



HAL
open science

Outbreak reconstruction with a slowly evolving multi-host pathogen: a comparative study of three existing methods on *Mycobacterium bovis* outbreaks

Hélène Duault, Benoit Durand, Laetitia Canini

► To cite this version:

Hélène Duault, Benoit Durand, Laetitia Canini. Outbreak reconstruction with a slowly evolving multi-host pathogen: a comparative study of three existing methods on *Mycobacterium bovis* outbreaks. 2024. hal-04369132

HAL Id: hal-04369132

<https://hal.science/hal-04369132>

Preprint submitted on 2 Jan 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 “Outbreak reconstruction with a slowly evolving multi-host
2 pathogen: a comparative study of three existing methods on
3 *Mycobacterium bovis* outbreaks.”

4 H el ene Duault^{1,2}, Benoit Durand¹ and Laetitia Canini^{1*}

5 ¹ Paris-Est University, Epidemiology Unit, Laboratory for Animal Health, Anses, Maisons-
6 Alfort, France

7

8 ² Universit e Paris-Saclay, Facult e de m edecine, Le Kremlin-Bic etre, France

9 * Corresponding author

10 Email: laetitia.canini@anses.fr (L.C.)

11 **Abstract**

12 In a multi-host system, understanding host-species contribution to transmission is key
13 to appropriately targeting control and preventive measures. Outbreak reconstruction methods
14 aiming to identify who-infected-whom by combining epidemiological and genetic data could
15 contribute to achieving this goal. However, the majority of these methods remain untested on
16 realistic simulated multi-host data. *Mycobacterium bovis* is a slowly evolving multi-host
17 pathogen and previous studies on outbreaks involving both cattle and wildlife have identified
18 observation biases. Indeed, contrary to cattle, sampling wildlife is difficult. The aim of our
19 study was to evaluate and compare the performances of three existing outbreak reconstruction
20 methods (seqTrack, *outbreaker2* and *TransPhylo*) on *M. bovis* multi-host data simulated with
21 and without biases.

22 Extending an existing transmission model, we simulated 30 bTB outbreaks involving
23 cattle, badgers and wild boars and defined six sampling schemes mimicking observation biases.
24 We estimated general and specific to multi-host systems epidemiological indicators. We tested
25 four alternative transmission scenarios changing the mutation rate or the composition of the
26 epidemiological system. The reconstruction of who-infected-whom was sensitive to the
27 mutation rate and seqTrack reconstructed prolific super-spreaders. *TransPhylo* and *outbreaker2*
28 poorly estimated the contribution of each host-species and could not reconstruct the presence
29 of a dead-end epidemiological host. However, the host-species of cattle (but not badger) index
30 cases was correctly reconstructed by seqTrack and *outbreaker2*. These two specific indicators
31 improved when considering an observation bias.

32 We found an overall poor performance for the three methods on simulated biased and
33 unbiased bTB data. This seemed partly attributable to the low evolutionary rate characteristic
34 of *M. bovis* leading to insufficient genetic information, but also to the complexity of the

35 simulated multi-host system. This study highlights the importance of an integrated approach
36 and the need to develop new outbreak reconstruction methods adapted to complex
37 epidemiological systems and tested on realistic multi-host data.

38 **Author summary**

39 Some pathogens like the one responsible for bovine tuberculosis can infect multiple
40 species. Identifying which species transmitted and to which other species in such an outbreak
41 presents a unique challenge, especially when difficult to observe wildlife species are concerned.
42 One way to tackle this issue would be to reconstruct who-infected-whom in an outbreak and
43 then identify the role each species played. However, methods that enable this type of
44 reconstruction have not been tested in the context of transmission between unevenly observed
45 species. Moreover, the pathogen responsible for bovine tuberculosis evolves slowly, which
46 further complicates the reconstruction of who-infected-whom. We thus simulated realistic and
47 complex bovine tuberculosis outbreaks on which we tested three widely used methods. We
48 found poor performances for all three tested methods, which highlights the need to develop new
49 methods adapted to outbreaks involving multiple species. Our results also underline the need to
50 combine multiple types of methods and data sources in addition to the reconstruction of who-
51 infected-whom, such as the reconstruction of phylogenetic trees or identifying possible
52 infectious contacts through investigations, when studying an outbreak.

53 **Introduction**

54 Over 60% of pathogens can infect more than one host-species [1,2]. This possible
55 contribution of multiple host-species to transmission dynamics complicates disease control and
56 surveillance for these multi-host pathogens, especially when one of the host-species to consider
57 is a free-ranging wildlife species. Indeed, quantifying contribution to transmission in order to
58 select appropriate control measures as well as the implementation of said measures can be

59 challenged by the lack of accurate estimations of wildlife population size, the impossibility to
60 restrain the entire wildlife population and the difficulty to prevent interactions between host-
61 species [3]. Multi-host pathogens can have important consequences on human health (*e.g.*
62 zoonotic diseases endemic in wildlife [4]), biodiversity (*e.g.* canine distemper in lions, *Panthera*
63 *leo*, in the Serengeti national park [5]) and animal trade economy (*e.g.* foot-and-mouth disease
64 and avian influenza [6]).

65 A prime example of a multi-host pathogen, for which the contribution of wildlife species
66 needs to be considered, is *Mycobacterium bovis*, the most frequent etiological agent of bovine
67 tuberculosis (bTB). Indeed, while *M. bovis* mainly affects cattle, which have been the target of
68 bTB control programs in the European Union since 1964 (EU directive 64/432/EEC), other
69 domestic and wildlife host-species can also be infected [7]. Furthermore, wildlife species have
70 even been implicated around the world as bTB reservoirs, *e.g.* badgers (*Meles meles*) in the
71 United Kingdom [8], wild boars (*Sus scrofa*) in Spain [9] and brush-tailed possums
72 (*Trichosurus vulpecula*) in New Zealand (10). In France, infected wildlife presenting the same
73 genotypes as nearby infected cattle have been reported by the wildlife surveillance program
74 since its implementation in 2012 [11], which suggests bTB transmission between wildlife and
75 cattle and therefore, the presence of bTB multi-host systems.

76 Studies have aimed to reconstruct phylogenetic trees from *M. bovis* whole genome
77 sequences, present in cattle and wildlife, in order to better understand transmission within these
78 multi-host systems [12–15]. In a phylogenetic tree, internal nodes correspond to hypothetical
79 common ancestors and, using Bayesian methods, the ancestral state (*e.g.* host-species [16,17]
80 or geographical location [18,19]) of these internal nodes can be estimated. These Bayesian
81 methods can therefore reconstruct the host-species of the most recent common ancestor of all
82 sampled sequences [20] as well as transitions between species or groups of individuals over
83 time [15,17], but not transmission events at an individual level. Phylogenetic trees thus differ

84 from transmission trees, in which each node represents an infected host and these infected hosts
85 are linked by directed edges representing transmission events [21]. Such a reconstruction of
86 who-infected-whom in the outbreak makes it possible to estimate transmission parameters
87 specific to each host-species (such as the number of transmission events due to an individual of
88 a particular host-species), and thus sheds more light on the transmission dynamics within the
89 studied multi-host system.

90 In principle, outbreak (here meaning transmission tree) reconstruction could be based
91 solely on epidemiological data obtained via contact tracing methods (*e.g.* [22]); however data
92 collected are not always reliable nor detailed enough to enable accurate reconstruction [23].
93 Therefore, some outbreak reconstruction methods have combined both genomic and
94 epidemiological data in transmission tree inference [24–29]. These outbreak reconstruction
95 methods can be divided into two categories according to how genomic data is treated [30], those
96 that consider a link between phylogenetic and transmission trees (generally by annotating
97 branches or internal nodes with infected hosts) [26,28,31,32] and those that solely consider
98 genetic distances [21,25,33]. While some outbreak reconstruction methods were developed to
99 study pathogen transmission within a specific multi-host system (*e.g.* foot-and-mouth disease
100 [34]), most were developed using the example of a single-host system, *e.g.* slowly evolving *M.*
101 *tuberculosis* [26,31], and more rapidly evolving pathogens like methicillin-resistant
102 *Staphylococcus aureus* [26,33] or SARS-CoV-1 [25] in a human population. However, the
103 development of outbreak reconstruction methods on single-host systems does not preclude them
104 from yielding insightful results in multi-host systems; for instance Willgert *et al.* recently
105 reconstructed the transmission history of SARS-CoV-2 in a human-deer system in Iowa (USA)
106 [35]. In a multi-host system, other than correctly reconstructing transmission events between
107 individuals and estimating outbreak size (general epidemiological indicators), we expect
108 outbreak reconstruction methods to allow accurate estimation of host-species contribution to

109 the outbreak and to identify the host-species of the index case (specific multi-host
110 epidemiological indicators).

111 While some outbreak reconstruction methods assume that all cases are known and
112 sampled [21,28,36], others account for the presence of unsampled cases by either allowing the
113 annotation of unsampled hosts in the phylogenetic tree [31] or the presence of intermediary
114 unsampled hosts between two sampled hosts [24,25]. When not all cases are sampled in the
115 outbreak, there exists a difference between the actual outbreak and the transmission tree these
116 methods can aim to reconstruct from sampled sequences. Indeed, even if the outbreak
117 reconstruction method accounts for the presence of unsampled hosts [25,31], these hosts can
118 only be inferred if they have descendant sampled hosts [35] and the transmission tree that can
119 be reconstructed is therefore a subtree induced by the sampling process.

120 The sampling process in a multi-host system that implicates a free-ranging wildlife
121 species can also result in incomplete or even biased data, when observation efforts differ
122 between host-species. For instance, *M. bovis* wildlife surveillance in France was implemented
123 later than cattle surveillance (2012 vs. 1954) and only investigates bTB infection in badgers,
124 boars, red deer (*Cervus elaphus*) and roe deer (*Capreolus capreolus*) [11]. However,
125 estimations of bTB infection rates in red foxes (*Vulpes vulpes*) have recently been investigated
126 in France and yielded similar results to those found in badgers and wild boars [37]. These
127 sampling biases between host-species could have an important impact on outbreak
128 reconstruction.

129 Our aim was to evaluate and compare performances of existing outbreak reconstruction
130 methods on bTB outbreaks in a multi-host system and study whether these performances were
131 affected by sampling biases. Therefore, we simulated bTB transmission within a multi-host
132 system situated in a previously studied area in the South-West of France. In this area, bTB
133 surveillance has reported *M. bovis* circulation in cattle, badgers and wild boars [38]. Multiple

134 sampling schemes were implemented to reflect the late implementation of wildlife surveillance
135 (temporal bias) and the fact that not all host-species are surveilled (species bias). In order to
136 evaluate the quality of reconstructed transmission trees, we calculated general as well as
137 specific multi-host epidemiological indicators.

138 **Materials and methods**

139 **1. Reference transmission trees**

140 **1.1 Transmission model**

141 We extended an existing model that simulated bTB transmission trees, for the 11
142 genotypes identified, in a badger-cattle system present in a study area in the South-West of
143 France, from January 2007 to January 2020 [39]. We narrowed our study to one of the two
144 genotypes of *M. bovis*, which were isolated in both wildlife and cattle within our study region.
145 Since infected wild boars have also been detected in this study area [40,41] and our aim was to
146 study a complex multi-host system, we added a wild boar meta-population to the modeled
147 epidemiological system (see details in S1 Appendix). Similarly to the badger population, wild
148 boars could either be susceptible (S) or infected (I) while cattle had an additional latent state
149 (E), when animals could be detected infected but could not transmit the pathogen [39].

150 Moreover, transmission trees simulated with the original model considered cattle farms
151 and badger social groups as epidemiological units whereas we aimed to reconstruct individual
152 transmission links. Therefore, we extended the model to randomly select infected animals
153 within these groups according to the SEI/SI system dynamics and thus, simulated animal-to-
154 animal transmission. The resulting transmission trees are termed below *reference transmission*
155 *trees* (terms written in italic are defined in Table 1).

156 1.2 Reference set of cases

157 We chose cattle as index cases and bTB spread in the multi-host system was simulated
158 during 13 years. We generated 30 reference transmission trees, in order to investigate various
159 simulated outbreaks while limiting the computational time. These 30 trees had to include less
160 than 500 infected hosts in total, for computational reasons, and at least 15 infected hosts from
161 each host-species, in order to be able to implement sampling schemes. A reference transmission
162 tree corresponded to a list of six variables: identification (id) of infector, id of infected, host-
163 species of infector, host-species of infected, date of infection and date of death.

164 We simulated genetic sequences along the reference trees according to a Hasegawa-
165 Kishino-Yano (HKY) substitution model (with transition/transversion ratio parameter, κ) [42],
166 since this substitution model was previously used to study *M. bovis* phylogenies [12–14], as
167 well as a fixed mutation rate (μ). We chose μ equal to 0.0024 substitutions per site per year and
168 κ equal to 5.9. Indeed, these values had been previously estimated on 167 *M. bovis* sequences
169 (171 SNPs in length) isolated in cattle and wildlife from this study area [12].

170 At $t = 0$, we considered that the index case was infected by a single sequence randomly
171 selected from the 167 sequences isolated in our study area [12]. Our substitution algorithm was
172 based on the Gillespie approach [43] implemented in the *phastSim* package [44] (Fig 1). Taking
173 into account the low genetic diversity observed in *M. bovis* sequences from the same region,
174 we assumed no within-host diversity by considering a single lineage per host but we allowed
175 within-host mutation.

176 We simulated sequences until February 2020. Then, the last simulated sequence was
177 recorded for each host, which corresponded to either the sequence present at the time of removal
178 or in February 2020, for infected hosts not yet removed at the end of the simulation. For each
179 reference transmission tree, we thus obtained a *reference set of cases* (Table 1), meaning a list

180 of four variables: id of infected, host-species of infected, date of death (or February 2020 if host
181 still alive) and sampled sequence.

182 **Fig 1. Sequence simulation procedure in two infected hosts A and B.** Host A, represented by the
183 grey rectangle on the left (infected at $T_{infection_A}$), transmitted the pathogen to host B at $T_{infection_B}$. This
184 transmission event is represented by the thick black arrow. Hosts were removed (represented by the
185 cross) respectively at $T_{removal_A}$ and $T_{removal_B}$. If the mutation time was inferior to the host removal time
186 (which was the case for $T_{mutation_A1}$ and $T_{mutation_A2}$ in host A), we then selected the nucleotide to mutate
187 (the 3rd nucleotide for the first mutation and the 2nd nucleotide for the second mutation in host A, shown
188 in white) and changed it according to a substitution model. If the mutation time was superior to the
189 removal time of the host (see host B), the sequence did not change until host removal and this sequence
190 was then the one sampled from the host.

191 2. Sampling schemes and reconstructed transmission trees

192 2.1 Sampling schemes

193 We first considered the hypothetical situation where all infected hosts are observed
194 (reference *sampling scheme*, Table 1), which corresponds to the reference set of cases. Then,
195 we simulated five sampling schemes that mimicked observation biases in bTB epidemiological
196 data, while also sampling all infected hosts unaffected by the scheme (even those not yet
197 removed at the end of the simulation). In scheme T (for “temporal bias”), the late
198 implementation of wildlife surveillance in the study region was simulated and we only
199 considered wildlife cases after 2012. Moreover, the fact that not all host-species are surveilled
200 was simulated in schemes S (for “species bias”): either wild boar cases were not considered,
201 scheme S_W , or badger cases, in scheme S_B . Finally, in scheme T+ S_W (or T+ S_B), we disregarded
202 cases before 2012 for the remaining wildlife species (respectively badgers and wild boars).

203 We thus simulated for each reference transmission tree one *biased set of cases* (Table
204 1) for each sampling scheme (T, S_B , S_W , T+ S_B , T+ S_W), that contained the same variables as the
205 reference set of cases. With 30 reference trees for each sampling scheme, we thus obtained a
206 total of $30*6 = 180$ sets of cases. For each of these sets of cases, we extracted from the reference
207 transmission tree, the *reconstructible outbreak* (Table 1), which is the subtree containing only
208 the cases that were sampled and their ancestors.

209 2.2 Transmission tree reconstruction

210 From our review on outbreak reconstruction methods [30], we identified three methods
211 (*seqTrack*, *outbreaker2* and *TransPhylo*) that were available in an R package and that needed
212 only sampling and/or removal times. In *seqTrack* and *outbreaker2*, transmission is estimated
213 based on pairwise genetic distances, while in *TransPhylo*, a link is established between
214 phylogenetic and transmission trees [30].

215 ○ *seqTrack*

216 Using Edmonds' algorithm, *seqTrack* computes the transmission tree in which the total
217 genetic distance between nodes is minimal, assuming that infectors are sampled before the host
218 they infected [21]. In order to use this method, we estimated pairwise genetic distances by using
219 the *dist.dna* function (*ape* R package v.5.4-1 [45]) with the F84 substitution model since it
220 closely resembles the HKY model [42]. *seqTrack* [21] is a function available in the *adegenet* R
221 package [46,47]. The format of the tree reconstructed by *seqTrack* was a table with five columns
222 corresponding to the following variables: *id* (indices of infected hosts), *ances* (indices of
223 infectors), *weight* (number of mutations separating infected hosts from their infectors), *date*
224 (sampling date of the infected host), *ances.date* (sampling date of their infector).

225 ○ *outbreaker2*

226 *outbreaker2* is a Bayesian method that considers four likelihoods: genetic, temporal,
227 reporting and contact [25]. In this method, probability of transmission is inferred from known
228 generation time (time between the infection of a case and the time of transmission from that
229 case to secondary cases) and sampling interval (time from infection to sampling) distributions.
230 Here, we assumed that generation time and sampling interval nonparametric distributions could
231 be obtained without bias by estimating them from the reference trees, which contained every
232 infected host, timed transmission event between hosts and host sampling time. We selected a
233 chain length of 100,000 iterations, a sampling frequency of 1 in 50 and a burn-in period of 10%

234 (for details on priors used and other arguments see S1 Appendix). We graphically checked for
235 convergence and independence of sampling (Effective Sample Size (ESS) above 200 for each
236 parameter), after estimation using the *coda* R package v.0.19-4 [48]. When the ESS were lower
237 than 200, we ran an additional 100,000 iterations and then checked the ESS again. This step
238 was repeated until every ESS was above 200.

239 Then, we built the consensus tree, as suggested by the authors, computing the most
240 frequent infector for each infected host in the posterior trees as well as the support (posterior
241 probability) for each transmission event. By construction, cycles can be present in this
242 consensus tree (which then becomes a directed graph), meaning that infected hosts can be both
243 the ancestors and the descendants of other infected hosts. Moreover, since this method considers
244 a reporting likelihood, the probability of sampling an infected host is estimated and unsampled
245 hosts are indirectly represented in the consensus tree, as a number of generations separating two
246 sampled hosts.

247 The format of the consensus tree reconstructed by *outbreaker2* was a table with five
248 columns corresponding to the following variables: from (indices of infectors), to (indices of
249 infected hosts), support (transmission probability), time (estimated time of transmission), date
250 (sampling date of the infected host) and generations (number of intermediary hosts + 1).

251 ○ *TransPhylo*

252 *TransPhylo*, another Bayesian method, affects infected hosts along branches in a
253 previously reconstructed phylogenetic tree [31] (for details on phylogenetic reconstruction see
254 S1 Appendix). We assumed that the generation time and sampling interval followed a Gamma
255 distribution and that the mean and standard deviation could be obtained without bias by
256 estimating them from the reference trees using the *epitrix* R package v.0.2.2 [49]. We selected
257 a number of iterations of 500,000, a sampling frequency of 1 in 50 and a burn-in period of 20%
258 (for details on priors used and other arguments see S1 Appendix). We used the same method as

259 with *outbreaker2* to check for convergence and independence of sampling, however we
260 considered a lower threshold for the ESS, 100 for each parameter as suggested by the authors
261 [50]. When the ESS were lower than 100, we ran an additional 500,000 iterations and then
262 checked the ESS again. This step was repeated until convergence and independence of sampling
263 parameters were satisfied or the number of iterations reached 2,500,000, we then discarded the
264 reference trees for which convergence was not obtained in every sampling scheme.

265 Then, as described by *Didelot et al.* [50], we computed the medoid transmission tree
266 (the transmission tree that is the least different from all other posterior trees according to a
267 distance metric defined by Kendall *et al.* [51]). This method accounts for the presence of
268 unsampled hosts when affecting hosts to branches in the phylogenetic tree, and unsampled hosts
269 are explicitly represented as nodes in the medoid transmission tree. This means that in the
270 medoid tree, contrary to the consensus tree in *outbreaker2*, unknown infected hosts can be
271 responsible for more than one transmission event. As in *outbreaker2*, *TransPhylo* estimates a
272 sampling probability.

273 The format of the medoid tree reconstructed by *TransPhylo* was a table with four
274 columns corresponding to the following variables: tinfecion (estimated time of infection),
275 tremoved (estimated time of removal of the infected host), infector_id (id of infector),
276 infected_id (id of infected).

277 From the sampled posterior trees, we also computed the n-by-n matrix of transmission
278 probability using the `computeMatWIW` function implemented in *TransPhylo*, where n is the
279 number of sampled infected hosts. Then, we identified for each infected host, its most likely
280 infector corresponding to the infector with the highest probability in the matrix of transmission
281 probabilities. If this probability was zero, we considered the most likely infector of the infected
282 host to be unknown. Note that this method of summarizing posterior trees can lead to the

283 presence of cycles, as in *outbreaker2*, and since time of infection is not estimated, no index case
284 can be inferred.

285 **3. Genetic information and epidemiological indicators**

286 **3.1 Genetic information**

287 To understand the impact of the sequence simulation model on outbreak reconstruction
288 and facilitate comparison with other works, we first quantified the genetic diversity present in
289 each simulated set of cases. We estimated the proportion of unique sequences in every set of
290 cases obtained with the reference sampling scheme as well as the mean transmission
291 divergence. Transmission divergence was defined in Campbell *et al.*'s work [52] as the number
292 of SNPs separating known transmission pairs, we used reference transmission trees to identify
293 transmission pairs and calculated the mean number of SNPs separating these transmission pairs
294 for every reference tree.

295 **3.2 Epidemiological indicators**

296 *TransPhylo* had two different outputs (the medoid tree and transmission probability
297 matrix). We used the transmission probability matrix when evaluating the method's accuracy
298 and the medoid tree for all other indicators.

- 299 ○ Accuracy

300 In order to evaluate the performance of all three reconstruction methods, we first
301 determined the correct transmission events that could be reconstructed between individuals
302 from each simulated set of cases. For the reference set of cases, the correct transmission events
303 were those present in the reference trees. However, for each biased set of cases, we considered
304 that the correct transmission events were those that connected observed cases to each other,
305 bypassing intermediary unobserved cases. For instance, the chain of transmission Sampled
306 subject #1 → Unobserved subject #2 → Sampled subject #3 would become Sampled subject

307 #1 → Sampled subject #3. For all three methods, we estimated whether reconstructed infector-
308 infected pairs (meaning every “id”-“ances” for seqTrack, “from”-“to” for *outbreaker2* and
309 “infector_id”-“infected_id” for the transmission matrix estimated from *TransPhylo*) were one
310 of the correct transmission events or not.

311 ○ Presence of super-spreaders

312 For all three methods, we considered super-spreaders to be present in a reconstructed
313 tree when less than 10% of infected hosts were responsible for over 80% of transmission events.
314 Moreover, when super-spreaders were present in a reconstructed tree, we identified the
315 maximum number of transmission events a single super-spreader could be responsible for as
316 well as the host-species of said super-spreader.

317 ○ Host-species of the index case

318 We evaluated the ability of all three methods to reconstruct the correct host-species of
319 the index case (*i.e.* cattle). Contrary to the *TransPhylo* medoid trees, in which identifying the
320 index case is straightforward (“infected_id” with the earliest “infection”), the presence of
321 cycles in *outbreaker2* and the multiples index cases possible in seqTrack complicated the
322 identification of the index case. For seqTrack, we considered the most frequent host-species
323 from the reconstructed index cases (“id” for whom the “ances” is unknown). For *outbreaker2*,
324 we considered the host-species of the index case to be the most frequent host-species among
325 cases infected at the earliest date (“to” with the earliest “time”).

326 ○ Outbreak size

327 We evaluated the ability of *outbreaker2* and *TransPhylo* to estimate the size of the
328 outbreak (seqTrack does not estimate outbreak size and was thus excluded for this indicator).
329 The simulated outbreak size was the number of infected hosts present in each reference tree.
330 We calculated the corresponding estimate by dividing the number of sampled hosts in each

331 reconstructed tree with the median of the sampling proportion provided by *outbreaker2* and
332 *TransPhylo*. In addition, we tested if the results for this indicator differed depending on whether
333 we were considering the reconstructible outbreak or the reference tree. Therefore, we also
334 calculated the number of infected hosts present in the reconstructible outbreak, and compared
335 it with the number of hosts (sampled and unsampled) present in the trees reconstructed by
336 *outbreaker2* and *TransPhylo*.

337 ○ Host-species contribution

338 Considering the importance of identifying the host-species that contributed the most to
339 transmission in a multi-host system, we evaluated the ability of *outbreaker2* and *TransPhylo* to
340 reconstruct the number of transmission events due to each host-species. Similarly to the
341 outbreak size, seqTrack was also excluded. The number of transmission events due to each
342 host-species was first calculated in the reference trees. As for the outbreak size, we calculated
343 the corresponding estimate by dividing the number of transmission events between sampled
344 hosts in each reconstructed tree with the median of the sampling proportion provided by
345 *outbreaker2* and *TransPhylo*. We then calculated the number of transmission events due to each
346 host-species in the reconstructible outbreak. This number was compared to the number of all
347 transmission events (to sampled and unsampled infected hosts) due to each host-species present
348 in the reconstructed trees.

349 ○ Statistical analysis

350 For the outbreak size and host-species contribution estimates, we obtained a credible
351 interval using the bounds of the 95%HPD (High Posterior Density) interval. For each
352 reconstructed tree, we evaluated whether the credible interval contained the simulated outbreak
353 size or number of transmission events due to each host-species. For all epidemiological
354 indicators except the presence of super-spreaders, we tested the effect on the indicator value of

355 the outbreak reconstruction method as well as its interaction with the effect of sampling scheme.
356 In order to account for the non-independence of reconstructed trees (six sets of cases are
357 constructed from the same reference tree), we fit mixed-effects models, using the id of the
358 reference tree as a random effect. For accuracy and index case, we selected a binomial
359 distribution and the probability of either reconstructing a correct transmission event or the
360 correct host-species for the index case was set as the outcome. Due to the overdispersion present
361 in the estimates of outbreak size and number of transmission events, we considered for both
362 indicators a negative binomial distribution. Since, for outbreak size and host-species
363 contribution, we aimed to compare estimates with the values in either the reference tree or the
364 reconstructible outbreak, these values were set as an offset and the intercept was set to zero.
365 The estimated incidence rates ratios (IRRs) could therefore be interpreted as multiplicative
366 factors of the outbreak size (or host contribution) in the reference tree (or reconstructible
367 outbreak).

368 **4. Alternative transmission scenarios**

369 We tested the influence of the low evolutionary rate, which is characteristic of *M. bovis*,
370 on our results. We simulated new sequences along the 30 reference trees having increased the
371 mutation rate by a factor of 10 ($\mu_n = 0.024$ substitutions per site per year) and implemented the
372 three outbreak reconstruction methods on the reference set of cases only.

373 To test whether the reconstruction of outbreak size and accuracy were influenced by the
374 complexity of the epidemiological system, we then simulated 30 new reference trees of a single-
375 host system, by setting transmission parameters to, between and from wildlife to 0, in order to
376 obtain cattle-only transmission trees. We simulated sequences along these 30 new trees with μ
377 (0.0024 substitutions per site per year), then implemented the three methods on these sequences.

378 We then analyzed whether asymmetrical roles within the multi-host system influenced
379 the reconstruction of the host-species contributions. With the same protocol (30 reference trees
380 and a low evolutionary rate), we tested a transmission scenario where one of the host-species
381 could be infected but could not play any role in transmission (dead-end epidemiological host).
382 We obtained a multi-host system where wild boars played no part in onward bTB transmission
383 by setting transmission parameters between and from wild boars to 0.

384 Finally, in order to evaluate the reconstruction of the host-species of the index case, we
385 simulated 30 new reference trees with badgers as index cases, in the multi-host system where
386 every host-species contributed to transmission.

387 **Table 1. Definition of terms used in the study (in order of appearance in the material and method).**

Reference (transmission) tree	A list of six variables (id of infector, id of infected, host-species of infector, host-species of infected, date of infection and date of death) obtained with the modified simulation model (first developed by Bouchez-Zacria <i>et al.</i> [39]).
Reference set of cases	A list of four variables (id of infected, host-species of infected, date of death and sampled sequence) obtained from the reference tree after sequence simulation.
Sampling scheme	One of six selection processes applied to a reference set of cases, five of which mimicked biases encountered on bTB data.
Biased set of cases	Set of cases obtained after applying a biased sampling scheme to a reference set of cases.
Reconstructible outbreak	A subset of the reference tree that contained only the sampled infected hosts and their ancestors.
Transmission scenario	Describes the combination of: the type of epidemiological system (multi- or single-host), whether all species contribute to transmission (presence or absence of dead-end hosts), the host-species of the index case (badger or cattle) and the mutation rate (low or high).

388 **Results**

389 **1. Transmission tree reconstruction**

390 While convergence was not a limiting factor for *outbreaker2*, it could not be obtained for
391 every set of cases in BEAST2 nor every consensus phylogenetic tree with *TransPhylo*. We were
392 thus restrained to 21 out of 30 reference trees (126 reconstructed trees in total). The reference

393 trees from which we could not reconstruct trees in *TransPhylo* showed a higher median number
394 of infected hosts compared to those whose set of sequences and trees converged (S1 Table).

395 Computational time varied greatly between sets of cases (or consensus phylogenetic
396 trees) and reconstruction methods: less than 10 min for all 126 trees reconstructed by seqTrack,
397 from less than 20 min (when only 100,000 iterations were needed) to two hours per tree
398 reconstructed by *outbreaker2*, and from less than an hour to over 12 hours (for 2,500,000
399 iterations) per tree reconstructed by *TransPhylo*. Moreover, phylogenetic reconstruction with
400 BEAST2 was needed to implement *TransPhylo* and computational time also varied between
401 sets of cases: from five hours to two days. In total, the computational time for these 378 (126
402 trees*3 methods) reconstructed trees was around three months.

403 The median proportion of unique sequences in the reference set of cases for which
404 convergence was obtained was 6.1%. The median of the mean transmission divergence was
405 0.19 (S1 Table) and the majority of transmission pairs shared the same sequence (S1 Fig).

406 All trees reconstructed by *outbreaker2* as well as all transmission probability matrices
407 estimated by *TransPhylo*, for which we kept the most probable infectors, presented cycles.

408 2. Epidemiological indicators

409 2.1 Accuracy

410 When all sequences were sampled, the median proportion of correctly reconstructed
411 transmission events (Fig 2) was 3.4% (range: 1.3-12.1) for trees reconstructed by seqTrack,
412 8.0% (2.2-11.3) for *outbreaker2* and 8.9% (6.0-16.8) for *TransPhylo* (S2 Table).

413 **Fig 2. Proportion of transmission events reconstructed from all sequences present in reference**
414 **trees according to method.**

415 Compared to *outbreaker2*, the probability of reconstructing a correct transmission event
416 was significantly lower for seqTrack (OR=0.51, p-value<0.001) but significantly higher for

417 *TransPhylo* (OR=1.30, p-value<0.001) (Table 2). In trees reconstructed by seqTrack, sampling
 418 schemes where wild boars were not sampled increased significantly the probability of
 419 reconstructing a correct transmission event (OR=1.37 and 1.30, p-value=0.001 and 0.008 for
 420 S_W and $T+S_W$ respectively). Results did not show a significant effect of the sampling scheme
 421 on accuracy for the other two methods (Table 2).

422 **Table 2. Presence of reconstructed transmission events in reference trees tested with a Binomial**
 423 **GLMM using reconstruction method, interaction between method and sampling scheme as fixed**
 424 **effects.**

Fixed effects	<i>outbreaker2</i>		seqTrack		<i>TransPhylo</i>	
	OR	p-value	OR	p-value	OR	p-value
Method	-	-	0.51	<0.001	1.30	<0.001
Method :T	0.99	0.89	0.94	0.54	0.91	0.17
Method : S_B	0.91	0.21	0.95	0.60	1.08	0.30
Method : $T+S_B$	0.92	0.32	0.94	0.57	1.11	0.14
Method : S_W	1.08	0.33	1.37	0.001	1.03	0.64
Method : $T+S_W$	1.07	0.41	1.30	0.008	1.01	0.88

425 OR stands for odds ratio. Results in bold mean that the p-value was <0.05. The *outbreaker2* method, the
 426 reference sampling scheme was set as reference, hence the “-“ present on the method line and the
 427 absence of the reference sampling scheme. T stands for “temporal bias”, S_B for “badger bias” and S_W
 428 for “wild boar bias”. $T+S_B$ ($T+S_W$) combined the temporal and the badger (wild boar) bias.

429 2.2 Super-spreaders

430 While in the reference trees the maximum number of transmission events a single
 431 infected host could be responsible for ranged from 9 to 27 (median: 14) and no super-spreaders
 432 were identified, all trees reconstructed by seqTrack presented super-spreaders. The median of
 433 the maximum number of transmission events a single super-spreader could be responsible for
 434 ranged from 90 to 108, while the median number of transmission events in the reconstructed
 435 trees ranged from 200 to 244 (S3 Table). The most frequent host-species responsible for this
 436 maximum number of transmission events was cattle (from 57% in the reference sampling
 437 scheme to 86% in the combined temporal and wild boars bias). None of the trees reconstructed
 438 by the two other methods presented super-spreaders.

439 2.3 Host-species of the index case

440 When all sequences were sampled, the proportion of correctly reconstructed host-
441 species of the index case (*i.e.* cattle) was 76% for trees reconstructed by seqTrack, 81% for
442 *outbreaker2* and 57% for *TransPhylo* (S4 Table). Except when considering the temporal bias
443 alone with the *TransPhylo* method, a temporal and a badger bias (combined or not) led to an
444 increase in the proportion of correctly reconstructed index cases.

445 2.4 Outbreak size

446 In the reference trees, the median number of infected hosts was 245 (S5 Table). Overall,
447 the simulated outbreak size was close to the credible interval estimated by *outbreaker2* (Fig 3).
448 Indeed, this credible interval contained the simulated outbreak size for all 21 trees reconstructed
449 with the reference and temporal sampling scheme. However, a species bias (combined or not
450 with a temporal bias) decreased the number of trees that correctly estimated the outbreak size
451 and led to a majority of trees that underestimated the outbreak size (20/21 with S_B and $T+S_B$,
452 16/21 for S_W and 18/21 for $T+S_W$). According to the statistical model, the outbreak size
453 estimated by *outbreaker2* was not significantly different to the reference tree size (IRR= 1.14,
454 p-value=0.43) and sampling schemes had no significant effect on outbreak size (Table 3).

455 **Fig 3. Outbreak size credible interval estimated by *outbreaker2* and *TransPhylo* compared to**
456 **simulated outbreak size.** The point corresponds to the simulated outbreak size. T stands for “temporal
457 bias”, S_B for “badger bias” and S_W for “wild boar bias”. $T+S_B$ ($T+S_W$) combined the temporal and the
458 badger (wild boar) bias.

459 *TransPhylo* could greatly overestimate the outbreak size and the difference between the
460 lower bound of the interval and the simulated outbreak size could exceed 10,000 infected hosts
461 (Fig 3). The reference and temporal sampling scheme led to an overestimation of the outbreak
462 size in the majority of reconstructed trees (19/21 and 16/21): the credible intervals contained
463 the simulated outbreak size in 3 and 5 out of 21 trees, respectively. The number of correct
464 estimations remained low for the other types of biases. The statistical model confirmed these

465 results, since the outbreak size estimated by *TransPhylo* was significantly higher than the
 466 simulated outbreak size (IRR= 2.92, p-value <0.001). Moreover, all biased schemes except for
 467 the temporal bias significantly lowered the estimated outbreak size (IRR ranging from 0.49 to
 468 0.68 and p-value <0.01).

469 **Table 3. Estimated outbreak size tested with a Negative Binomial GLMM using reconstruction**
 470 **method, interaction between method and sampling scheme as fixed effects.**

Fixed effects	<i>outbreaker2</i>		<i>TransPhylo</i>	
	IRR	p-value	IRR	p-value
Method	1.14	0.43	2.92	<0.001
Method :T	0.98	0.90	0.96	0.80
Method :S _B	0.83	0.24	0.50	<0.001
Method :T+S _B	0.83	0.23	0.49	<0.001
Method :S _W	0.87	0.34	0.68	0.01
Method :T+S _W	0.85	0.28	0.65	0.005

471 IRR stands for incidence rates ratio. Results in bold mean that the p-value was <0.05. The reference tree
 472 size was set as the offset and the reference sampling scheme was set as reference. T stands for “temporal
 473 bias”, S_B for “badger bias” and S_W for “wild boar bias”. T+S_B (T+S_W) combined the temporal and the
 474 badger (wild boar) bias.

475 2.5 Host-species contribution to transmission

476 The median number of transmission events due to each host-species in the reference
 477 trees was 175 for cattle, 24 for badgers and 40 for wild boars (S6 Table).

478 In the reference sampling scheme, the credible interval contained the simulated number
 479 of transmission events due to each host-species in few of the trees reconstructed by *outbreaker2*
 480 (2/21 trees for cattle, none for badger and wild boars) and *TransPhylo* (5/21 trees for cattle,
 481 4/21 for badgers and 3/21 for wild boars) (Fig 4). Otherwise, the number of transmission events
 482 in the majority of the remaining trees was either underestimated (cattle: 14/21 trees for
 483 *outbreaker2* and 13/21 trees for *TransPhylo*), overestimated (badgers: 19/21 trees for
 484 *outbreaker2* and 13/21 trees for *TransPhylo*) or no particular trend was observed (wild boars).
 485 Similar results were obtained with the other five sampling schemes (S2-S4 Figs).

486 **Fig 4. Credible interval of host-species contribution compared to simulated outbreaks.** The credible
 487 interval was either estimated by *outbreaker2* or by *TransPhylo*. The point corresponds to the number of
 488 transmission events due to each host-species in the simulated outbreak. Only the reference sampling
 489 scheme is considered here.

490 According to the statistical model, the underestimation of the number of reconstructed
 491 transmission events due to cattle (Fig 4) was not significant for either method (Table 4). The
 492 statistical model confirmed the results obtained for badgers, since the number of transmission
 493 events due to badgers estimated by both methods was significantly higher than the simulated
 494 number (IRR=2.06 for *outbreaker2* and 1.70 for *TransPhylo*, p-value<0.001) (Table 4). Results
 495 did not show a significant effect of the sampling scheme on badger contribution for
 496 *outbreaker2*. However, the sampling scheme with the least number of sampled hosts (temporal
 497 and wild boar biases combined) significantly decreased the number of transmission events due
 498 to badgers compared to the reference sampling scheme in trees reconstructed by *TransPhylo*.
 499 Finally, the number of transmission events due to wild boars estimated by both methods was
 500 not significantly different to the simulated number in the reference tree (Table 4).

501 **Table 4. Number of transmission events due to each host-species tested with a Negative Binomial**
 502 **GLMM per host-species using method and interaction between method and sampling scheme as**
 503 **fixed effects.**

Fixed effects	<i>outbreaker2</i>		<i>TransPhylo</i>	
	IRR	p-value	IRR	p-value
1. Cattle contribution				
Method	0.86	0.09	0.95	0.58
Method :T	1.04	0.58	1.02	0.77
Method :S _B	1.06	0.41	0.95	0.45
Method :T+S _B	1.06	0.39	0.91	0.20
Method :S _W	1.01	0.91	1.03	0.63
Method :T+S _W	1.06	0.37	1.09	0.20
2. Badger contribution				
Method	2.06	<0.001	1.70	<0.001
Method :T	0.80	0.09	0.84	0.20
Method :S _W	1.12	0.39	0.91	0.47
Method :T+S _W	0.87	0.27	0.74	0.02
3. Wild boar contribution				
Method	1.33	0.12	1.04	0.85
Method :T	0.92	0.62	1.29	0.13
Method :S _B	1.08	0.65	1.06	0.72
Method :T+S _B	1.03	0.87	1.05	0.80

504 IRR stands for incidence rates ratio. Results in bold mean that the p-value was <0.05. The number of
 505 transmission events in the reference tree was set as the offset and the reference sampling scheme was
 506 set as reference. T stands for “temporal bias”, S_B for “badger bias” and S_W for “wild boar bias”. T+S_B
 507 (T+S_W) combined the temporal and the badger (wild boar) bias.

508 3. Alternative transmission scenarios

509 3.1 Higher mutation rate

510 As expected, sequences simulated with a higher mutation rate presented a higher
511 proportion of unique sequences (median: 33.4%) and a higher mean transmission divergence
512 (median: 0.69) (S7 Table).

513 A higher mutation rate increased markedly the median accuracy for all three methods:
514 25.7% (+17.7, range: 15.9-33.3) for *outbreaker2*, 15.3% (+11.9, range: 8.2-33.3) for seqTrack
515 and 21.2% (+12.3, range: 13.2-29.3) for *TransPhylo* (S8-S9 Tables). While the majority of trees
516 reconstructed by seqTrack again contained super-spreaders (20/21), the median of the
517 maximum number of transmission events due to a single super-spreader was lower when
518 considering a higher mutation rate (35 vs. 108, S10 Table).

519 The credible interval contained the simulated outbreak size for 16/21 trees reconstructed
520 by *outbreaker2* and in only 4/21 trees reconstructed by *TransPhylo*, otherwise the outbreak size
521 was overestimated (Fig 5 and S11 Table). For both methods, the credible interval contained the
522 number of transmission events due to each host-species in only 4/21 trees for cattle, 3/21
523 (*outbreaker2*) and 5/21 (*TransPhylo*) for badgers and 1/21 for wild boars (Fig 6). Otherwise,
524 cattle contribution was underestimated by *TransPhylo* (16/21 trees in Fig 6, and S12 Table) and
525 wildlife contribution was overestimated by *outbreaker2* (13/21 trees for badgers and wild boars
526 in Fig 6, and S12 Table).

527 **Fig 5. Outbreak size credible interval compared to simulated outbreak size in the high mutation**
528 **rate scenario.** The credible interval was either estimated by *outbreaker2* or by *TransPhylo*. The point
529 corresponds to the simulated outbreak size.

530 **Fig 6. Credible interval of host-species contribution compared to simulated outbreaks in the high**
531 **mutation rate scenario.** The credible interval was either estimated by *outbreaker2* or by *TransPhylo*.
532 The point corresponds to the number of transmission events due to each host-species in the simulated
533 outbreak.

534 **3.2 Single-host system**

535 Sequences simulated within a single-host system presented a lower proportion of unique
536 sequences (median: 3.6%) and a lower mean transmission divergence (median: 0.14) (S7
537 Table).

538 Similarly to the multi-host systems, the accuracy was the highest for *TransPhylo* (6.5%),
539 then *outbreaker2* (5.5%) and the lowest for seqTrack (2%) (S8-S9 Table). Super-spreaders were
540 present in all trees reconstructed by seqTrack but also in trees reconstructed by *outbreaker2*
541 (5/26 trees, median of maximum 39 transmission events due to a single super-spreader) and
542 *TransPhylo* (10/26 trees, median: 39.5) (S10 Table).

543 The credible interval contained the simulated outbreak size in all 26 trees reconstructed
544 by *outbreaker2* and in 16/26 trees reconstructed by *TransPhylo*, otherwise the outbreak size
545 was overestimated (Fig 7 and S11 Table).

546 **Fig 7. Outbreak size credible interval compared to simulated outbreak size in the single-host**
547 **system scenario.** The credible interval was either estimated by *outbreaker2* or by *TransPhylo*. The point
548 corresponds to the simulated outbreak size.

549 **3.3 Dead-end epidemiological host**

550 The credible interval never contained the simulated number of transmission events due to
551 wild boars and wild boar contribution was overestimated in all 17 reconstructed trees (Fig 8).
552 Otherwise, similarly to the multi-host systems without a dead-end epidemiological host, cattle
553 contribution tended to be underestimated by both methods (17/17 trees for *outbreaker2* and
554 10/17 for *TransPhylo*) and badger contribution, overestimated by *outbreaker2* (15/17 trees in
555 Fig 8, and S12 Table).

556 **Fig 8. Credible interval of host-species contribution compared to simulated outbreaks in the dead-**
557 **end epidemiological host scenario.** The credible interval was either estimated by *outbreaker2* or by
558 *TransPhylo*. The point corresponds to the number of transmission events due to each host-species in the
559 simulated outbreak.

560 **3.4 Badger index**

561 The proportion of correctly reconstructed badger index cases compared to cattle index
562 cases was markedly lower for *outbreaker2* (28%), seqTrack (28%) and *TransPhylo* (11%).

563 For all transmission scenarios, even the reference multi-host scenario, similar results
564 were obtained when considering the reference tree or the reconstructible outbreak (S2 Appendix
565 and S11-S12 Tables).

566 **Discussion**

567 In this work, we evaluated and compared the performances of three outbreak
568 reconstruction methods on simulated *M. bovis* data in a multi-host system, as well as the impact
569 of observation biases on these performances. *M. bovis*, characterized by a low mutation rate, is
570 a prime example of a multi-host pathogen for which sampling biases complicate the estimation
571 of host-species contribution to transmission, an estimation which is however necessary to select
572 appropriate measures for disease control. Contrary to previous evaluations of outbreak
573 reconstruction methods, the transmission model we used to simulate our data was not tailored
574 to a specific method [25,31,52] but to the slowly evolving multi-host pathogen. Moreover, the
575 epidemiological indicators we estimated were also relevant in a multi-host system and not just
576 general performance indicators [53].

577 Reconstructing transmission trees can have multiple objectives according to the studied
578 pathogen and epidemiological system, the most obvious objective is the accurate reconstruction
579 of who-infected-whom. The proportion of correctly reconstructed transmission events (which
580 we called accuracy) has previously been used to evaluate performances of outbreak
581 reconstruction methods [25,53]. With the low mutation rate characteristic of *M. bovis*, we
582 estimated poor accuracies (median accuracy lower than 9% for all three methods). Sobkowiak
583 *et al.* compared these outbreak reconstruction methods on real *M. tuberculosis* data, which is

584 also a slow-evolving pathogen, and estimated the positive predictive value (PPV), meaning the
585 number of epidemiologically linked case-contact pairs that were correctly identified (preprint,
586 [54]). Contrary to the accuracy indicator we estimated, the links between cases were not
587 directed, we thus expected this study to estimate a higher number of correctly reconstructed
588 cases. The PPV estimated by Sobkowiak *et al.* was 15% for *TransPhylo*, 11% for *outbreaker2*
589 and 10% for seqTrack. These PPV values were in the range of values we estimated for accuracy
590 and the ranking of methods was the same as the one we obtained (with *TransPhylo* as the best,
591 followed by *outbreaker2*).

592 Accuracy was little influenced by the sampling biases or the complexity of the
593 epidemiological system, however it was greatly dependent on the mutation rate. When the
594 mutation rate was multiplied by a factor of 10 ($\sim 6.6 \times 10^{-5}$ substitutions per site per day), the
595 accuracies we estimated more than doubled. In the study that presented and tested *outbreaker2*,
596 Campbell *et al.* estimated the average proportion of transmission pairs correctly inferred when
597 using solely temporal and genetic information from simulated Ebola virus (mutation rate: 0.31
598 $\times 10^{-5}$ per site per day) and SARS-CoV-1 (1.14×10^{-5} per site per day) outbreaks [25]. Moreover,
599 Firestone *et al.* compared *TransPhylo* and *outbreaker2* on six FMDV outbreaks simulated with
600 a high mutation rate (2.2×10^{-5} per site per day) and estimated the proportion of infected hosts
601 (premises) for which the most likely source predicted was the true source [53]. Since both
602 indicators corresponded to the accuracy we estimated, we expected similar results. However,
603 Campbell *et al.* estimated an average accuracy of 29% (from the simulated Ebola data) and 70%
604 (SARS-CoV-1). In addition, when genomic data was available for all infected hosts, the
605 accuracy estimated by Firestone *et al.* was 4% for *TransPhylo* and 35% for *outbreaker2*. While
606 these values were respectively higher for *outbreaker2* and lower for *TransPhylo* compared to
607 the range of values we calculated, the ranking of methods obtained by Firestone *et al.* was the
608 same as the one we obtained (with *outbreaker2* as the better of the two).

609 The lowest accuracy always being estimated for seqTrack could be due to the fact that
610 this method does not consider a transmission model [21], but simply sampling dates and genetic
611 distances. As mentioned by Nigsch *et al.*, seqTrack is thus strongly dependent on the temporal
612 order of sampling dates and when the sampling order does not necessarily coincide with the
613 infection order (here, because of imperfect case detection and sampling protocol varying
614 according to host-species), “the order of ancestries cannot be inferred with certainty” [55].
615 Contrary to what we observed with *outbreaker2* and *TransPhylo*, trees reconstructed by
616 seqTrack presented super-spreaders with extreme numbers of transmission events due to a
617 single infected host (over a hundred transmissions) that lowered when considering a higher
618 mutation rate. The low genetic diversity combined with the lack of a transmission model could
619 therefore account for the reconstruction of super-spreaders, which in turn could contribute to
620 the low accuracy. Similarly, the lower genetic diversity obtained with the single-host system
621 could explain the presence of less prolific super-spreaders in trees reconstructed with
622 *TransPhylo* and *outbreaker2*.

623 While we estimated poor accuracies for all three methods, a high proportion of correctly
624 reconstructed directed transmission events is difficult to obtain and might not be the main
625 objective when studying a multi-host system implicating wildlife or with a low sampling
626 proportion. However, the presence of super-spreaders is an important indicator to consider since
627 it highlighted the fact that seqTrack reconstructed unrealistic transmission dynamics with
628 prolific super-spreaders.

629 Other than reconstructing who-infected-whom, outbreak reconstruction can aim to
630 estimate epidemiological indicators, from which practical measures can be directly inferred.
631 The first we studied was the outbreak size, which could by comparison with the number of
632 sampled cases be informative *e.g.* of the need to increase the sampling effort [35]. Outbreak
633 size estimation was sensitive to sampling biases, the complexity of the epidemiological system

634 and also the mutation rate. The outbreak size was correctly estimated by *outbreaker2* but
635 consistently overestimated by *TransPhylo*, even though we considered the same non-
636 informative prior for the sampling proportion when implementing both methods. This
637 overestimation could therefore be due to the fact that Didelot *et al.* developed this method to
638 study partially sampled *M. tuberculosis* outbreaks and account for within-host diversity [31],
639 whereas we assumed all cases sampled in the reference scheme and no within-host diversity in
640 the sequence simulation. Furthermore, when not all sequences were sampled, better results were
641 obtained for *TransPhylo* and the estimated outbreak size significantly lowered.

642 With a higher mutation rate, *TransPhylo* also overestimated the outbreak size, but to a
643 lesser extent. Xu *et al.* developed in 2019 a method of simultaneous inference on multiple *M.*
644 *tuberculosis* clusters based on *TransPhylo*. From this study, Xu *et al.* discussed the link between
645 mutation rate and sampling proportion, explaining that with a faster assumed clock, the
646 branches in the phylogenetic trees are shorter and *TransPhylo* is therefore less likely to place
647 unsampled cases along them [56]. This could explain the lower effect we estimated.

648 Some epidemiological indicators are relevant only in the context of a multi-host system
649 and reveal the host-species that should be primarily targeted, such as the identification of the
650 host-species responsible for the outbreak and the accurate reconstruction of each host-species'
651 contribution to the outbreak. The index case indicator was sensitive to sampling biases and to
652 the host-species of the index case. The proportion of correctly reconstructed host-species of the
653 index case was high for *outbreaker2* and seqTrack (over 75%) when considering cattle index
654 cases. However, the fact that *TransPhylo* could designate unsampled hosts as index cases,
655 combined with a tendency to overestimate the outbreak size and thus, the number of unsampled
656 hosts, could explain this method's poorer performance. Moreover, biased sampling schemes
657 generally led to a higher proportion of correctly reconstructed host-species of the index case,
658 which could be explained by the fact that only non-index cases (wildlife) were concerned by

659 these sampling schemes. Finally, this indicator was sensitive to the host-species responsible for
660 the outbreak and had a poorer performance when the index case was a badger.

661 Host contribution estimation was influenced by the sampling biases and the complexity
662 of the epidemiological system but not the mutation rate. With either mutation rates, *outbreaker2*
663 and *TransPhylo* poorly reconstructed the contribution of each host-species and tended to
664 underestimate the host-species that contributed the most to transmission (cattle) while
665 overestimating those that contributed the least (wildlife). Both outbreak reconstruction methods
666 were developed and tested on single-host systems [21,25,31], and not on multi-host systems
667 where each host-species play a different role. While *TransPhylo* has previously been applied to
668 multi-host systems, a human-deer SARS-CoV-2 system [35] and two badger-cattle bTB
669 systems [57,58], the estimation of host-species contribution to transmission in these systems
670 was not straightforward. The high number of unsampled cases estimated in the human-deer
671 system (mean sampling proportion of 0.1%) complicated the inference of transmission events
672 and while phylogenetic evidence seemed to support multiple human-to-deer spillover events,
673 deer-to-human transmission could not be ruled out [35]. In a badger-cattle system in the South-
674 West of England, van Tonder *et al.* were interested in between-species transmission and as such
675 *TransPhylo* was implemented in addition to a Bayesian ancestral state reconstruction method
676 (BASTA, [16]), which was primarily used to estimate the number of within- and between-
677 species transitions [57]. Finally, Akhmetova *et al.* also implemented *TransPhylo* in addition to
678 Bayesian phylogenetic methods in a badger-cattle system in Northern Ireland and highlighted
679 a mostly cattle-driven (over 90% of strongly supported reconstructed transmission events)
680 epidemic in the region [58].

681 Biases simulated with the sampling schemes resulted in a decrease in the number of
682 infected hosts for which contribution estimates was overestimated. Therefore, when sampling
683 schemes had a significant effect on host contribution, they tended to yield better results with

684 this particular host-system and either lowered (for wildlife) or increased (for cattle) the
685 estimated number of transmission events. In addition, neither *outbreaker2* nor *TransPhylo*
686 could accurately reconstruct asymmetrical roles between host-species, *i.e.* the presence of a
687 dead-end epidemiological host.

688 In the epidemiological multi-host system we extended, the basic reproduction number
689 varied according to the combination of host-species considered [39]. Moreover, Bouchez-
690 Zacia *et al.* calculated inter- and intra-species generation time distributions that showed a more
691 rapid spread from cattle farms than from badger groups. We added to the transmission model,
692 a third population of host-species (wild boars) that could transmit or not the pathogen. The
693 complexity of this multi-host system could have contributed to the poor results we obtained for
694 host-species contribution. Indeed, both *outbreaker2* and *TransPhylo* considered a single
695 generation time, sampling time and/or offspring distribution for all three host-species, not
696 accounting for host-species variation in the natural history of the disease nor the uneven
697 transmission dynamics. A multi-host system where all three host-species contributed unevenly
698 to transmission is not unusual, results obtained from Bayesian ancestral state reconstructions
699 (Mascot, [19]) in other French regions point to the presence of similarly complex bTB multi-
700 host systems [59]. Furthermore, the impact of said complexity on method performance does not
701 only concern systems with multiple host-species, pathogens for which different categories (*e.g.*
702 age groups and/or vaccination status [60]) of hosts can be defined (according to infectiousness
703 or duration of infection) also constitute complex epidemiological systems. Finally, considering
704 what can be reconstructed by the method (reconstructible outbreak) instead of the reference tree
705 did not improve results for the outbreak size nor the host contribution indicators.

706 We were limited by practical considerations and the ensuing choices we made. With the
707 *M. bovis* data we simulated, convergence was a limiting factor for *TransPhylo* but not for
708 *outbreaker2*. Indeed, in order to limit the computational time, we fixed a maximum number of

709 iterations, which narrowed the number of reconstructed trees we could compare to those that
710 converged in less than 48 hours in BEAST2 and 12 hours in *TransPhylo*. Moreover, in order to
711 better compare reconstructions, we used the same evolutionary model for the phylogenetic
712 reconstruction in BEAST2. A more adapted evolutionary model could lead to a more accurate
713 phylogenetic tree reconstruction and thus, a better performance from *TransPhylo*.

714 With the sequence simulation model we implemented, we simulated a low proportion
715 of unique sequences from 13-year-long outbreaks, which is consistent with *M. bovis* low
716 mutation rate. In the study on 167 *M. bovis* sequences in the South-West of France from which
717 we selected the value of the mutation rate [12], the proportion of unique sequences isolated
718 (37.1%) was around six times higher than the median proportion we simulated. The higher
719 proportion of unique sequences in this previous study could be due to the fact that not all
720 sequences are sampled in real data and that the outbreak lasted longer as suggested by the
721 MRCA which was estimated to have been circulating 27 years earlier. In the sequence
722 simulation model, we also considered the same mutation rate within all three host-species,
723 however whether *M. bovis* evolves the same way within different host-species remains
724 unknown. Indeed, *M. tuberculosis* mutation rates in humans may decrease during periods of
725 latency, which differs from what was observed in non-human primates [61]. A similar
726 phenomenon could lead to variability in the evolution of *M. bovis* within and between host-
727 species [62] and thus, to the difference in the proportion of unique sequences observed and
728 simulated.

729 We chose to compare results from the same epidemiological and genetic data for all
730 three methods. While difficult to implement when studying a slowly evolving multi-host system
731 that implicates wildlife, contact data can be directly incorporated in the transmission tree
732 inference with *outbreaker2*. The addition of contact data led to higher accuracies than those
733 obtained with only temporal and genetic data in simulated Ebola virus and SARS-CoV-1

734 outbreaks [25]. Similarly, when limited genetic diversity was expected in their study on *M.*
735 *avium* ssp *paratuberculosis*, Nigsch *et al.* took advantage of the fact that seqTrack can
736 incorporate additional data in the form of weighting matrices [55]. They thus resolved equally
737 probable ancestries using known exposure time or susceptibility based on accepted
738 epidemiological knowledge. Even when the method does not allow additional epidemiological
739 data, Xu *et al.* mentioned that one of the strengths in their study on *M. tuberculosis* transmission
740 within a Spanish cohort lied in the extensive contact investigation data that allowed them to
741 validate the results of their genomic and *TransPhylo* analysis [63]. Using these available
742 features and strategies could have improved results obtained for *outbreaker2* and seqTrack as
743 well as help evaluate those from *TransPhylo*. However, since limited real contact data can be
744 obtained for wildlife, we chose not to include additional epidemiological data in this study.
745 Furthermore, we limited our study to only three methods, available in a package and that only
746 needed sampling times as epidemiological data. Additional methods would be interesting to test
747 on this simulated data, especially methods that simultaneously inferred phylogenetic and
748 transmission trees like *phybreak* [26], since none were considered here. Finally, the simulated
749 multi-host data could also be used to test Bayesian ancestral state reconstruction methods like
750 Mascot [19], previously used to study complex bTB multi-host systems [59].

751 The overall poor performances we obtained for accuracy and host-species contribution,
752 even without biased sampling schemes, suggest that when studying the transmission of a slowly
753 evolving pathogen in complex multi-host systems, outbreak reconstruction methods should not
754 be implemented alone but as a complement to epidemiological and phylogenetic methods. The
755 difficulty in estimating host-species contribution highlights the need to develop new outbreak
756 reconstruction methods adapted to complex epidemiological systems as well as evaluate these
757 methods on data simulated in multi-host systems and not specific to the each method.

758 **Availability of data and materials**

759 All simulated data and code used to simulate data, reconstruct outbreaks and evaluate
760 methods are available on Github (<https://github.com/duaulthe1/bTBtreereconstruction.git>).

761 **Competing interests**

762 The authors declare that they have no competing interests.

763 **Funding**

764 This work was financially supported by the Université Paris-Saclay, which funded
765 H.D.'s PhD grant.

766 **Acknowledgements**

767 Not applicable.

References

1. Cleaveland S, Laurenson MK, Taylor LH. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. *Philos Trans R Soc Lond B Biol Sci.* 29 juill 2001;356(1411):991-9.
2. Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. *Philos Trans R Soc Lond B Biol Sci.* 29 juill 2001;356(1411):983-9.
3. Portier J, Ryser-Degiorgis MP, Hutchings MR, Monchâtre-Leroy E, Richomme C, Larrat S, et al. Multi-host disease management: the why and the how to include wildlife. *BMC Vet Res.* 14 août 2019;15:295.
4. Cross AR, Baldwin VM, Roy S, Essex-Lopresti AE, Prior JL, Harmer NJ. Zoonoses under our noses. *Microbes Infect.* 2019;21(1):10-9.
5. Viana M, Cleaveland S, Matthiopoulos J, Halliday J, Packer C, Craft ME, et al. Dynamics of a morbillivirus at the domestic–wildlife interface: Canine distemper virus in domestic dogs and lions. *Proc Natl Acad Sci U S A.* 3 févr 2015;112(5):1464-9.
6. Gortázar C, Ferroglio E, Höfle U, Frölich K, Vicente J. Diseases shared between wildlife and livestock: a European perspective. *Eur J Wildl Res.* 1 nov 2007;53(4):241-56.
7. O'Reilly LM, Daborn CJ. The epidemiology of *Mycobacterium bovis* infections in animals and man: a review. *Tuber Lung Dis.* août 1995;76 Suppl 1:1-46.
8. Simpson VR. Wild Animals as Reservoirs of Infectious Diseases in the UK. *The Veterinary Journal.* 1 mars 2002;163(2):128-46.
9. Naranjo V, Gortazar C, Vicente J, de la Fuente J. Evidence of the role of European wild boar as a reservoir of *Mycobacterium tuberculosis* complex. *Veterinary Microbiology.* févr 2008;127(1-2):1-9.
10. Kean JM, Barlow ND, Hickling GJ. Evaluating potential sources of bovine tuberculosis infection in a New Zealand cattle herd. *New Zealand Journal of Agricultural Research.* janv 1999;42(1):101-6.
11. Réveillaud É, Desvaux S, Boschioli ML, Hars J, Faure É, Fediaevsky A, et al. Infection of Wildlife by *Mycobacterium bovis* in France Assessment Through a National Surveillance System, Sylvatub. *Front Vet Sci.* 2018;5:262.
12. Duault H, Michelet L, Boschioli ML, Durand B, Canini L. A Bayesian evolutionary model towards understanding wildlife contribution to F4-family *Mycobacterium bovis* transmission in the South-West of France. *Veterinary Research.* 2 avr 2022;53(1):28.
13. Salvador LCM, O'Brien DJ, Cosgrove MK, Stuber TP, Schooley AM, Crispell J, et al. Disease management at the wildlife-livestock interface: Using whole-genome sequencing to study the role of elk in *Mycobacterium bovis* transmission in Michigan, USA. *Molecular Ecology.* 2019;28(9):2192-205.
14. Crispell J, Zadoks RN, Harris SR, Paterson B, Collins DM, de-Lisle GW, et al. Using whole genome sequencing to investigate transmission in a multi-host system: bovine tuberculosis in New Zealand. *BMC Genomics.* 16 2017;18(1):180.

15. Crispell J, Benton CH, Balaz D, De Maio N, Ahkmetova A, Allen A, et al. Combining genomics and epidemiology to analyse bi-directional transmission of *Mycobacterium bovis* in a multi-host system. *eLife*. 17 déc 2019;8.
16. De Maio N, Wu CH, O'Reilly KM, Wilson D. New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics*. 12 août 2015;11(8).
17. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. *Ferguson NM, éditeur. eLife*. 16 janv 2018;7:e31257.
18. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian Phylogeography Finds Its Roots. *PLoS Comput Biol*. 25 sept 2009;5(9).
19. Müller NF, Rasmussen D, Stadler T. MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics*. 15 nov 2018;34(22):3843-8.
20. Müller NF, Rasmussen DA, Stadler T. The Structured Coalescent and Its Approximations. *Mol Biol Evol*. nov 2017;34(11):2970-81.
21. Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*. 2010/06/17 éd. févr 2011;106(2):383-90.
22. Varia M, Wilson S, Sarwal S, McGeer A, Gournis E, Galanis E, et al. Investigation of a nosocomial outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada. *CMAJ: Canadian Medical Association Journal*. 19 août 2003;169(4):285-92.
23. Garry M, Hope L, Zajac R, Verrall AJ, Robertson JM. Contact tracing: a memory task with consequences for public health. *Perspectives on Psychological Science*. janv 2021;16(1):175-87.
24. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS computational biology*. 2014/01/28 éd. janv 2014;10(1):e1003457.
25. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS computational biology*. 2019/03/30 éd. mars 2019;15(3):e1006930.
26. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS computational biology*. 2017/05/26 éd. mai 2017;13(5):e1005495.
27. Didelot X, Gardy J, Colijn C. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution*. 2014/04/10 éd. juill 2014;31(7):1869-79.
28. Cottam EM, Thebaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings Biological sciences*. 2008/01/31 éd. 22 avr 2008;275(1637):887-95.
29. Morelli MJ, Thebaud G, Chadoeuf J, King DP, Haydon DT, Soubeyrand S. A Bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS computational biology*. 2012/11/21 éd. 2012;8(11):e1002768.

30. Duault H, Durand B, Canini L. Methods Combining Genomic and Epidemiological Data in the Reconstruction of Transmission Trees: A Systematic Review. *Pathogens*. 15 févr 2022;11(2):252.
31. Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution*. 2017/01/20 éd. 1 avr 2017;34(4):997-1007.
32. Hall M, Woolhouse M, Rambaut A. Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLoS Comput Biol*. déc 2015;11(12):e1004613.
33. Worby CJ, O'Neill PD, Kypraios T, Robotham JV, De Angelis D, Cartwright EJ, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *The annals of applied statistics*. 2016/04/05 éd. mars 2016;10(1):395-417.
34. Firestone SM, Hayama Y, Lau MSY, Yamamoto T, Nishi T, Bradhurst RA, et al. Transmission network reconstruction for foot-and-mouth disease outbreaks incorporating farm-level covariates. *PloS one*. 2020/07/16 éd. 2020;15(7):e0235660.
35. Willgert K, Didelot X, Surendran-Nair M, Kuchipudi SV, Ruden RM, Yon M, et al. Transmission history of SARS-CoV-2 in humans and white-tailed deer. *Sci Rep*. 15 juill 2022;12:12094.
36. Sashittal P, El-Kebir M. Sampling and summarizing transmission trees with multi-strain infections. *Bioinformatics (Oxford, England)*. 2020/07/14 éd. 1 juill 2020;36(Supplement_1):i362-70.
37. Richomme C, Réveillaud E, Moyen JL, Sabatier P, De Cruz K, Michelet L, et al. Mycobacterium bovis Infection in Red Foxes in Four Animal Tuberculosis Endemic Areas in France. *Microorganisms*. 17 juill 2020;8(7):1070.
38. Desvaux S, Réveillaud É, Richomme C, Boschioli ML, Delavenne C, Calavas D, et al. Sylvatub: Bilan 2015-2017 de la surveillance de la tuberculose dans la faune sauvage. *Bulletin épidémiologique*. 2019;91(14):10.
39. Bouchez-Zacria M, Ruelle S, Richomme C, Lesellier S, Payne A, Boschioli ML, et al. Analysis of a multi-type resurgence of Mycobacterium bovis in cattle and badgers in Southwest France, 2007-2019. *Veterinary Research*. 3 mai 2023;54(1):41.
40. Hauer A, Michelet L, Cochard T, Branger M, Nunez J, Boschioli ML, et al. Accurate Phylogenetic Relationships Among Mycobacterium bovis Strains Circulating in France Based on Whole Genome Sequencing and Single Nucleotide Polymorphism Analysis. *Front Microbiol*. 2019;10.
41. Hauer A, Cruz KD, Cochard T, Godreuil S, Karoui C, Henault S, et al. Genetic Evolution of Mycobacterium bovis Causing Tuberculosis in Livestock and Wildlife in France since 1978. *PLOS ONE*. 6 févr 2015;10(2):e0117103.
42. Hasegawa M, Kishino H, Yano T aki. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*. oct 1985;22(2):160-74.
43. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*. 1 déc 1977;81(25):2340-61.
44. De Maio N, Boulton W, Weilguny L, Walker CR, Turakhia Y, Corbett-Detig R, et al. phastSim: efficient simulation of sequence evolution for pandemic-scale datasets. *bioRxiv*. 23 sept 2021;2021.03.15.435416.

45. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*. 1 févr 2019;35(3):526-8.
46. Jombart T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. 1 juin 2008;24(11):1403-5.
47. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*. 1 nov 2011;27(21):3070-1.
48. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and output analysis for MCMC. *R News*. mars 2006;6(1):7-11.
49. Jombart T, Cori A, Finger F. Small Helpers and Tricks for Epidemics Analysis [Internet]. [cité 10 mars 2022]. Disponible sur: <http://www.repidemicsconsortium.org/epitrix/>
50. Didelot X, Kendall M, Xu Y, White PJ, McCarthy N. Genomic Epidemiology Analysis of Infectious Disease Outbreaks Using TransPhylo. *Curr Protoc*. févr 2021;1(2):e60.
51. Kendall M, Ayabina D, Xu Y, Stimson J, Colijn C. Estimating Transmission from Genetic and Epidemiological Data: A Metric to Compare Transmission Trees. *Statistical Science*. 2018;33(1):70-85.
52. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? *PLoS Pathogens* [Internet]. 2018;14(2). Disponible sur: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85042693214&doi=10.1371%2fjournal.ppat.1006885&partnerID=40&md5=42f90951988d807bd8023c13e5ef71da>
53. Firestone SM, Hayama Y, Bradhurst R, Yamamoto T, Tsutsui T, Stevenson MA. Reconstructing foot-and-mouth disease outbreaks: a methods comparison of transmission network models. *Scientific reports*. 2019/03/20 éd. 18 mars 2019;9(1):4809.
54. Sobkowiak B, Romanowski K, Sekirov I, Gardy JL, Johnston J. Comparing transmission reconstruction models with *Mycobacterium tuberculosis* whole genome sequence data. *bioRxiv*; 2022. p. 2022.01.07.475333.
55. Nigsch A, Robbe-Austerman S, Stuber TP, Pavinski Bitar PD, Gröhn YT, Schukken YH. Who infects whom?-Reconstructing infection chains of *Mycobacterium avium* ssp. *paratuberculosis* in an endemically infected dairy herd by use of genomic data. *PLoS One*. 2021;16(5):e0246983.
56. Xu Y, Cancino-Munoz I, Torres-Puente M, Villamayor LM, Borrás R, Borrás-Manez M, et al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. *PLoS medicine*. 2019/11/02 éd. oct 2019;16(10):e1002961.
57. van Tonder AJ, Thornton MJ, Conlan AJK, Jolley KA, Goolding L, Mitchell AP, et al. Inferring *Mycobacterium bovis* transmission between cattle and badgers using isolates from the Randomised Badger Culling Trial. *PLoS Pathog*. nov 2021;17(11):e1010075.
58. Akhmetova A, Guerrero J, McAdam P, Salvador LCM, Crispell J, Lavery J, et al. Genomic epidemiology of *Mycobacterium bovis* infection in sympatric badger and cattle populations in Northern Ireland. *Microbial Genomics*. 2023;9(5):001023.

59. Canini L, Modenesi G, Courcoul A, Boschirolu ML, Durand B, Michelet L. Deciphering the role of host species for two *Mycobacterium bovis* genotypes from the European 3 clonal complex circulation within a cattle-badger-wild boar multihost system. *MicrobiologyOpen*. 2023;12(1):e1331.
60. Xue Y, Chen D, Smith SR, Ruan X, Tang S. Coupling the Within-Host Process and Between-Host Transmission of COVID-19 Suggests Vaccination and School Closures are Critical. *Bull Math Biol*. 2023;85(1):6.
61. Colangeli R, Arcus VL, Cursons RT, Ruthe A, Karalus N, Coley K, et al. Whole Genome Sequencing of *Mycobacterium tuberculosis* Reveals Slow Growth and Low Mutation Rates during Latent Infections in Humans. *PLoS One*. 11 mars 2014;9(3):e91024.
62. Kao RR, Price-Carter M, Robbe-Austerman S. Use of genomics to track bovine tuberculosis transmission. *Rev - Off Int Epizoot*. avr 2016;35(1):241-58.
63. Séraphin MN, Didelot X, Nolan DJ, May JR, Khan MSR, Murray ER, et al. Genomic investigation of a *Mycobacterium tuberculosis* outbreak involving prison and community cases in Florida, United States. *American Journal of Tropical Medicine and Hygiene*. 2018;99(4):867-74.

Supporting information

S1 Appendix. Details on transmission tree simulation, phylogenetic and transmission tree reconstruction.

S2 Appendix. Results on outbreak size and host contribution indicators using the reconstructible outbreak as a reference.

S1 Fig. Proportion of transmission pairs with 0, 1 and 2 SNPs between their sequences according to transmission scenario. Reference stands for the complex multi-host system where cattle are index cases and wild boars contribute to transmission. High mutation rate is the same scenario as the reference except for the higher mutation rate used to simulate sequences. Single-host stands for the only-cattle scenario. Dead-end host stands for the scenario where wild boars did not contribute to transmission and badger index, the reference scenario with badger as index cases.

S2 Fig. Credible interval of the number of transmission events due to cattle estimated by *outbreaker2* and *TransPhylo* compared (color) to the number in the simulated outbreak (point) according to sampling scheme. T stands for “temporal bias”, SB for “badger bias” and SW for “wild boar bias”. T+SB (T+SW) combined the temporal and the badger (wild boar) bias.

S3 Fig. Credible interval of the number of transmission events due to badgers estimated by *outbreaker2* and *TransPhylo* compared (color) to the number in the simulated outbreak (point) according to sampling scheme. T stands for “temporal bias”, SW for “wild boar bias” and T+SW combined the temporal and the wild boar bias.

S4 Fig. Credible interval of the number of transmission events due to wild boars estimated by *outbreaker2* and *TransPhylo* compared (color) to the number in the simulated outbreak (point) according to sampling scheme. T stands for “temporal bias”, SB for “badger bias” and T+SB combined the temporal and the badger bias.

S1 Table. Comparison between reference trees that converged in BEAST2 and *TransPhylo* and those that did not.

S2 Table. Proportion of reconstructed transmission events that were present in the reference trees according to method and sampling scheme.

S3 Table. Maximum number of transmission events a single super-spreader could be responsible for in a tree reconstructed by seqTrack and their host-species according to sampling scheme.

S4 Table. Proportion (%) of correctly reconstructed host-species of the index case according to method and sampling scheme.

S5 Table. Number of infected hosts present in the induced subtrees and reconstructed trees according to method and sampling scheme.

S6 Table. Number of transmission events due to each host-species in the reconstructible outbreak and reconstructed trees according to method and sampling scheme.

S7 Table. Comparison between reference trees that converged in BEAST2 and *TransPhylo* and those that did not, according to transmission scenario.

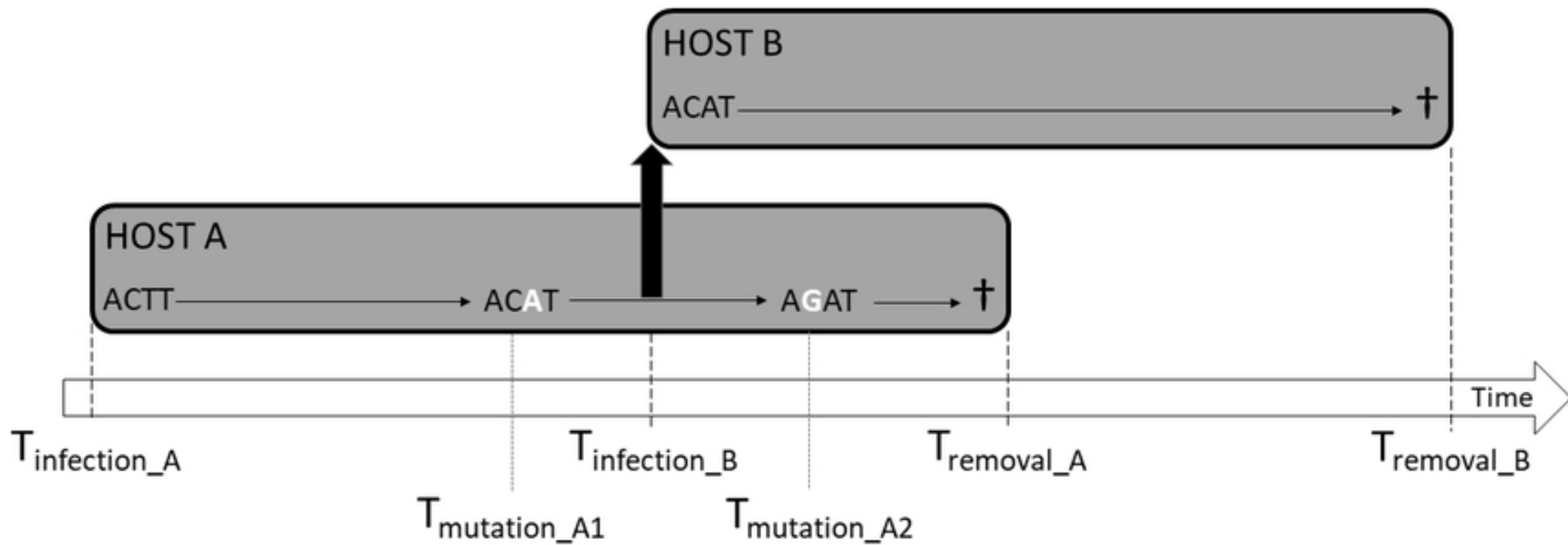
S8 Table. Proportion (%) of reconstructed transmission events that were present in the reference trees according to method and transmission scenario.

S9 Table. Accuracy tested with a Binomial GLM using method as the explanatory variable, according to transmission scenario.

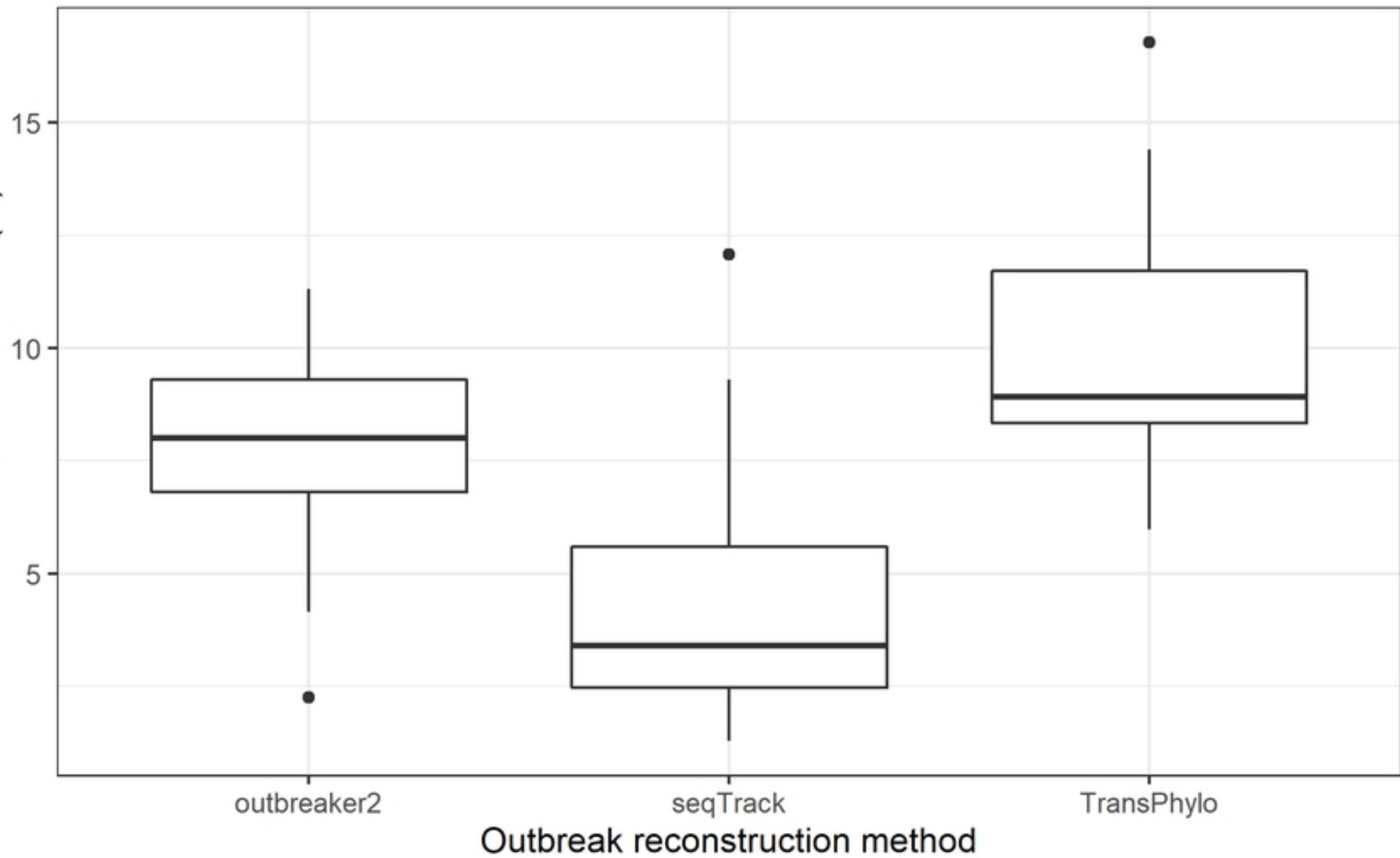
S10 Table. Number of trees reconstructed where super-spreaders were present and the maximum number of transmission events a single super-spreader could be responsible for according to method and transmission scenario.

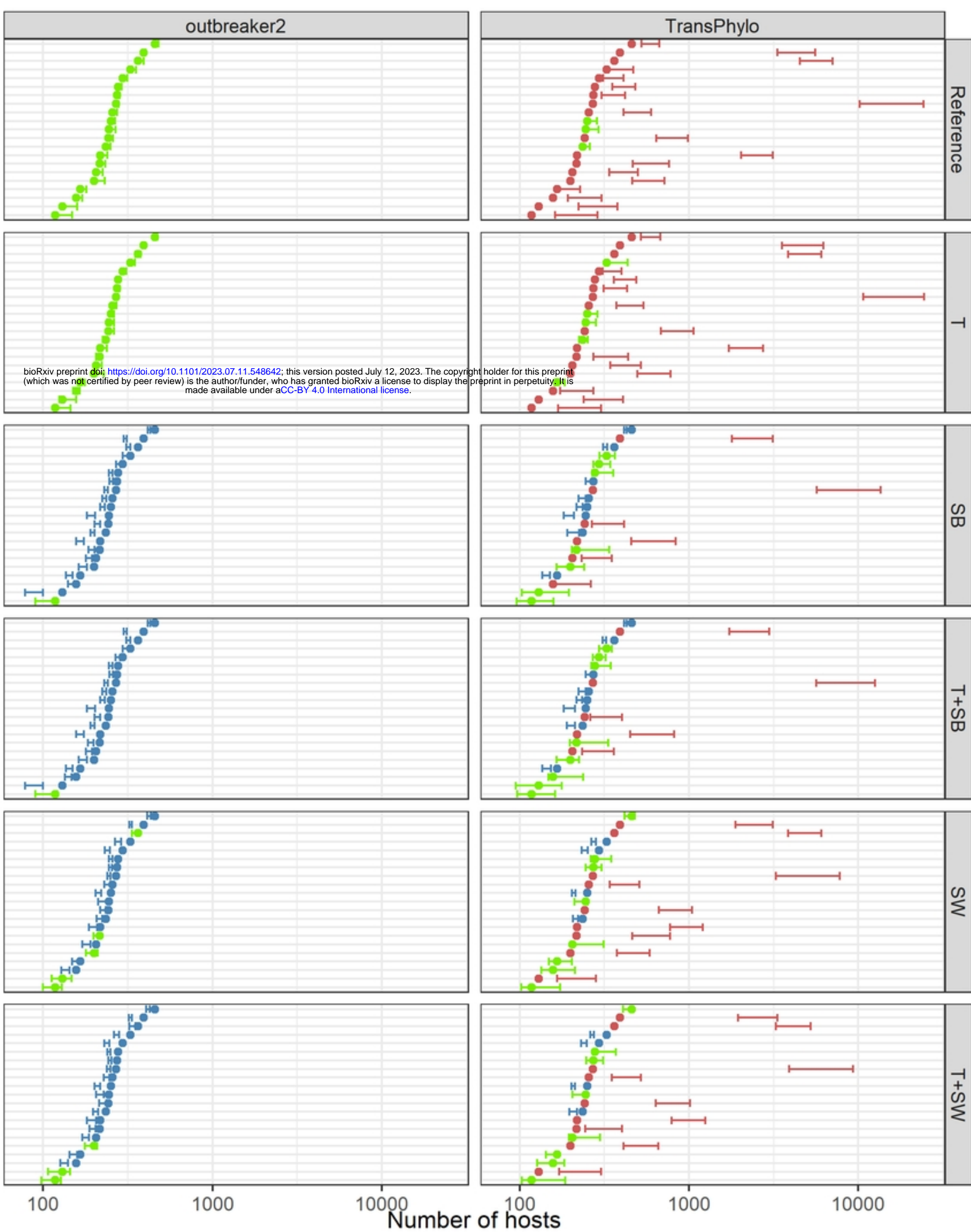
S11 Table. Outbreak size with a Negative Binomial GLM using method as the explanatory variable, according to transmission scenario.

S12 Table. Host contribution tested with a Negative Binomial GLM using method as the explanatory variable, according to transmission scenario.

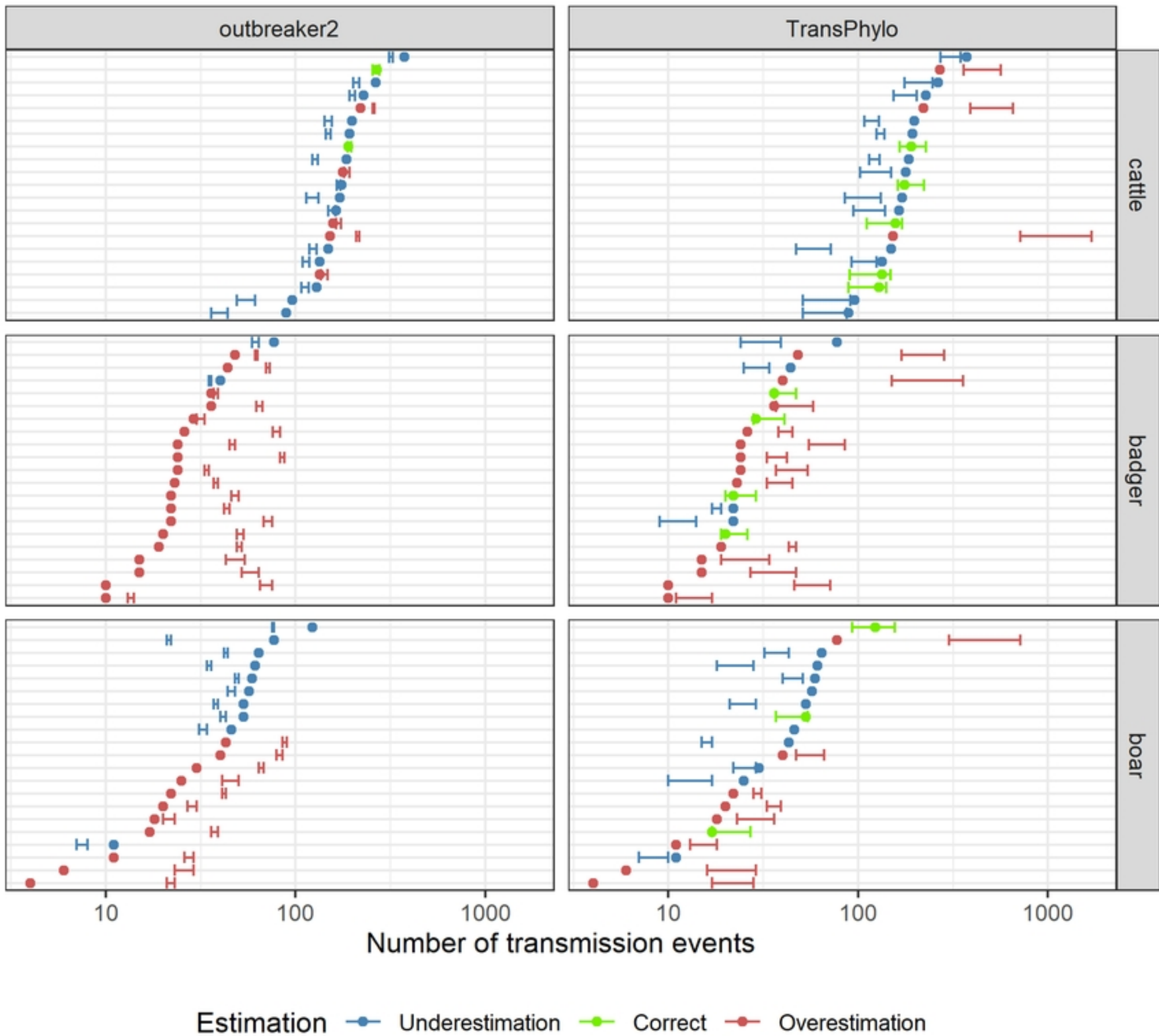


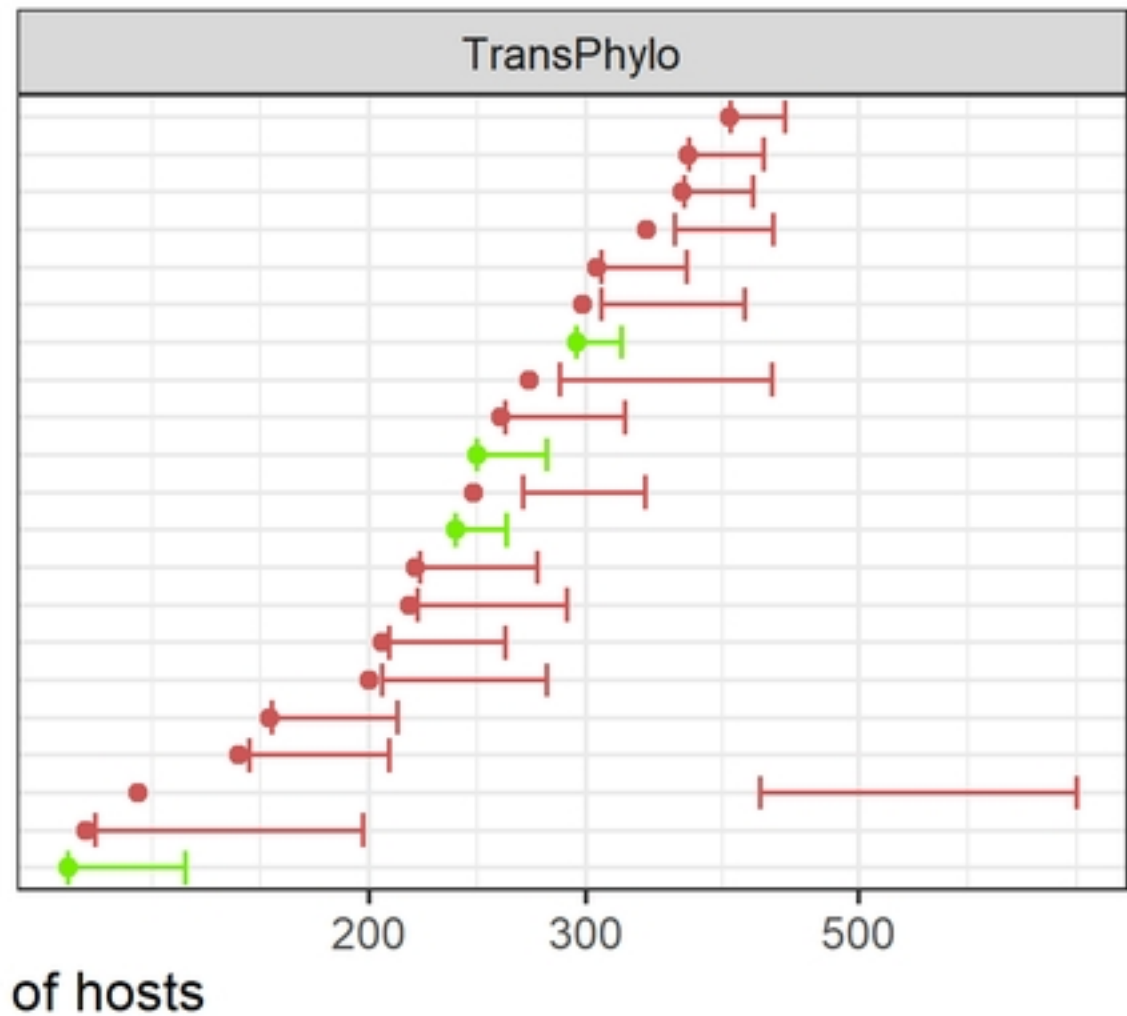
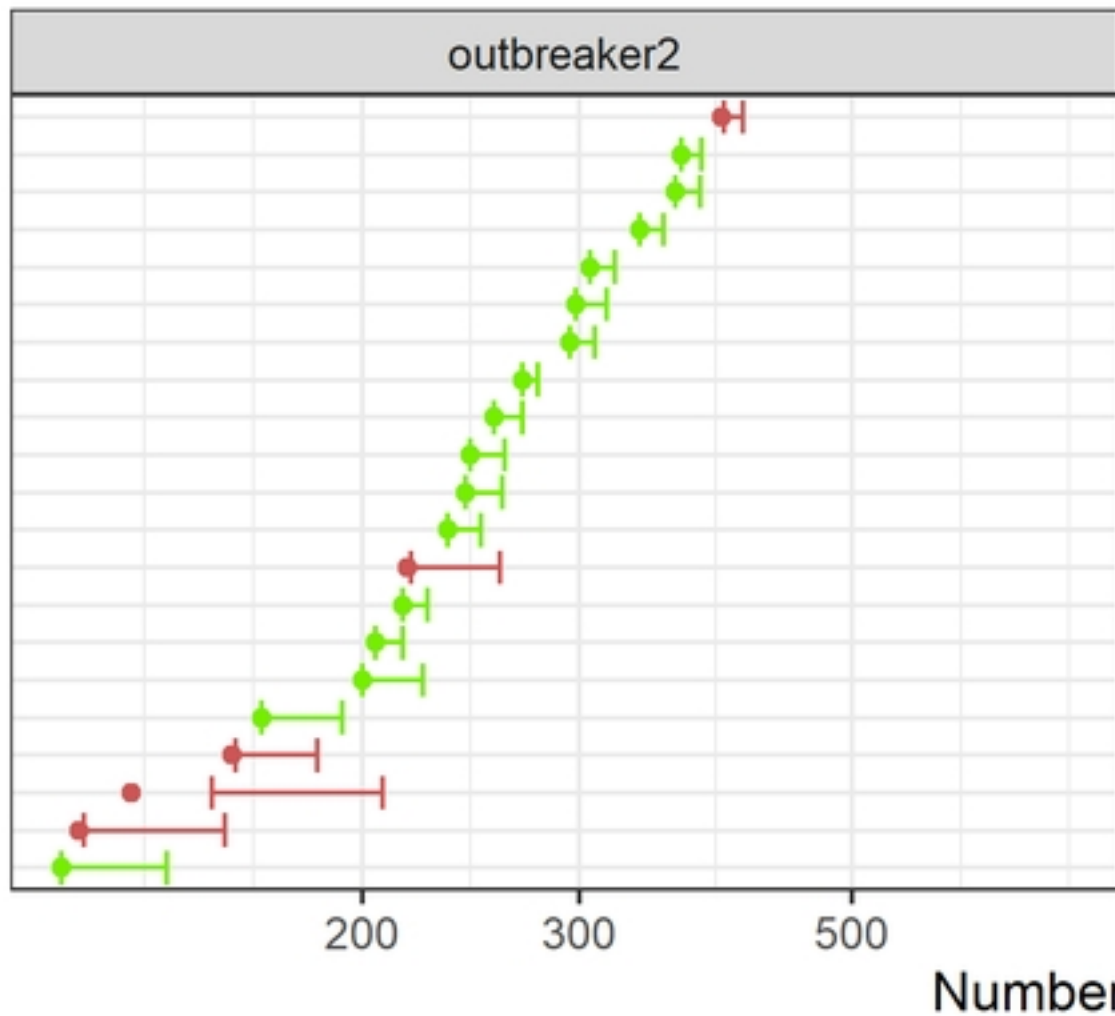
Proportion of correctly identified transmission events (%)



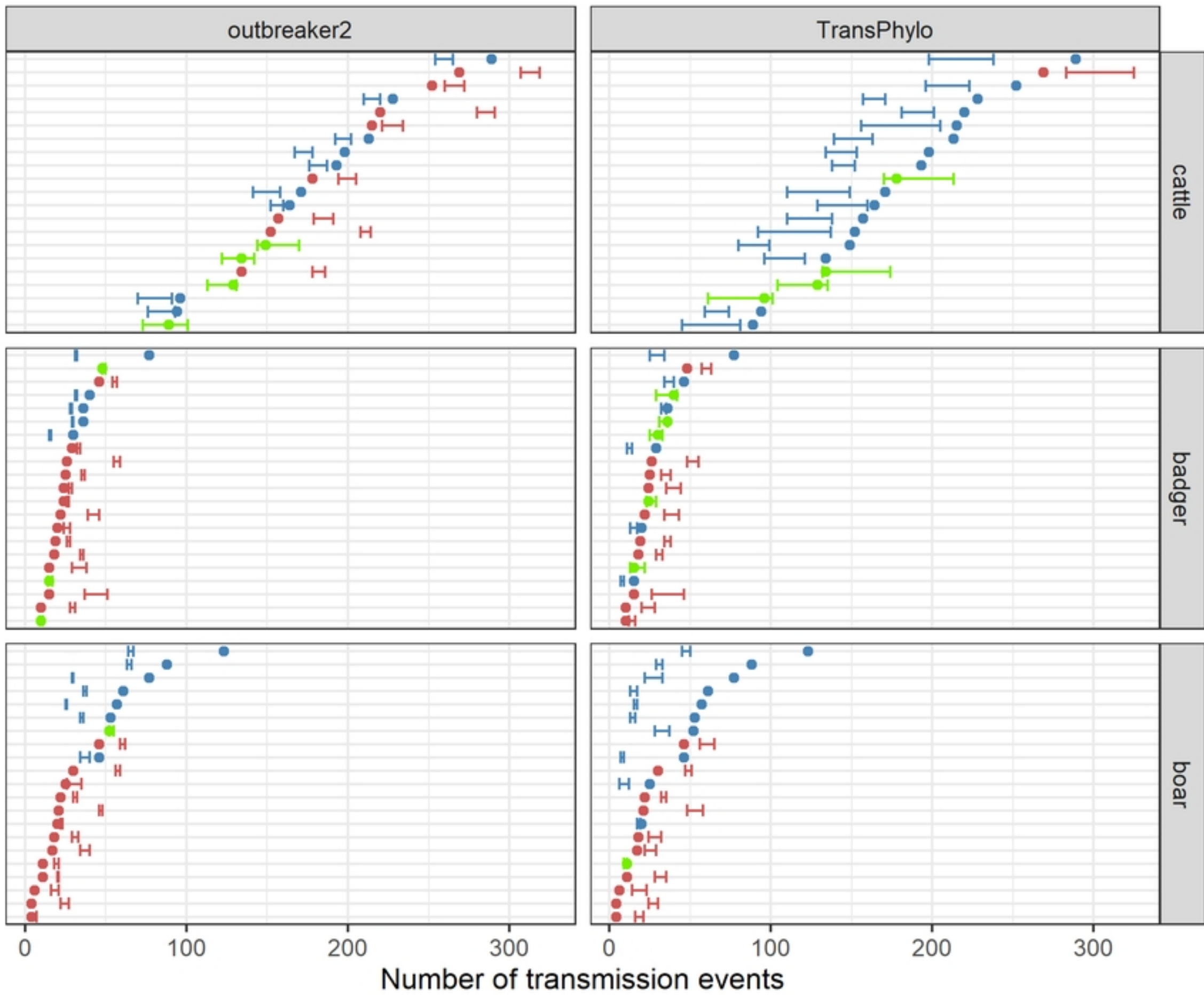


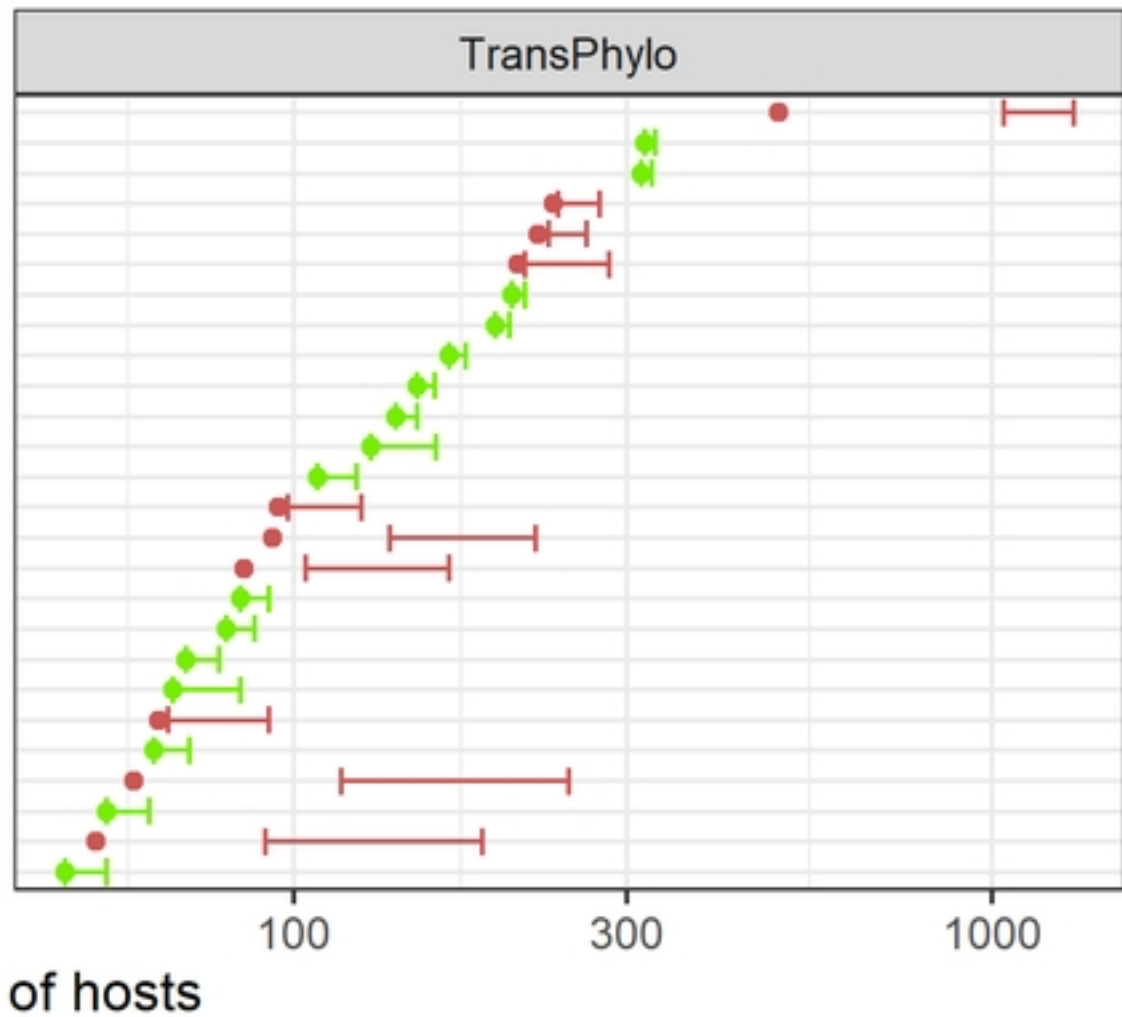
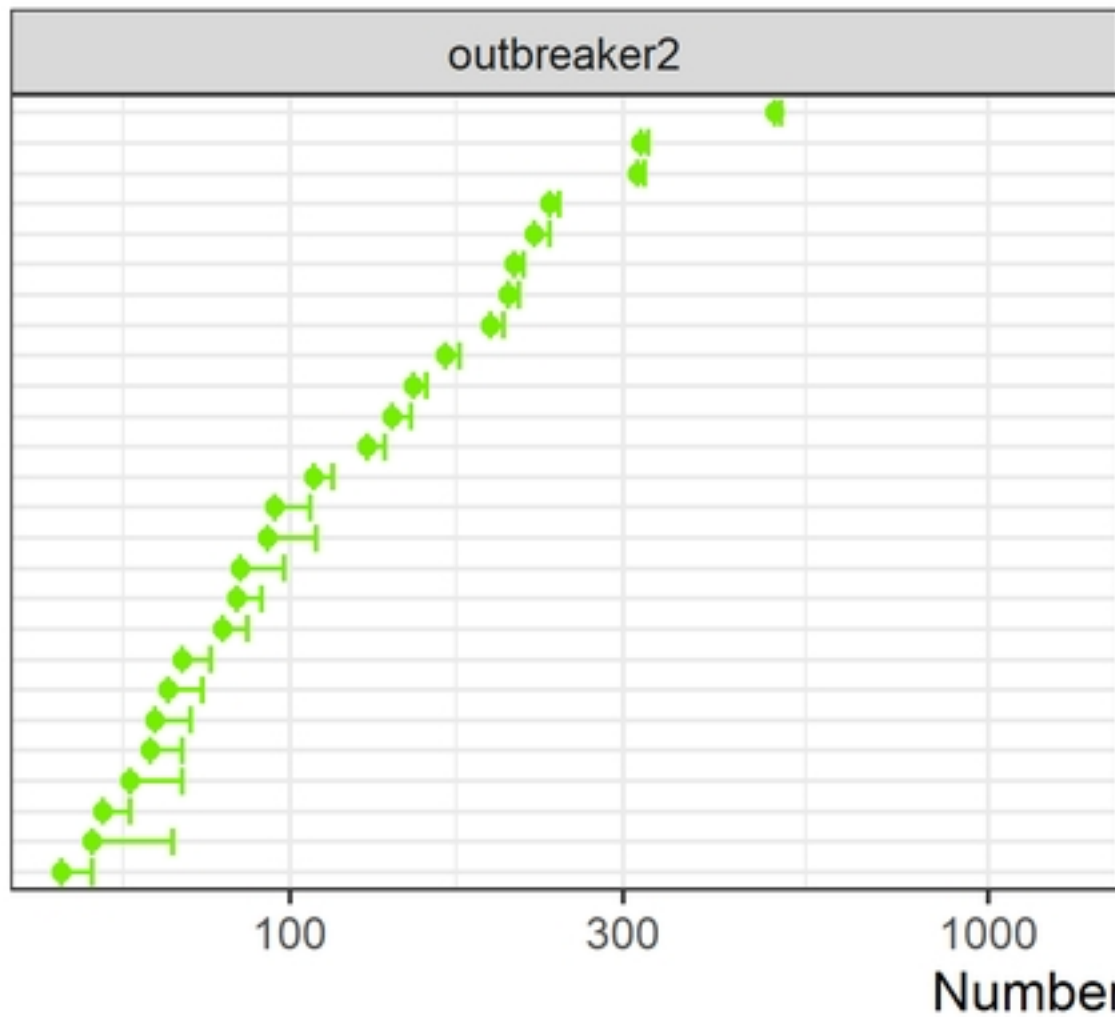
Estimation — Underestimation — Correct — Overestimation



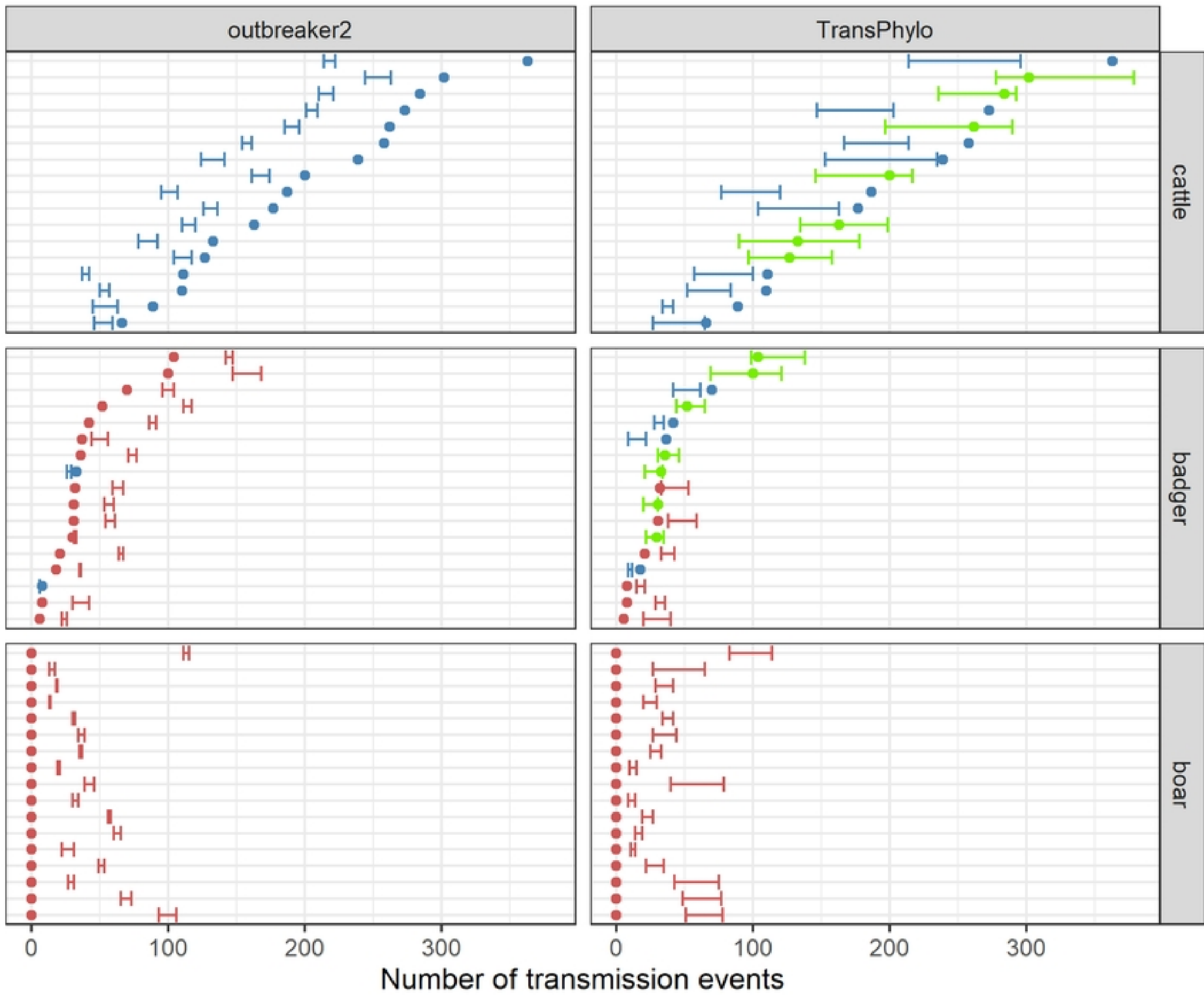


Estimation ● Underestimation ● Correct ● Overestimation





Estimation ● Underestimation ● Correct ● Overestimation



Estimation — Underestimation — Correct — Overestimation