formation and open questions on the role of pathogen-
641 host immunity interplay for the start of bacterial shedding of the recipient cow (50). Therefore, we
642 chose an approach where we investigated with three birth and three shed scenarios, two opposing
643 temporal indicators as proxies for the time of infection and time of becoming infectious.

644 **Value of field data**

645 Unquestionably, availability of sequencing data allows us to enter a new era of transmission
646 studies. This study aims to contribute to the efforts towards investigation of endemic infection spread.
647 In this field, MAP poses great challenges as its detection and cultivation are difficult and time-
648 consuming. For this study, although one of the most detailed MAP cohorts published to date, ideally
649 (at least) one more generation of cows should have been sampled to reduce the impact of both right-
650 and left-censored data on the results. As MAP prevalence decreases, sampling effort increases even
651 more compared to each WGS sequence gained in return. In this study, on average 33 faecal, tissue
652 or environmental samples were cultured per sequenced isolate; towards the end of the study this
653 number increased towards 100 samples for one recovered sequence. Particularly for studying low
654 prevalent MAP, field studies will quickly reach their resource limit (in the absence of diagnostic
655 methods to identify MAP infected individuals more efficiently). However, only field data can show the
656 true variation in incidence, change of prevalence and effects of control interventions on transmission
657 in a non-controlled environment. This information underlies the basis for forming hypotheses to be

658 followed up with experimental studies in animal models under controlled conditions and for
659 parameterization of modeling exercises to complement field studies.

660 **Limitations**

661 Although all cows were sampled periodically, the within-herd MAP population was only
662 incompletely sampled. Low test sensitivity will have led to unobserved infected cows or additional
663 MAP genotypes shed by known MAP-shedders. Since infection source is only attributed to sampled
664 isolates, adding such non-detected isolates to the analysis could either result in new common sources
665 of infection for separate transmission trees or cause additional generations between (wrongly)
666 inferred direct ancestries within a transmission tree. Nevertheless, the original isolates would remain
667 in the same transmission tree. The reproductive ratios at genotype level R_{GT} would have been higher;
668 the above presented values thus represent estimates at the low end. With more cows potentially
669 contributing to MAP spread, the R_A would have varied for individual cows. Assuming that all
670 genotypes were affected to a similar extent by the non-perfect test characteristics, the authors believe
671 that the main conclusions of this work remain valid.

672 Genotypes isolated from the environment could be simultaneously detected in individual cow
673 samples in only 9% of cases, and for a small proportion of ancestries no epidemiological link was
674 recorded at all, indicating unobserved cases. The observed strain diversity in tissues sampled in
675 parallel, highlights the value of sequencing more than one individual colony from a sample to
676 understand within-host diversity and to infer ancestries with more certainty. An optimized combination
677 of sampling and diagnostics is needed to capture genomic diversity with minimized sampling effort.

678 As youngstock were not sampled, no genomic data to directly confirm calf-to-calf or adult-to-
679 calf transmission were available. From the data it could not be determined whether genotypes that led
680 to the initial infection were among the available collection of isolates from a cow. In addition, temporal
681 order of isolation was not necessarily in the order of genomic evolution: mutated genotypes might
682 have been sampled before their ancestor. The sensitivity of SeqTrack to time was taken into account
683 by calculating scenarios with two different time indicators: it could be shown that the order of isolates

684 within an infection chain was time dependent. However, isolates were assigned to the same infection
685 chain, independent of their date variable.

686 SeqTrack is highly sensitive to genomic distance between isolates. SNP distances may vary
687 for technical reasons, and any errors leading to different calls at SNP positions could potentially lead
688 to different ancestries. Assuming similar error rates across isolates, the relevant contributors to
689 infection spread could nevertheless be construed. When comparing maximum parsimony across
690 transmission trees, it should be kept in mind that the maximum genomic parsimony is likely present in
691 the model without any epidemiological limitations. As a consequence, the decision on the best fitting
692 model is based on a combination of epidemiological and biological knowledge and maximum
693 parsimony within this defined epidemiological and biological transmission framework. Actually, a
694 similar parsimony in models with epidemiological and biological constraints and the basic model
695 without these constraints is an outcome that heavily favours the outcome of the model with the
696 constraints.

697 A distance threshold of 6 SNPs was used as a cut-off to rule in/out direct ancestry between
698 isolates. This approach was applied to several pathogens, including tuberculosis in humans for which
699 the most commonly employed cut-off is based on the finding that epidemiologically linked patients
700 were genetically linked by ≤ 5 SNPs, with an upper bound of 12 SNPs between any two linked
701 isolates (51). Even when this analysis was largely insensitive to the mutation rate, it needs to be kept
702 in mind that rates of evolution (and as a result also distance thresholds) may differ across lineages of
703 the same species (21).

704 **Conclusions**

705 Mixed infections of dairy cows with MAP appear to be common, and some strains are more
706 successful than others in terms of transmission. To the best of the authors' knowledge, the high level
707 of within-host MAP diversity observed in this study with up to 5 genotypes sampled from a cow has
708 not been previously reported in the literature. Cows infected with one or more strains remain
709 susceptible to infections with other MAP genotypes. Transmission studies are therefore expected to

710 benefit from strain-specific transmission parameterisation. To be able to observe the full range of
711 diversity in samples with heterogeneous MAP populations, methods for pathogen isolation are
712 needed which support detection and quantification of multiple genotypes. This work presents a
713 method for reconstructing “who infects whom” based on genomic data with greater epidemiological
714 and statistical support. Reconstructed infection chains confirmed high shedding and exposure to
715 shedders over longer time periods as risk factors for MAP transmission on the investigated dairy farm.
716 We believe that the method will be useful for further studies on the relevance of transmission routes
717 and role of individuals expressing distinct disease phenotypes in infection dynamics of endemic
718 disease.

719 WGS is invaluable in studying pathogen transmission, both with outbreaks and in endemic
720 settings. However, WGS is not a solution for low test sensitivity which leads to non-observed isolates.
721 In addition, especially for MAP-like pathogens the question remains, when does an animal become
722 infected and infectious? The method presented in this work is able to indicate where infection cycles
723 went undetected. This information can be used to adapt sampling to better capture underlying
724 transmission processes. Knowledge of pathogen biology and availability of precise longitudinal data
725 are crucial to maximise benefits of WGS and validly reconstruct infection chains.

726 **Acknowledgements**

727 The authors thank all staff and veterinarians of the participating farm for their invaluable input
728 and support in collecting samples and data. We also acknowledge the teams of collaborators who
729 performed all samplings on the farm and performed MAP cultures in the course of this large cohort
730 study. We also would like to acknowledge Mart de Jong for his valuable comments on analyses.

731 **References**

732 1. Marce C, Ezanno P, Weber M, Seegers H, Pfeiffer D, Fourichon C. Invited review: modeling
733 within-herd transmission of *Mycobacterium avium* subspecies *paratuberculosis* in dairy cattle: a
734 review. *J Dairy Sci*, 2010;93(10):4455–4470.

735 2. Behr MA. Paratuberculosis: Organism, Disease, Control. In: Behr DM, Collins MA, editors. Paratuberculosis: organism, disease, control. Cambridge, MA, USA: CABI; 2010. p. 40–49.

737 3. Windsor PA, Whittington RJ. Evidence for age susceptibility of cattle to Johne's disease. *Vet J*, 2010;184(1):37–44.

739 4. Van Roermund H, Bakker D, Willemsen P, De Jong M. Horizontal transmission of Mycobacterium avium subsp. paratuberculosis in cattle in an experimental setting: calves can 740 transmit the infection to other calves. *Vet Microbiol*, 2007;122(3-4):270–279.

742 5. Whittington RJ, Begg DJ, Silva K de, Plain KM, Purdie AC. Comparative immunological and 743 microbiological aspects of paratuberculosis as a model mycobacterial infection. *Vet Immunol Immunop*, 2012;148(1):29–47.

745 6. Mortier RA, Barkema HW, De Buck J. Susceptibility to and diagnosis of Mycobacterium avium 746 subspecies paratuberculosis infection in dairy calves: a review. *Prev vet med*, 2015;121(3-4):189–198.

748 7. Pradhan AK, Mitchell RM, Kramer AJ, Zurkowski MJ, Fyock TL, Whitlock RH, et al. Molecular 749 epidemiology of Mycobacterium avium subsp. paratuberculosis in a longitudinal study of three 750 dairy herds. *J Clin Microbiol*, 2011;49(3):893–901.

751 8. Schukken YH, Whitlock RH, Wolfgang D, Grohn Y, Beaver A, VanKessel J, et al. Longitudinal 752 data collection of Mycobacterium avium subspecies Paratuberculosis infections in dairy herds: 753 the value of precise field data. *Vet res*, 2015;46(1):65p.

754 9. SDBbio. Review of On-Farm Bovine Johne's Disease Management Strategies for Victorian 755 Cattle Herds. Final project report 2014 [Internet]. Herd Health, Scott Williams Consulting, SDB 756 Bio; 2014. Available from: <https://www.vff.org.au/>

757 10. Mitchell RM, Schukken Y, Koets A, Weber M, Bakker D, Stabel J, et al. Differences in 758 intermittent and continuous fecal shedding patterns between natural and experimental 759 Mycobacterium avium subspecies paratuberculosis infections in cattle. *Vet res*, 2015;46(1):66.

760 11. Whitlock R, Sweeney R, Fyock T, Smith J. MAP Super-shedders: another factor in the control 761 of Johne's disease. In: SS N, editor. 8th International Colloquium on Paratuberculosis. The 762 Royal Veterinary and Agricultural University Copenhagen, 2005. p. 42.

763 12. Benedictus A, Mitchell R, Linde-Widmann M, Sweeney R, Fyock T, Schukken Y, et al. 764 Transmission parameters of Mycobacterium avium subspecies paratuberculosis infections in a 765 dairy herd going through a control program. *Prev vet med*, 2008;83(3-4):215–227.

766 13. Li L, Katani R, Schilling M, Kapur V. Molecular epidemiology of Mycobacterium avium subsp. 767 paratuberculosis on dairy farms. *Annu Rev Anim Biosci*, 2016;4:155–176.

768 14. Mitchell R, Whitlock R, Stehman S, Benedictus A, Chapagain P, Grohn Y, et al. Simulation 769 modeling to evaluate the persistence of Mycobacterium avium subsp. paratuberculosis (MAP) 770 on commercial dairy farms in the United States. *Prev vet med*, 2008;83(3-4):360–380.

771 15. Stevenson K. Genetic diversity of Mycobacterium avium subspecies paratuberculosis and the 772 influence of strain type on infection and pathogenesis: a review. *Vet res*, 2015;46(1):64.

773 16. Collins MT, Morgan IR. Simulation model of paratuberculosis control in a dairy herd. *Prev vet 774 med*, 1992;14(1-2):21–32.

775 17. Groenendaal H, Nielen M, Hesselink JW. Development of the Dutch Johne's disease control 776 program supported by a simulation model. *Prev vet med*, 2003;60(1):69–90.

777 18. Möbius P, Luyven G, Hotzel H, Köhler H. High genetic diversity among *Mycobacterium avium*
778 subsp. *paratuberculosis* strains from German cattle herds shown by combination of IS900
779 restriction fragment length polymorphism analysis and mycobacterial interspersed repetitive
780 unit-variable-number tandem-repeat typing. *J Clin Microbiol*, 2008;46(3):972–981.

781 19. Davidson FW, Ahlstrom C, De Buck J, Whitney HG, Tahlan K. Examination of *Mycobacterium*
782 *avium* subspecies *paratuberculosis* mixed genotype infections in dairy animals using a whole
783 genome sequencing approach. *PeerJ*. 2016;4.

784 20. Kao RR, Haydon DT, Lycett SJ, Murcia PR. Supersize me: how whole-genome sequencing and
785 big data are transforming epidemiology. *Trends microbiol*, 2014;22(5):282–291.

786 21. Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission.
787 *BMJ*, 2015;350:h1314.

788 22. Worby CJ, Lipsitch M, Hanage WP. Within-host bacterial diversity hinders accurate
789 reconstruction of transmission networks from genomic distance data. *PLoS Comput Biol*,
790 2014;10(3):e1003549.

791 23. Al-Mamun MA, Smith RL, Nigsch A, Schukken YH, Gröhn YT. A data-driven individual-based
792 model of infectious disease in livestock operation: A validation study for *paratuberculosis*. *PLoS*
793 *ONE*, 2018;13(12):e0203177.

794 24. USDA APHIS. vSNP. USDA APHIS Veterinary Services pipeline for *Mycobacterium*
795 tuberculosis complex and *Brucella* sp. Genotyping from high throughput sequence providing
796 SNP tables and phylogenetic trees with output to aid in SNP validation. [Internet]. USDA
797 APHIS; 2018. Available from: <https://github.com/USDA-VS/vSNP>

798 25. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform.
799 *Bioinformatics*, 2009;25(14):1754–1760.

800 26. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for
801 variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*,
802 2011;43(5):491.

803 27. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome
804 Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.
805 *Genome Res*, 2010;20(9):1297–1303.

806 28. Auwera GA Van der, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al.
807 From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices
808 pipeline. *Curr Protoc Bioinformatics*, 2013;43(1):11–10.

809 29. Richards VP, Nigsch A, Pavinski Bitar P, Sun Q, Stuber T, Ceres K, et al. Evolutionary genomic
810 and bacteria GWAS analysis of *Mycobacterium avium* subsp. *paratuberculosis* and dairy cattle
811 Johne's disease phenotypes. *Appl Environ Microbiol*, 2021; forthcoming.

812 30. Kumar S, Stecher G, Koichiro T. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0
813 for Bigger Datasets. *Mol Biol Evol*, 2016;33(7):1870–1874.

814 31. Jombart T, Eggo R, Dodd P, Balloux F. Reconstructing disease outbreaks from genetic data: a
815 graph approach. *Heredity*, 2011;106(2):383–390.

816 32. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data.
817 *Bioinformatics*. 2011;27(21):3070-3071.

818 33. R Development Core Team. R: A Language and Environment for Statistical Computing
819 [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2008. Available from:
820 <http://www.R-project.org>

821 34. Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, et al. Integrating
822 genetic and epidemiological data to determine transmission pathways of foot-and-mouth
823 disease virus. *Proc R Soc Lond B Biol Sci*, 2008;275(1637):887–895.

824 35. De Maio N, Wu C-H, Wilson DJ. SCOTTI: efficient reconstruction of transmission within
825 outbreaks with the structured coalescent. *PLoS Comput Biol*, 2016;12(9):e1005130.

826 36. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-
827 genome sequence data. *Mol Biol Evol*, 2014;31(7):1869–1879.

828 37. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of
829 disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol*,
830 2014;10(1):e1003457.

831 38. Lieberman TD, Michel J-B, Aingaran M, Potter-Bynoe G, Roux D, Davis MR, et al. Parallel
832 bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Genet*,
833 2011;43(12):1275–1280.

834 39. Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. *Proc
835 Natl Acad Sci U S A*, 2016;113(10):2690–2695.

836 40. Volz EM, Frost SD. Inferring the source of transmission with phylogenetic data. *PLoS Comput
837 Biol*, 2013;9(12):e1003397.

838 41. Aldrin M, Lyngstad T, Kristoffersen A, Storvik B, Borgaon Ø, Jansen P. Modelling the spread of
839 infectious salmon anaemia among salmon farms based on seaway distances between farms
840 and genetic relationships between infectious salmon anaemia virus isolates. *J R Soc Interface*,
841 2011;8(62):1346–1356.

842 42. Hall M, Woolhouse M, Rambaut A. Epidemic reconstruction in a phylogenetics framework:
843 transmission trees as partitions of the node set. *PLoS Comput Biol*, 2015;11(12):e1004613.

844 43. Mollentze N, Nel LH, Townsend S, Le Roux K, Hampson K, Haydon DT, et al. A Bayesian
845 approach for inferring the dynamics of partially observed endemic infectious diseases from
846 space-time-genetic data. *Proc R Soc Lond B Biol Sci*, 2014;281(1782):20133251.

847 44. Morelli MJ, Thébaud G, Chadoeuf J, King DP, Haydon DT, Soubeyrand S. A Bayesian
848 inference framework to reconstruct transmission trees using epidemiological and genetic data.
849 *PLoS Comput Biol*, 2012;8(11):e1002768.

850 45. Ypma RJ, Ballegooijen WM van, Wallinga J. Relating phylogenetic trees to transmission trees
851 of infectious disease outbreaks. *Genetics*, 2013;195(3):1055–1062.

852 46. Ypma RJ, Bataille A, Stegeman A, Koch G, Wallinga J, Van Ballegooijen WM. Unravelling
853 transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc
854 R Soc Lond B Biol Sci*, 2012;279(1728):444–450.

855 47. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using
856 timing of symptoms, pathogen genomes and contact data. *PLoS Comput Biol*,
857 2019;15(3):e1006930.

858 48. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences
859 informative of transmission events? *PLoS Pathog*, 2018;14(2):e1006885.

860 49. Leavitt SV, Lee RS, Sebastiani P, Horsburgh CR, Jenkins HE, White LF. Estimating the relative
861 probability of direct transmission between infectious disease patients. *Int J Epidemiol*,
862 2020;49(3):764–775.

863 50. Koets AP, Eda S, Sreevatsan S. The within host dynamics of *Mycobacterium avium* ssp.
864 paratuberculosis infection in cattle: where time and place matter. *Vet res*, 2015;46(1):61.

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

868

869 **Supporting information**

870 Supporting information related to this article can be found, in the online version, at doi: ###

871 **S1 file. SNP alignment of 150 MAP sequences with 1,472 SNPs of the core genome.**

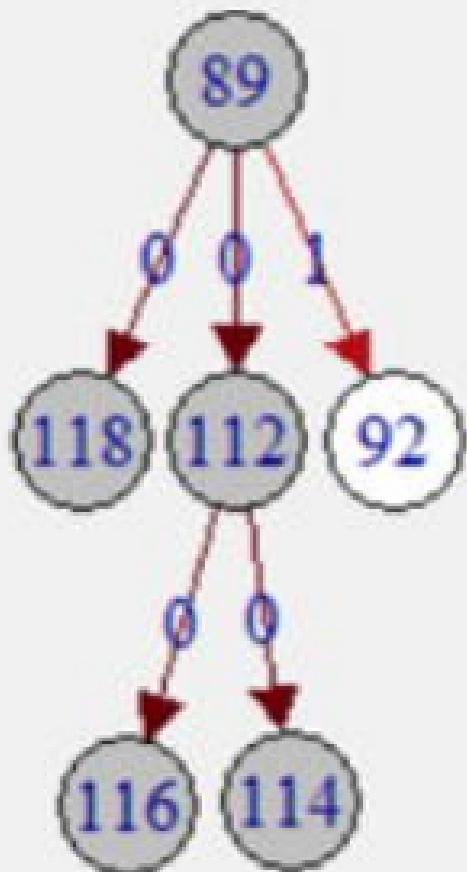
872 **S2 file. R code of analysis.**

873 **S1 Fig. Reconstructed transmission tree of three scenarios (n = 128 isolates).** (A) [shed_Basic],
874 (B) [shed_E], (C) [shed_S]. Isolates sampled from the same cow are labelled with successive
875 numbers and are shown in vertices of same colour and outline. White vertices represent cows with
876 only one isolate. Dark green vertices (labelled 99–120) represent environmental samples. Edge labels
877 and edge colour indicate number of SNPs difference between ancestor and descendant.

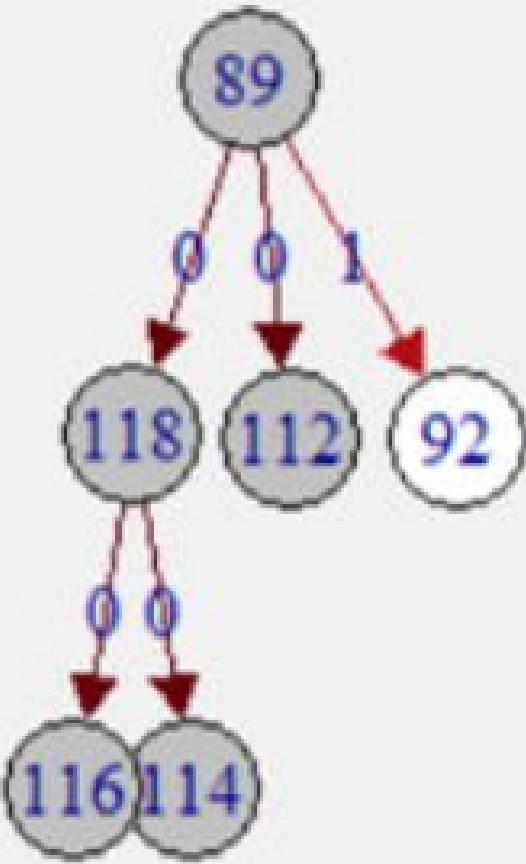
878 **S1 Table. Statistical support for inferred ancestries, by scenario.** Statistical support for each
879 inferred ancestry is expressed as a p-value, calculated based on maximum likelihood. Summary
880 statistics are presented at the bottom of the table. Descendants with no ancestors can be identified as
881 singleton isolates in the figures of the reconstructed transmission trees Fig. 5 and S2 Fig. Singleton
882 isolates have >6 SNP difference to any other isolate).

883 **S2 Fig. Numbers of recipients produced by individual cows.** Boxplots with numbers of recipients
884 produced by individual cows, by disease phenotype and scenario. **(A)** shedding level (0 – always
885 faecal culture negative, 1 - low, 2 – high), **(B)** age at first shedding (0 - ≤3 years, 1 - >3 years, 2 - ante
886 mortem negative), **(C)** infection progress (0 - ante mortem negative, 1 - non-progressor, 2 -

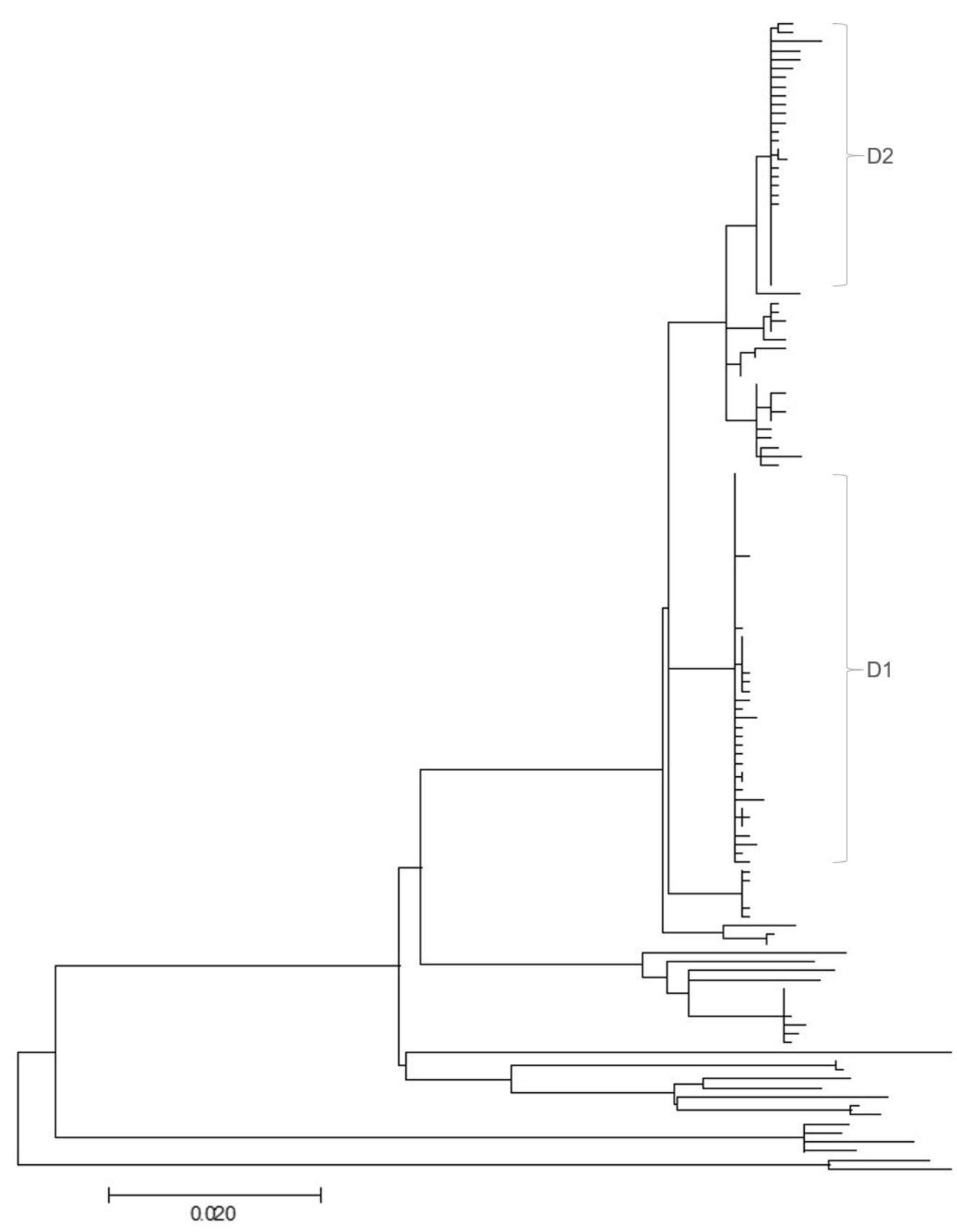
887 progressor), **(D)** serostatus (0 - ELISA-negative, 1 - ELISA-positive). Scenarios (from left-most to
888 right-most column): [birth_Basic], [birth_E], [birth_S], [shed_Basic], [shed_E] and [shed_S].

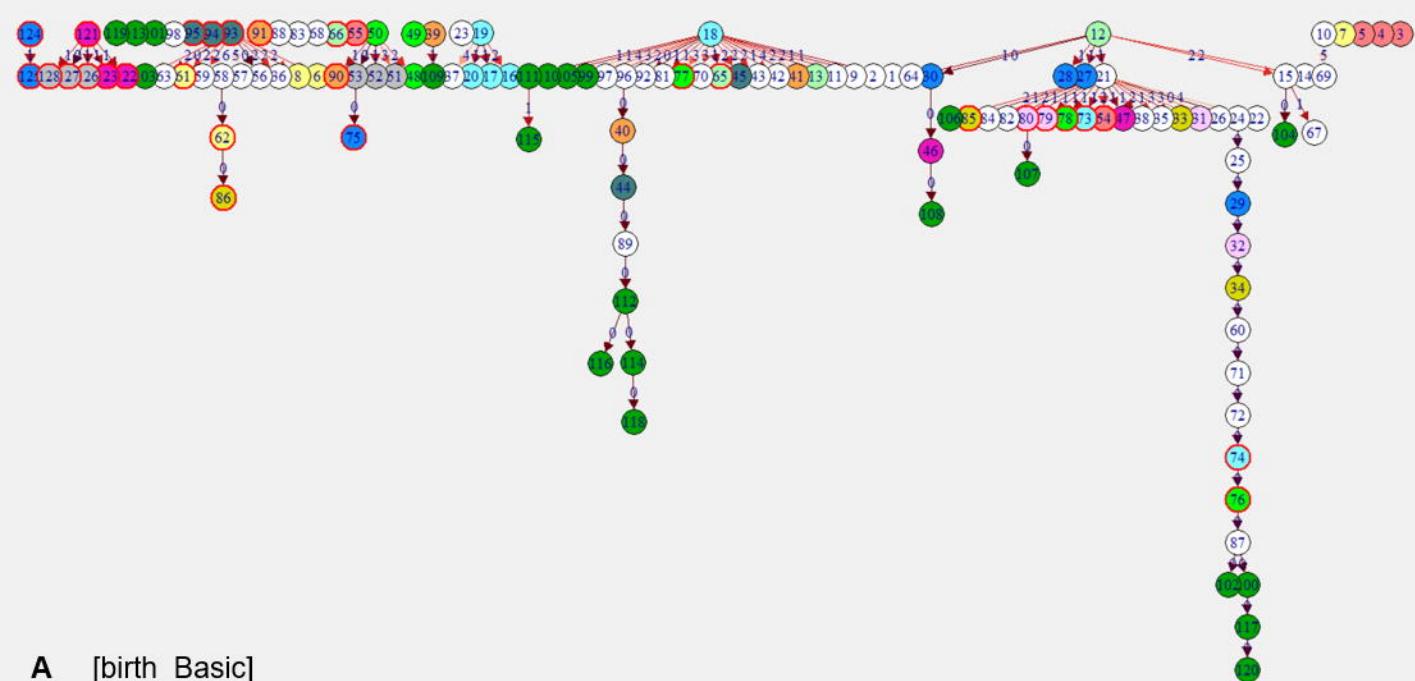


A

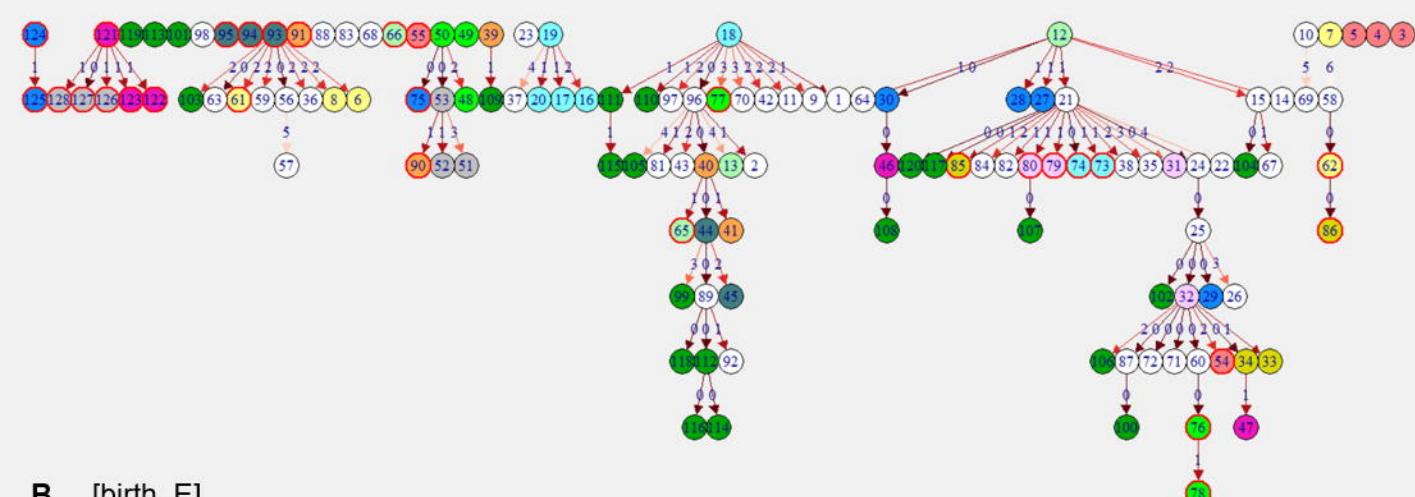


B

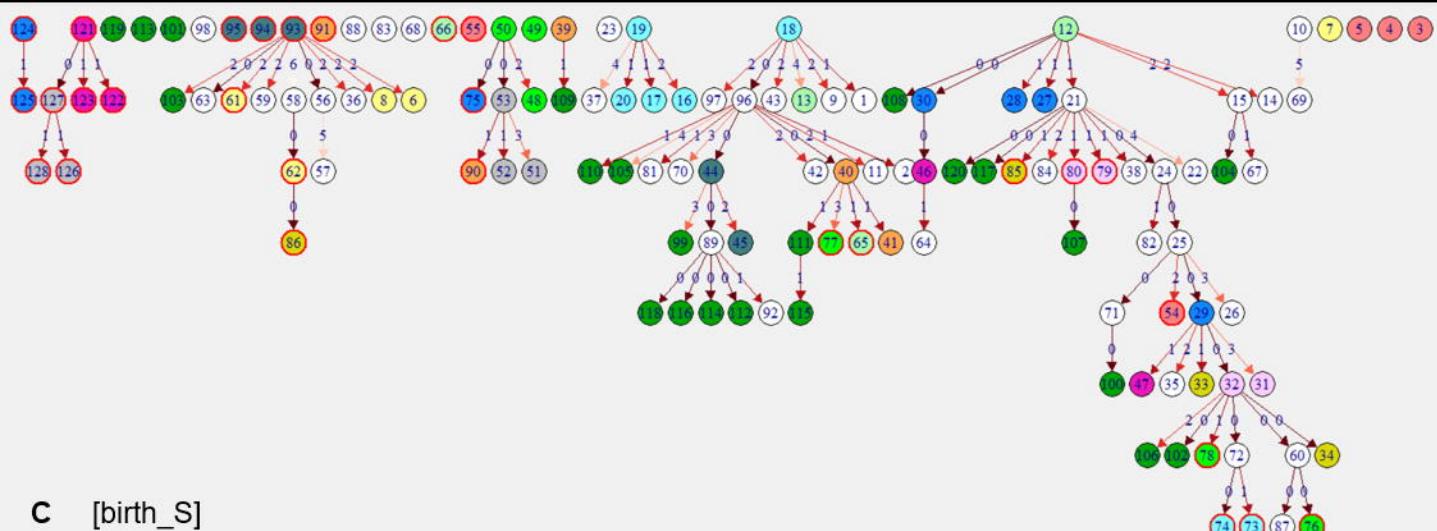




A [birth_Basic]

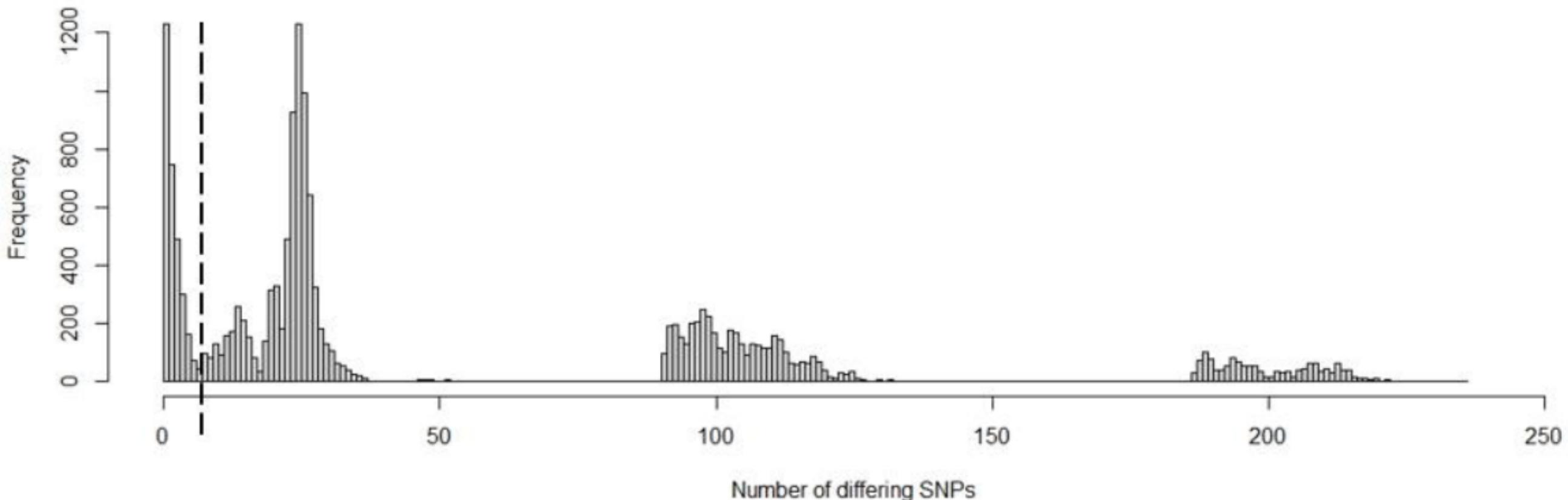


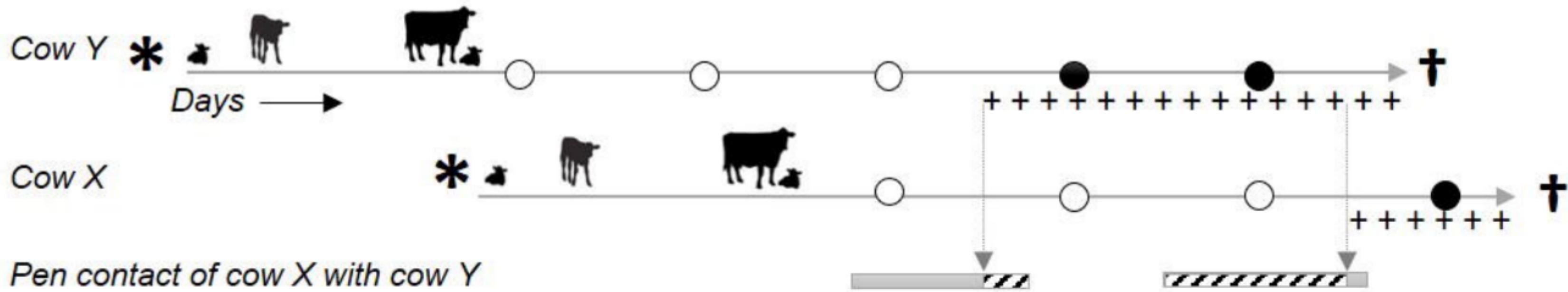
B [birth_E]



C [birth_S]

Distribution of pairwise genomic distances





* Birth † Death ○ MAP negative sample ● MAP positive sample + + Infectious period
 ■ Pen contact outside exposure time ☐ [E]: Pen contact during exposure time