RH: MULTILOCUS DELIMITATION USING ROOTED TRIPLETS

A rapid and scalable method for multilocus species delimitation using Bayesian model comparison and rooted triplets

Tomochika Fujisawa[1], Amr Aswad[2,*], Timothy G. Barraclough[2]

[1] *Department of Zoology, Kyoto University, Sakyo, Kyoto, 606-8502, Japan*

[2] *Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot, Berkshire, SL5 7PY, United Kingdom.*

* Current address*: Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, United Kingdom*

Corresponding author: Tomochika Fujisawa, Department of Zoology, Kyoto University, Sakyo, Kyoto, 606-8502, Japan.

Email: t.fujisawa05@gmail.com

Telephone: +81(0)75-753-4092

Abstract

Multilocus sequence data provide far greater power to resolve species limits than the single locus data typically used for broad surveys of clades. However, current statistical methods based on a multispecies coalescent framework are computationally demanding, because of the number of possible delimitations that must be compared and time-consuming likelihood calculations. New methods are therefore needed to open up the power of multilocus approaches to larger systematic surveys. Here, we present a rapid and scalable method that introduces two new innovations. First, the method reduces the complexity of likelihood calculations by decomposing the tree into rooted triplets. The distribution of topologies for a triplet across multiple loci has a uniform trinomial distribution when the 3 individuals belong to the same species, but a skewed distribution if they belong to separate species with a form that is specified by the multispecies coalescent. A Bayesian model comparison framework was developed and the best delimitation found by comparing the product of posterior probabilities of all triplets. The second innovation is a new dynamic programming algorithm for finding the optimum delimitation from all those compatible with a guide tree by successively analyzing subtrees defined by each node. This algorithm removes the need for heuristic searches used by current methods, and guarantees that the best solution is found and potentially could be used in other systematic applications. We assessed the performance of the method with simulated, published and newly generated data. Analyses of simulated data demonstrate that the combined method has favourable statistical properties and scalability with increasing sample sizes. Analyses of empirical data from both eukaryotes and prokaryotes demonstrate its potential for delimiting species in real cases.

Keywords: Multilocus species delimitation, Bayesian model comparison, Dynamic programming, Bacterial species

INTRODUCTION

Species constitute the basic taxonomic unit for exchanging information about biological diversity. Defining species boundaries in a consistent manner is therefore of major importance to a broad range of biological disciplines. DNA-based delimitation provides a universal method to detect the signature of species existence applicable to various organisms. Consequently, methods to delimit species from DNA sequences alone have been actively developed over the last decade. For early applications of DNA-based delimitation, available markers were limited to a handful of barcoding loci customized for each type of organism (such as *cox1* for animals, Hebert et al. 2003), and therefore delimitation methods were designed to handle these single locus sequences (Pons et al. 2006; Puillandre et al. 2012; Fujisawa and Barraclough 2013; Zhang et al. 2013). However, as the cost of sequencing large amounts of DNA has dramatically decreased, and the ease of developing nuclear markers from genome data has increased, the focus has naturally shifted from single to multiple locus approaches.

There has been huge progress recently in the development of statistical methods for multilocus species delimitation, driven by theoretical advances in the multispecies coalescent model (Rannala and Yang 2003; Degnan and Rosenberg 2009). By comparing alternative delimitation hypotheses and finding the best one based on probability distributions of gene trees under the multispecies coalescent model, species can be delimited robustly even with incomplete lineage sorting. Several methods using Bayesian or information theoretic frameworks have been published so far (O'Meara 2010; Yang and Rannala 2010; Ence and Carstens 2011). Empirical studies have evaluated these methods using taxonomically difficult groups (Carstens and Dewey 2010; Hambäck et al. 2013; Satler et al. 2013). Now, the multispecies coalescent model is becoming a standard for multilocus DNA-based delimitation,

69 and there are attempts to integrate these methods with morphology and geography in order to
70 achieve integrative taxonomy (Fujita et al. 2012; Edwards and Knowles 2014).

71 One drawback of methods based on the multispecies coalescent model is their limited
72 scalability: they rely on the calculation of the probability of obtaining gene trees (or a
73 sequence alignment) given a population tree under the coalescent model, which is relatively
74 time consuming. Also, the joint evaluation of species boundaries and species phylogeny
75 requires searches through an enormous parameter space (Yang and Rannala 2014) , and
76 computation becomes challenging even with small numbers of sampled individuals. Thus,
77 current procedures for multilocus delimitation often require prior assignments of samples to
78 populations, and they are therefore restricted to validation of candidate delimitations based on
79 the assignments. Delimitation without any a priori assignment (species discovery, Ence and
80 Carstens 2011) is feasible only with a limited number of samples, though techniques to reduce
81 search space are being actively studied (Yang and Rannala 2010; Satler et al. 2013). With the
82 increasing ease of sequencing massive multiple nuclear markers (e.g. transcriptome, RAD;
83 Baird et al. 2008; anchored hybrid enrichment; Lemmon et al. 2012), the need for rapid and
84 scalable delimitation methods is becoming more urgent.

85 An alternative strategy for potentially scalable multilocus species delimitation is to use
86 genealogical concordance. The congruence of between-species branching across gene trees
87 reconstructed for separate loci versus incongruence within species has been used as a
88 signature of reproductive isolation and thereby species diversification (Barraclough et al.
89 2003). Early attempts that used topological congruence to detect species included the
90 delimitation of cryptic fungi using concordance of gene trees inferred from five loci
91 (Koufopanou et al. 1997). The "Genealogical Concordance Phylogenetic Species
92 Recognition" (GCPSR, Taylor et al. 2000) is now commonly used to delimit fungal species
93 which often lack morphological or environmental information (Vialle et al. 2013; Millanes et

al. 2014). A disadvantage of using concordance measures between multiple gene trees is that it is hard to treat them under statistical models of evolution. It has been known that a set of multiple gene trees do not necessarily "concord" with each other even if they are generated under the same species tree because of the stochastic nature of the coalescent process. Moreover, the consensus topology of gene trees may not be congruent even with the species tree that generated the gene trees (the anomalous gene tree problem, Degnan and Rosenberg 2006). Thus, the degree of concordance at which one can confidently infer species is not as simple as first perceived. Modelling the distribution of congruence of trees is intrinsically difficult as it must incorporate calculations of the probability of gene trees under a given species tree. Only one non-parametric method with a simulated null model has been devised for statistical delimitation based on topological congruence (O'Meara 2010).

Here, we develop a new method for multilocus species delimitation using gene tree congruence, which employs a likelihood model based on the distribution of triplets. We define a triplet as a partial rooted tree consisting of three tips. Using the distribution of rooted triplets is a promising approach to model congruence between gene trees under the coalescent framework for two reasons. First, the number of triplets with congruent topology is an intuitive measure of topological similarity between trees. Second, the distribution of triplets is readily tractable under the multispecies coalescent framework (Pamilo and Nei 1988) and has been used successfully for rapid inference of phylogenetic trees (Liu et al. 2010). The distribution model for triplet topology is simple and can be extended for intuitive and rapid model-based delimitation. We tested the performance of the new method with various data sets including simulated gene genealogies and both published and newly generated sequence data from both eukaryotes and microbes. The method provides a tractable approach for multilocus delimitation that is scalable to samples with hundreds of individuals across large clades.

METHODS

Calculation of the likelihood of triplet distributions

We employ common assumptions of the multispecies coalescent model (Rannala and Yang 2003; Degnan and Rosenberg 2009): there is neutral random coalescence without structure within species (i.e. panmixia), no gene flow or horizontal transfer between species, and loci evolve independently without intra-locus recombination. In addition, to simplify, we assume initially that the topology of the gene tree is known without error. Under these assumptions, the distribution of triplet topologies is modelled by a simple trinomial distribution as follows.

A bifurcating tree with K tips can be decomposed into $\binom{K}{3}$ rooted triplets. For a given triplet of three individuals, a, b and c, there are three possible topologies, ab|c, ac|b and bc|a. When genealogies from N independent loci are sampled, the numbers of gene trees that conform to each topology - represented by n_1, n_2, and n_3 - are modelled by a trinomial distribution for each triplet. When individuals a, b and c belong to a single species, then under our assumption that the species is panmictic, there is an equal probability of observing each of the three triplet topologies because coalescent events of any pair are equally likely in a panmictic population. Therefore, the distribution of counts of the three topologies is represented by an equiprobable trinomial distribution with likelihood:

$$P_W(n_1, n_2 n_3) = \frac{N}{n_1!\, n_2!\, n_3!}\left(\frac{1}{3}\right)^N$$

(Eq.1)

When individuals are sampled from 2 or 3 distinct species, under the assumptions of the multispecies coalescent process above, the probability of observing triplets congruent with the species tree topology is $1-2e^{-\lambda}/3$, where λ is the length of the internal branch measured by coalescent time units on the species tree, and the probability of observing an incongruent triplet is $e^{-\lambda}/3$ (Pamilo and Nei 1988; Degnan and Rosenberg 2006). Hence, the distribution of triplet counts follows the skewed trinomial distribution.

$$P_B(n_1, n_2, n_3 | v = n_1, \lambda) = \frac{N}{n_1!\, n_2!\, n_3!}\left(1 - \frac{2}{3}e^{-\lambda}\right)^{n_1}\left(\frac{1}{3}e^{-\lambda}\right)^{n_2+n_3}$$

(Eq. 2)

In the equation above, $\nu = n_1$ is the count of triplets congruent with the species tree topology (dominant topology) while n_2 and n_3 denote the counts of incongruent triplets (minority topologies). Note that this distribution does not distinguish the two-species case from the three-species case. Therefore, it is impossible to split a pair of species only represented by two samples but possible to split a species represented by a single sample from species with 2 or more samples.

A Bayesian model comparison framework

In the absence of prior knowledge of the species tree, the observer cannot know a priori which triplet is the triplet concordant with the species tree. Choosing the most frequently observed triplet and using its count as ν in the above equation introduces a bias toward the three-species case and increases the rate of false positives (Supplementary figure

S1, available at Dryad: dryad. 3cb25). We therefore develop a Bayesian model comparison framework to take the unknown species tree into account.

When the species tree is unknown, there are three models that conform to the three-species case described above. We call these three models, $T_B = \{\tau_{b1}, \tau_{b2}, \tau_{b3}\}$, each of which is associated with one of three possible topologies of the underlying species tree. The likelihood functions of the models in T_B are described by P_B in equation (2), with the dominant triplet ν matching n_1, n_2 and n_3 for τ_{b1}, τ_{b2} and τ_{b3}, respectively.

We also consider three models for the case of a single species, following the scheme of Yang and Rannala (2014). We call the set of the three models, $T_W = \{\tau_{w1}, \tau_{w2}, \tau_{w3}\}$. Each model in T_W is again associated with one of the three possible topologies and has its counterpart in T_B (Yang and Rannala 2014). The likelihood functions are P_W in equation (1) and they are identical across models.

With the six candidate models, the joint posterior probability of τ (model) and λ (branch length) given triplet counts X = (n_1, n_2, n_3) is

$$P(\lambda, \tau \,|X) = \frac{P(X|\lambda, \tau)\pi(\lambda)\pi(\tau)}{\int \sum_{\tau \in T_B \cup T_W} P(X|\lambda, \tau)\pi(\lambda)\pi(\tau)\, d\lambda}$$

(Eq.3)

, where π(τ) and π(λ) are posterior probabilities of τ and λ. We obtain the posterior probability of τ by marginalizing the joint posterior by λ.

$$P(\tau|X) = \frac{\int P(X|\lambda, \tau)\pi(\lambda)\pi(\tau) d\lambda}{\int \sum_{\tau \in T_B \cup T_W} P(X|\lambda, \tau)\pi(\lambda)\pi(\tau)\, d\lambda}$$

(Eq. 4)

To simplify the expression for the posterior, we now employ simple uniform priors, $\pi(\tau) = \frac{1}{6}$ and $\pi(\lambda) = {}^{1}/_{L}\ [0 \leq \lambda \leq L]$. We use a prior range up to L=5 throughout this study, which covers a realistic range of frequency of dominant triplets, $0.33 \leq 1\text{-}2e^{-\lambda}/3 \leq 0.996$. The posterior probability of the model with the uniform priors is,

$$P(\tau|\mathrm{X}) = \frac{\int_0^L P(X|\lambda,\tau)d\lambda}{\int_0^L \sum_{\tau\in \mathrm{T}_B \cup \mathrm{T}_W} P(X|\lambda,\tau)d\lambda}$$

(Eq.5)

The integration of P(X| λ, τ) over λ has a tractable analytical solution, therefore a reversible jump MCMC is not required to characterize this posterior distribution. When τ is one of the three models of T_W, the integration over λ is trivial.

$$\int_0^L P(X|\lambda,\tau \in \mathrm{T}_W)d\lambda = L \cdot P_w(X)$$

When τ belongs to T_B, the integration of the likelihood function is represented by the incomplete beta function. When the dominant triplet ν is n_1,

$$\int_0^L P(X|\lambda,\tau \in T_B)d\lambda = \left(\frac{1}{2}\right)^{n_2+n_3} C \int_{\frac{2e^{-L}}{3}}^{\frac{2}{3}} (1-x)^{n_1} x^{n_2+n_3-1} dx$$

$$= \left(\frac{1}{2}\right)^{n_2+n_3} C \left\{\beta\left(\frac{2}{3};\ n_2+n_3, n_1+1\right) - \beta\left(\frac{2e^{-L}}{3};\ n_2+n_3, n_1+1\right)\right\}$$

where β(x; a, b) is the incomplete beta function and C is the multinomial coefficient in equation (2). Replacing n_1 with n_2 or n_3 gives solutions for $\nu = n_2$ or n_3.

The models in T_B are supporting the three-species delimitation, B, (that is, samples are from three distinct species); therefore the posterior probability of the delimitation B is a sum of the three posterior probabilities of the models in T_B.

$$P(B|\mathrm{X}) = \frac{\sum_{\tau \in \mathrm{T}_B} \int_0^L P(X|\lambda,\tau)d\lambda}{\sum_{\tau \in \mathrm{T}_B} \int_0^L P(X|\lambda,\tau)d\lambda + 3L \cdot P_W(X)}$$

(Eq.6)

, and the posterior probability of the single-species case delimitation, *W*, (samples are from a single species) is,

$$P(W|\mathrm{X}) = \frac{3L \cdot P_W(X)}{\sum_{\tau \in \mathrm{T}_B} \int_0^L P(X|\lambda,\tau)d\lambda + 3L \cdot P_W(X)}$$

(Eq.7)

With a given hypothesis of delimitation, each of the $\binom{K}{3}$ triplets is assigned to one of the two categories defined above, i.e. a, b and c either belong to the same species or to multiple species. The overall posterior probability of a given delimitation for all K taxa is the product of the posterior probabilities of all triplet counts of two categories. For a set of triplet counts, $\boldsymbol{w}$, which is assigned to delimitation *W*, and a set $\boldsymbol{b}$, which is assigned to delimitation *B*, the log-posterior probability of a delimitation *D* is as follows.

$$\log P(\mathrm{D}|\mathrm{X}) = \sum_{(n_1,n_2,n_3)\in \boldsymbol{w}} \log P(W|n_1,n_2,n_3) + \sum_{(n_1,n_2,n_3)\in \boldsymbol{b}} \log P(B|n_1,n_2,n_3)$$

(Eq.8)

We use this quantity as the posterior probability score. Note that it is not a true posterior probability of delimitation since it ignores the mutual dependence of the parameters of the triplet distribution caused by overlapping membership of some triplets. However, the similar approximation of likelihood functions has been used successfully in statistical phylogenetic inference (Liu et al. 2010) and we test its performance by simulation here.

Finding the best delimitation model

The posterior probability score described above is used to find the optimal delimitation from a set of delimitations of samples. The number of all possible delimitations of K samples is represented by the Bell number, $\sum_{i=1}^{K}\left\{{K \atop i}\right\}$ (Bell 1934), where $\left\{{K \atop i}\right\}$ is a Stirling number of second kind, defined as the number of all possible ways to split K items into i groups. This number is common to partitioning problems and intractably large. An approach taken to reduce the number of delimitations considered is using a guide tree (Yang and Rannala 2010), which gives a hierarchical structure of multiple delimitations. Different combinations of splitting and lumping of lineages on a given guide tree are searched to find the best delimitation. Conventional search methods with the guide tree approach use either reversible-jump Markov chain Monte Carlo for characterizing posterior probabilities of competing delimitations (Yang and Rannala 2010, 2014) or heuristic search algorithms to find the optimal combinations of splits and lump of lineages (O'Meara 2010; Satler et al. 2013).

We now consider only the problem of finding the best delimitation on a fixed guide tree without tree rearrangement. The number of all possible delimitations under a given guide tree with S tips is approximately$\lfloor 1.5^{S} \rfloor$ (floor of 1.5^{S}) in the worst case and S in the best case (Fujisawa and Barraclough 2013). The size of the space is reduced compared with the Bell number, but still grows exponentially with the number of species in the worst case. We developed a new dynamic programming algorithm to rapidly find the best combination of lineages on the guide tree, taking advantage of the optimal substructure of the likelihood model and the guide tree structure.

Given a delimitation under a guide tree, D, its posterior probability score can be decomposed into a sum of the scores of the delimitations of two subtrees descending from the root, $\log P(D_L)$ and $\log P(D_R)$, and a constant factor since equation (8) is additive.

$$\log P(D) = \log P(D_L) + \log P(D_R) + c$$

$$c = \sum_{(n_1,n_2,n_3)\in \boldsymbol{b_{root}}} \log P(B|n_1, n_2, n_3)$$

(Eq.9)

where *c* is a constant representing a score for triplets crossing over 2 subtrees descending from the root node of the guide tree. Since triplets are not shared between subtrees and their posterior probabilities are independently calculated, the optimal solutions for each subtree must be included in the global solution. Therefore, finding the global optimal solution can be reduced to finding solutions to subtrees' delimitations, and iteratively solving and combining them yields the global solution. An exception is the case where logP(D) is represented by the root of the guide tree; that is, all samples are from the same species. In the case of root delimitation, equation (9) does not hold because the constant of the third term must be represented by $P(W|X)$ not $P(B|X)$. So, the dynamic programming algorithm must compare the "root" delimitation with the aggregated solution of subtrees in each step. This leads to the algorithm described in Figure 1 and supplementary text S1 (Online Appendix). This algorithm calculates the global optimal posterior probability score from a guide tree, and the best delimitation was obtained by keeping the set of nodes producing the best score. The algorithm reduces the number of likelihood calculations to twice the number of the nodes on the guide tree.

255 We implemented the method in a program called "tr2" (Trinomial distribution of
256 Triplets) for calculation of posterior probability scores for delimitation hypotheses and the
257 search algorithm for the best delimitation given a guide tree. The program is implemented in
258 Python and can run on any operating system (Distributed at
259 https://bitbucket.org/tfujisawa/tr2-delimitation, archived version for paper is available at
260 Dryad: dryad.3cb25).

261

262 *Simulations and case studies*

263 We used simulated and real gene trees to test the performance of the method. First, we
264 performed coalescent simulations with species trees with 3 and 10 tips. Then, we analyzed a
265 published data set of rattlesnakes with 29 individuals and a newly sequenced data set of 144
266 *Bacillus cereus* isolates.

267

268 *Three-species simulations*.— In order to test the performance of the delimitation
269 model, we first conducted a simple 3 species simulation and assessed the error rates of the
270 model. In this simulation, we assume gene trees are known without error. Gene genealogies
271 were simulated within a species tree with three tips and fixed branch lengths, T_1=4000 and
272 T_0=8000 generations (Figure 2a). The number of samples per species was set to 10, totalling
273 30 individual samples. The effective population sizes were set to 1/2*T_1 to 8*T_1 for all
274 species (T_1=1/8Ne – 2Ne generations). Coalescent trees within the species tree were
275 simulated using SIMCOAL (Excoffier et al. 2000) assuming that one species represents one
276 population and populations merge on speciation events. Custom scripts were used to generate
277 input files for SIMCOAL from species trees. Twenty-five independent loci were simulated

100 times, which resulted in 2500 gene trees in total. The posterior probability for a global delimitation *W* (all individuals are from a single species) and the three alternative models representing correct delimitation (a), over-splitting (b) and under-splitting (c) (Figure 2) were calculated with increasing numbers of loci between 5 and 25 with step 5. Error rates, i.e., the frequency of choosing an incorrect model as the best model, were recorded for each iteration.

Ten-species simulations.— The ten-species simulation considers more realistic conditions. Species trees with 10 tips were simulated under the Yule model with a constant speciation rate. The total depth of species trees, T, was rescaled to 20000 generations, and the effective population size (Ne) of species was set as 1, 1/2, 1/4, 1/8 and 1/16 times of T (Ne=1250-20000). These parameter settings cover speciation rates and effective population sizes observed in various eukaryotic groups (Coyne and Orr 2004; Charlesworth 2009) including extreme cases of rapid radiations. Gene trees with 10 samples per species (100 total samples) for 40 independent loci were simulated using SIMCOAL and the custom scripts. Simulations were replicated 100 times.

In the first simulation, hereafter simulation A, we assume that the topology of the guide tree and assignment of terminals to species groups is known. This simulation tests whether the method can correctly find the positions of nodes which define species from multiple competing combinations on a guide tree. The tr2 program was run with the species tree as a guide and simulated gene trees as inputs. The number of loci used ranged from 5 to 40.

In the second simulation, B, delimitation was conducted solely from sets of gene trees (species discovery approach). A consensus tree was built from gene trees from multiple loci using the rooted triple consensus (Ewing et al. 2008). Then, the consensus tree was used as

302 the guide tree in the delimitation step. This guide tree contains all possible hierarchical
303 delimitations, from each individual representing a separate species to all individuals
304 representing a single species. Polytomies on consensus trees were randomly resolved by the
305 "multi2di" function in the "ape" package (Paradis et al. 2004). In addition, we performed a
306 set of simulations to assess the effect of increasing numbers of loci and individual samples.
307 Gene trees were simulated within the same species trees as above with Ne=T/4, but the total
308 number of samples was reduced to 50 (5 per species) and the number of loci was doubled,
309 keeping the total sample size (number of loci X number of samples) constant. Delimitation
310 with tr2 was conducted in the same procedure as simulation B.

311 The third simulation, C, considers conditions where gene trees and species trees are
312 estimated from DNA sequences. Sequences were simulated along the branches of the gene
313 trees simulated above using Seq-Gen (Rambaut and Grassly 1997) assuming HKY+G model
314 (Ts/Tv = 2.5 and $\alpha = 0.1$) and 3% of overall genetic variations. These parameters were chosen
315 to be comparable to the case studies described in the next sections. Sequence length was set to
316 a constant length of 750bp. Gene trees were reconstructed from the simulated sequences using
317 RAxML with a GTR+G model (Stamatakis 2014) and rooted by the "-I f" option of RAxML .
318 Guide trees were estimated from the reconstructed gene trees with the rooted triple consensus,
319 and delimitation was conducted with tr2. Under the parameter settings above, within-species
320 genetic variation of simulated sequences ranged from 0.3% to 1.4% depending on Ne, and
321 between-species variation was 3.0%.

322 The number of estimated species and the number of exact matches between estimated
323 and true species were measured as the accuracy of delimitation. The elapsed time for each
324 trial was also recorded. The numbers of non-monophyletic species were counted to measure
325 the degree of incomplete lineage sorting. The effects of Ne, the number of loci, simulation
326 type (A, B and C) and their second interaction terms on the proportion of exact matches were

327 tested using GLM. For simulation B, the effect of the two sampling strategies was also tested.
328 Simulations and delimitations were run on a Linux personal computer with a 2.3 GHz Intel i5
329 quad-core processor and 4GB memory.

330

331 *Case study one: Sistrurus Rattlesnakes*.— Kubatko et al. (2011) sampled 18 nuclear
332 loci and one mtDNA locus of *Sistrurus* rattlesnakes. The data set of the nuclear loci included
333 58 phased sequences from 29 individuals of six known subspecies of *S. catenatus*/*S. miliarius*
334 and two outgroups. Kubatko et al. (2011) reported that one subspecies, *S. catenatus catenatus*,
335 exhibited signatures of a distinct species status while the other five subspecies did not show
336 significant evidence of independent species based on the monophyly-based test described by
337 Rosenberg (2007). We reanalyzed this data set. The gene trees and an alignment matrix of 18
338 nuclear loci were downloaded from TreeBase (accession:TB2:S11174). The trees were
339 randomly resolved with "multi2di". Then, a consensus tree was built using the rooted triple
340 consensus from them, and the best delimitation was determined with the consensus as the
341 guide tree. A re-sampling procedure of loci was conducted by progressively adding single loci
342 in random order. Polytomies were randomly resolved in each iteration. The re-sampling was
343 repeated 50 times to characterize the effect of increasing number of loci on the delimitation.
344 Genetic variation within subspecies was 0.2% and between species 2.2%.

345

346 *Case study two: Bacillus Multilocus Sequence Typing*.— We tested the applicability of
347 the tr2 method to bacterial species using a multilocus sequence typing (MLST) data set of the
348 *Bacillus cereus* complex. Multilocus sequence typing (MLST) is a typing scheme for bacterial
349 species/subspecies using a few (typically seven) loci (Maiden et al. 1998; Maiden 2006). It is
350 widely used in clinically relevant bacteria and occasionally in environmental prokaryotes to

351 delimit species (e.g. Papke et al. 2007). Although bacterial reproduction is largely clonal, in
352 many bacteria including *Bacillus cereus*, genetic exchange also occurs (Vos and Didelot
353 2009). If there was frequent gene exchange within a group of closely related individuals, but
354 none between distantly related groups, this could lead to units equivalent to reproductively
355 isolated species in sexual eukaryotes (Didelot et al. 2011; Barraclough et al. 2012).The tr2
356 method should be able to delimit such a group as a putative species. However, in clonal
357 bacteria without any recombination, the delimitation method based on gene tree congruence
358 would delineate all individuals as separate species because the true genealogy of each locus
359 would be identical. Another complication is that horizontal transfer might occur rarely
360 between otherwise distinct species. This could introduce additional incongruence among loci
361 between otherwise separate species. We were interested to see how the method coped with a
362 prokaryotic clade that might display these complications.

363 Our sample comprised 144 isolates originally collected from evenly spaced quadrats
364 in the walled garden at Silwood Park for the study by Collier et al. (2005). In brief, freezer
365 isolates were regrown on *B. cereus* selective agar and DNA extracted using Chelex Instagene
366 matrix method. The 7 house-keeping genes used for standard *B. cereus* MLST (Jolley et al.
367 2004) were PCR amplified and Sanger sequenced using primers and conditions at the MLST
368 database (http://pubmlst.org/bcereus/info/primers.shtml). Sequences were edited in Geneious
369 and trimmed to the lengths used at the MLST database. Full details are provided elsewhere
370 (Collier et al. 2005; Barraclough et al. in preparation), and sequences are available at Genbank
371 (Accession: KT806485-KT807462).

372 Alignment lengths of the MLST sequences ranged from 348 to 504bp, and there were
373 29 to 55 unique haplotypes at each locus (maximum of 55 for purH and minimum of 29 for
374 glpF and gmk). The complete data matrix excluding missing loci contained 2806 bp from
375 each of 114 isolates, which included 99 unique multilocus sequence types. Overall genetic

376 variation was 4.0%. Sequences from the seven loci were separately aligned with MUSCLE
377 3.8 (Edgar 2004). Gene genealogies of the seven loci were estimated using BEAST 1.80
378 (Drummond et al. 2012). Ten million generations of MCMC sampling were run with a
379 GTR+G substitution model and the log-normal relax clock model (Drummond et al. 2006).
380 Twenty percent of the MCMC samples were discarded as burn-in. The convergence of the
381 parameters was checked by effective sampling size using Tracer (Rambaut and Drummond
382 2007), and the maximum clade credibility trees (MCC trees) were extracted from the MCMC
383 runs using TreeAnnotator.

384 Two methods were used to obtain guide trees for delimitation of the *Bacillus* group.
385 First, a consensus tree was constructed using the rooted triple consensus from the MCC trees
386 of seven loci. Second, in order to account for the effects of horizontal transfer on the guide
387 tree estimation, we ran ClonalFrame (Didelot and Falush 2007) on the concatenated
388 alignment. ClonalFrame estimates the most likely clonal genealogy by removing putative
389 horizontally transferred regions. An MCMC of ClonalFrame was run with 800 thousand
390 generations, and 50% of the chain was discarded as burn-in. Convergence of parameters was
391 examined by checking effective sample size using Tracer. The 50%-majority consensus from
392 the ClonalFrame MCMC was used as a second guide tree. Re-sampling of loci was
393 conducted 50 times using these two guide trees. To further account for the uncertainty of tree
394 building, 100 trees were sampled from the MCMC chain from BEAST for each locus and
395 from the chain of ClonalFrame, and delimitation was repeated with these 100 sets. The
396 frequency for each pair of samples to be grouped in the same species was recorded. (Sequence
397 alignments and trees are available at Dryad: dryad. 3cb25)

398

399 RESULTS

Three-species simulations

The overall false positive rate (FPR, rate of over-splitting) in the three-species simulations is 0.0 in all iterations with all numbers of loci between 5 and 25. False negative rates (FNR, rate of under-splitting) decrease as the number of loci used increases (Figure 3). FNR of less than 30% were attained with only 5 loci when Ne was 2000 and 4000 (equivalent to T=2Ne and Ne) whereas the FNRs reached 30% with 20 loci when Ne was 8000 (T=1/2Ne). With larger Ne values, the decrease of FNRs was much slower, and the method was not able to correctly delimit species within the range of loci used in the simulations when Ne was 16000 and 32000 (T=1/4Ne-1/8Ne). The average time required for one trial was 0.5 seconds.

Ten-species simulations

When true species trees are given as the guide tree, the method appeared to delimit species consistently. The proportion of exact matches increased with the number of loci used (Figure 4, A), and the number of estimated species approached the true number of species, 10 (Figure 5, A). With low Ne value (Ne=1250), the median number of exact matches reached 10 when 25 or more loci were used. The increase in the number of exact matches slowed down with larger Ne values, for example, when Ne≥5000, 40 loci were not enough to attain 100% exact matches.

The accuracy was slightly reduced when the guide trees were estimated by the consensus method (Figure 4 and 5, B). However, the effect of simulation type was not significant ($z = -0.33$, $p = 0.74$ for simulation type, GLM with binomial errors) while Ne and the number of loci were highly significant ($p << 0.001$ for both Ne and the number of loci). In addition to the under-split observed in the simulation A, a few oversplits occurred especially when the number of loci was small. In 0.9% of trials, the method estimated more than 10 species. Overall accuracy still increased when more loci were added. When the gene trees

were estimated from the simulated sequences, the accuracy further decreased, especially when the number of loci was small (Figure 4, C). The accuracy was significantly lower than other simulation types ($z = -4.42$, $p << 0.001$, GLM with binomial error). Even more frequent oversplits were observed: the number of trials with >10 estimated species reached 2.0%. (Figure 5, C).

The time required for a delimitation process increased nearly linearly with the number of loci (Supplementary Figure S2). Median time ranged from 23 to 47 seconds for 10 tip guide trees and from 135 sec to 162 seconds for guide trees with 100 tips. Average proportions of non-monophyletic species were between 0.34 for Ne=1250 and 0.97 for Ne=20000 (Supplementary Figure S3), indicating non-monophyly is prevalent even for small Ne values. The accuracy of delimitation was significantly lower when fewer loci and more samples were used ($z=-3.27$, $p=0.001$, GLM with binomial error, Supplementary Figure S7).

Rattlesnakes

The method delimited 4 putative species of the *Sistrurus* rattlesnakes, including two ingroup and two outgroup species (Figure 6 Left, Supplementary Figure S4). The two ingroup species matched with the known taxonomic species, *S.catenatus* and *S.miliarius*. Random resampling of loci indicated the number of estimated species does not saturate within the range of loci used in this study (Figure 6 Left). With 18 loci, 28% of repeated delimitations split *S.catenatus* into two groups: one group exclusively consisted of a subspecies *S.c.catenatus* and another group consisted of *S.c.edwardsii* and *S.c.tergeminus*. Three subspecies of *S.miliarius* were always grouped together into a single species.

447 *Bacillus MLST*

448 Delimitation using the seven MCC trees and rooted triple consensus tree resulted in 7 putative
449 species while the delimitation with ClonalFrame consensus resulted in 11 species. The
450 majority of nodes on the rooted triple consensus were unresolved (Supplementary Figure S5).
451 ClonalFrame robustly recovered three clades, two of which were unresolved in the rooted
452 triple consensus (Clade A, B and C in Figure 7). The difference between the two approaches
453 is consistent with horizontal transfer affecting topologies deeper in the tree; we mainly focus
454 on the result of delimitation using the ClonalFrame guide tree. Re-sampling of loci showed
455 that there was substantial variation in the number of estimated species (Figure 6 Right): the
456 sample of 7 loci might be too few for robust delimitation in this case. Repeated delimitations
457 run on 100 sets of MCMC tree samples exhibited 18 species that were consistently delimited
458 (Figure 7). While clades A and B were grouped into three or four large clusters, clade C was
459 more frequently separated into small singleton species. Frequencies for isolates to be grouped
460 in species with other isolates within these clades were on average 61% and 40% for clade A
461 and B and 35% for clade C.

462 We estimated linkage disequilibrium (LD) within subsets of these groups to test for
463 variation in recombination rate. Samples were taken from within the largest clusters in clade
464 A and B respectively, and randomly from within clade C, and LD of variable sites was
465 calculated for each group by the "LD" function of an R package "pegas" (Paradis 2010). The
466 test calculates the correlation between pairs of variable sites (Zaykin et al. 2008). There are
467 distinctive linkage patterns between and within the seven loci in the three groups
468 (Supplementary Figure S6). In clade A and B, strong to moderate LD within each locus and
469 LD between a few pairs of loci were observed, but LDs between loci were small (Median
470 within-locus R^2=0.49 and 0.16 and Median between-locus R^2 = 0.07 and 0.08 for clades A
471 and B, respectively). This is consistent with recombination among separate loci, but linkage

472 within loci. On the other hand, in clade C, there were moderate or high levels of LD between
473 most loci (Median within-locus R^2=0.38 and between-locus R^2=0.29 for clade C), consistent
474 with low rates of recombination even between loci.

475

476 DISCUSSION

477 Congruence between gene trees provides intuitive and readily tractable statistical models for
478 multilocus species delimitation. In this paper, we developed a method to delimit species based
479 on topological congruence or incongruence of triplets quantified by two types of trinomial
480 distribution models. These models were derived from the multispecies coalescent framework
481 and can be used for robust delimitation of species from gene trees with incomplete lineage
482 sorting. The simulation studies confirmed that the method can consistently delimit species
483 without monophyly, and its performance increased with the number of loci and decreased
484 with larger effective population size relative to divergence time.

485 The accuracy of the method is slightly lower than the reported performance of
486 conventional multilocus delimitation methods (Camargo et al. 2012); more than 25 loci were
487 required to delimit with 95% success rate under the condition T=0.5*Ne (Figure 2) while
488 Camargo et al. (2012) reported 60 - 100 % success with 10 loci by conventional methods. The
489 advantage of tr2 appears to be its speed and applicability to large data. According to
490 Camargo et al. (2012), with a four species guide tree, SpedeSTEM (Ence and Carstens 2011)
491 ran in 30 seconds with 20 samples and BP&P (Yang and Rannala 2010) with 80 samples in
492 6.5hrs. The order of speed of the tr2 (30 seconds with 100 samples and 10 species with known
493 gene trees) is comparable to the fastest conventional method. When a sequence alignment is
494 used, additional time for tree reconstruction is required (e.g. approximately 1 minute per locus
495 by RAxML), but the reconstruction - delimitation procedure can still scale to large data sets.

496 In addition, the dynamic programming algorithm finds the global solution on a given guide
497 tree while most heuristic optimizations do not guarantee it. The method was sufficiently
498 conservative to over-splitting, which is a favorable property for DNA-based species
499 delimitation methods (Carstens et al. 2013).

500 The simulation studies also showed that the accuracy of the guide tree is crucial for
501 accurate multilocus delimitation. It has been known that the incorrect assignment of samples
502 on guide trees results in oversplits of species in multilocus delimitation (Leaché and Fujita
503 2010; Zhang et al. 2014). The oversplits observed in the discovery approach (simulation B
504 and C) are likely to have resulted from the incorrect placement of samples on guide trees.
505 However, except for the excess of oversplits, the effect of unknown guide tree was minimal.
506 The number of exact matches was not significantly different between known and unknown
507 guide tree simulations, and even when DNA sequences were used, accuracy was comparable
508 to the other simulations with a sufficient number of loci. It appears that, when the consensus
509 species tree estimation can resolve a particular node on a guide tree, tr2 does not erroneously
510 merge or split species on the node. This is a useful property since there are discrepancies
511 between the number of loci required for correct delimitation, guide tree estimation and initial
512 population assignment in the conventional delimitation procedures (Zhang et al. 2014). The
513 inaccurate estimate of guide tree and delimitation may be mediated simply by adding more
514 loci as the number of loci is not a major computational obstacle.

515 The delimitation results for *Sistrurus* rattlesnakes were partially consistent with the
516 reported results in Kubatko et al. (2011). Though the two known taxonomic species,
517 *S.catenatus* and *S.miliarius*, were consistently delimited as putative species, only about 30%
518 of iterations supported the distinctiveness of the subspecies *S.c.catenatus*. Considering the
519 number of loci necessary to delimit species in the simulation studies, 18 nuclear markers
520 appear to be insufficient to fully delimit this group with the present method. The resampling

521 also indicates that polytomies are an important source of uncertainty on delimitations. The
522 two alternative outcomes with 18 loci resulted solely from the different resolutions of
523 polytomies. The lack of mutations and resulting polytomies do not positively mislead the
524 delimitation when the identical sequences are randomly inserted or polytomies are randomly
525 resolved. Nevertheless, simulations and case studies show the use of uninformative loci
526 compromise the power of species detection and introduce uncertainty. We used repeated
527 delimitations with randomly resolved gene trees and guide trees, and this approach was able
528 to capture the level of uncertainty of gene tree reconstruction in the rattlesnakes. Resampling
529 trees from bootstrap trees or MCMC runs, as done in the *Bacillus* data set, is an alternative
530 way to handle the uncertainty.

531 The results of re-sampling analysis of *Bacillus* complex indicate more uncertainty in
532 their delimitation than the rattlesnakes. The reduced number of species observed on the rooted
533 triple consensus may partly result from the unresolved guide tree due to horizontal transfer
534 between distantly related groups. However, distinctive patterns of bacterial diversification
535 were still observed. Clade A and B were consistently delimited into large groups while clade
536 C mainly consisted of weakly connected singletons. Samples from these two categories
537 exhibited a contrasting pattern of linkage disequilibrium patterns. Especially, low LD between
538 loci observed in samples from the largest clusters detected in clades A and B indicates that
539 there is frequent gene exchange between members of those groups. Homologous
540 recombination creates local topological discordance on bacterial genomes (Didelot et al.
541 2010), and if the recombination events are localized only within closely related groups, the
542 mutually recombining groups can be detected by tr2 through genealogical discordance. The
543 clusters delimited in the clades A and B are likely to be such groups. Clade C has low
544 recombination rates and methods based on recombination and gene congruence are
545 inappropriate. It may still be possible to identify independently evolving groups in such clades

546 using alternative concepts and methods developed for clonal bacteria and asexuals (Cohan
547 2001; Barraclough et al. 2003). Clearly, the mixture of high and low-recombining lineages in
548 the *Bacillus* data adds complexities to species delimitation (which we will address in detail
549 elsewhere) and the number of loci may not be large enough to fully elucidate diversification
550 patterns. However, the result demonstrates the potential for detecting 'recombinationally
551 isolated' groups in prokaryotes.

552 The parameters to be considered for the computational complexity of delimitation are
553 the number of samples (K), the number of species (S) and the number of loci (N). The
554 dynamic programming algorithm introduced in this paper finds the best delimitation and
555 reduces the complexity of search through a guide tree to time scale linear to S, $O(S)$, which
556 allows a thorough search of a guide tree. For example, using a guide tree that assigns every
557 individual into a distinct species has often been prohibitive with large samples, but, in our
558 simulations, tr2 was able to process guide trees with 100 tips within 150 seconds. Combined
559 with good performance with respect to other parameters - cubic dependency of time on
560 overall sample size, $O(K^3)$ and linear for loci, $O(N)$ - the method could be used to provide a
561 rapid search method through candidate delimitation hypotheses before applying more
562 statistically rigorous methods to large datasets. Current next generation sequencing projects
563 often target a large number of loci from relatively few individuals. The tr2 method is suitable
564 for this type of sampling design since the impact of increasing loci on computations is smaller
565 than increasing individuals. A simulation shows that higher accuracy is achieved with more
566 loci than with more individuals when total sample size (loci X individuals) is fixed
567 (Supplementary Figure S7). This demonstrates a potential use of current sequence
568 technologies for species delimitation though the optimal sampling strategy is yet to be
569 investigated. A final point is that the dynamic programming algorithm introduced in this
570 study may be applied for other optimal partitioning problems using hierarchical structure,

571 such as finding optimal partitioning of sequence alignments for phylogenetic inference (Li et
572 al. 2008; Lanfear et al. 2012).

573 In this study, we did not consider possible violations of the assumption of the
574 multispecies coalescent model including gene flow between populations. Gene flow between
575 sister species reduces the number of dominant triplets and increases two minority triplets
576 equally, and may compromise the accuracy of the method. Incorporating branch lengths into
577 the model may be required to tease apart the effects of gene flow and incomplete linage
578 sorting. Introgression events from distantly related groups may be detected as an increase in
579 one of two minority triplet counts. Indeed, deviation from equal counts of minority triplets is
580 used for tests of introgressive gene flow (Durand et al. 2011; Zwickl et al. 2014). Violation of
581 the model assumption of panmixia could be detected in a similar manner by extending the
582 trinomial distribution model used in this study to a three-rate model.

583 The method now uses estimated gene trees as inputs. In addition, it uses a guide tree
584 estimated from the given gene trees or other independent methods. This procedure does not
585 take the uncertainty of gene tree and species tree inference into account. Also, most
586 computational time of the delimitation procedure was spent on the tree-building steps (a
587 BEAST run on 1 locus of *Bacillus* took 2.5hrs while the tr2 ran in 2 minutes with > 100
588 samples). For guide tree inference, one possible solution would be to incorporate joint
589 inference of species tree and delimitation using triplets. Triplet- or quartet-based phylogenetic
590 inference methods using known gene trees under the multispecies coalescent framework have
591 been developed and implementations to handle large datasets already exist (Liu et al. 2010;
592 Mirarab et al. 2014). The delimitation step based on the trinomial distributions could be easily
593 integrated into these procedures. Also, gene tree inference could be bypassed by directly
594 counting triplets estimated from 3 corresponding sequences and an outgroup as done in some

phylogenetic inference programs (DeGiorgio and Degnan 2010). Combining these methods could potentially lead to a highly scalable joint estimation of species tree and delimitation.

In conclusion, we present a method for species delimitation from multilocus data that can potentially scale to the kind of sample sizes that are currently only feasible for single-locus approaches. The method uses exact methods derived from the multispecies coalescent, but by splitting the problem into triplets it circumvents the computational challenges. As it becomes easier to sequence non-model genomes, and consequently to assay variable nuclear markers across clades, we envisage a growth in the number of studies using standardized multiple unlinked markers across entire clades, equivalent to current DNA barcoding sample regimes. Our method is designed with these scenarios in mind to complement more intensive methods.

AUTHOR CONTRIBUTIONS

TF devised the new methods, wrote the software, ran the analyses and wrote the manuscript. AA generated the *Bacillus* sequence data. TGB provided advice on the methods, helped with some analyses and wrote the manuscript.

ACKNOWLEDGEMENTS

We thank Richard Ellis for providing the frozen *Bacillus* samples and Yumi Moon and Kevin Balbi for help with that project. TGB and AA were supported by BBSRC grant BB/G004250/1.

617 REFERENCES

618 Baird N.A., Etter P.D., Atwood T.S., Currey M.C., Shiver A.L., Lewis Z.A., Selker E.U.,
619 Cresko W.A., Johnson E.A. 2008. Rapid SNP discovery and genetic mapping using
620 sequenced RAD markers. PLoS One. 3:e3376.

621 Barraclough T.G., Balbi K.J., Ellis R.J. 2012. Evolving Concepts of Bacterial Species. Evol.
622 Biol. 39:148–157.

623 Barraclough T.G., Birky JR. C.W., Burt A. 2003. Diversification in sexual and asexual
624 organisms. Evolution (N. Y). 57:2166–2172.

625 Bell E.T. 1934. Exponential Numbers. Am. Math. Mon. 41:411–419.

626 Camargo A., Morando M., Avila L.J., Sites J.W. 2012. Species delimitation with ABC and
627 other coalescent-based methods: a test of accuracy with simulations and an empirical
628 example with lizards of the liolaemus darwinii complex (Squamata: Liolaemidae).
629 Evolution (N. Y). 66:2834–2849.

630 Carstens B.C., Dewey T.A. 2010. Species delimitation using a combined coalescent and
631 information-theoretic approach: an example from north american myotis bats. Syst. Biol.
632 59:400–14.

633 Carstens B.C., Pelletier T.A., Reid N.M., Satler J.D. 2013. How to fail at species delimitation.
634 Mol. Ecol. 22:4369–83.

635 Charlesworth B. 2009. Effective population size and patterns of molecular evolution and
636 variation. Nat. Rev. Genet. 10:195–205.

637 Cohan F.M. 2001. Bacterial species and speciation. Syst. Biol. 50:513–524.

638 Collier F.A., Elliot S.L., Ellis R.J. 2005. Spatial variation in Bacillus thuringiensis/cereus
639 populations within the phyllosphere of broad-leaved dock (Rumex obtusifolius) and
640 surrounding habitats. FEMS Microbiol. Ecol. 54:417–425.

641 Coyne J.A., Orr H.A. 2004. Speciation. Sunderland, MA.: Sinauer Associates. Inc.

642 DeGiorgio M., Degnan J.H. 2010. Fast and consistent estimation of species trees using
643 supermatrix rooted triples. Mol. Biol. Evol. 27:552–69.

644 Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene
645 trees. PLoS Genet. 2:e68.

646 Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the
647 multispecies coalescent. Trends Ecol. Evol. 24:332–40.

648 Didelot X., Bowden R., Street T., Golubchik T., Spencer C., McVean G., Sangal V., Anjum
649 M.F., Achtman M., Falush D., Donnelly P. 2011. Recombination and population
650 structure in salmonella enterica. PLoS Genet. 7.

651 Didelot X., Falush D. 2007. Inference of Bacterial Microevolution Using Multilocus
652 Sequence Data. Genetics. 175:1251–1266.

653 Didelot X., Lawson D., Darling A., Falush D. 2010. Inference of homologous recombination
654 in bacteria using whole-genome sequences. Genetics. 186:1435–49.

655 Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with
656 BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–73.

657 Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for Ancient Admixture
658 between Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

659 Edgar R.C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and
660 space complexity. BMC Bioinformatics. 5:113.

661 Edwards D.L., Knowles L.L. 2014. Species detection and individual assignment in species
662 delimitation: can integrative data increase efficacy? Proc. Biol. Sci. 281:20132765.

663 Ence D.D., Carstens B.C. 2011. SpedeSTEM: a rapid and accurate method for species
664 delimitation. Mol. Ecol. Resour. 11:473–80.

665 Ewing G.B., Ebersberger I., Schmidt H.A., von Haeseler A. 2008. Rooted triple consensus
666 and anomalous gene trees. BMC Evol. Biol. 8:118.

667 Excoffier L., Novembre J., Schneider S. 2000. SIMCOAL : A General Coalescent Program
668 for the Simulation of Molecular Data. J. Hered. 91:506–510.

669 Fujisawa T., Barraclough T.G. 2013. Delimiting species using single-locus data and the
670 Generalized Mixed Yule Coalescent approach: a revised method and evaluation on
671 simulated data sets. Syst. Biol. 62:707–24.

672 Fujita M.K., Leaché A.D., Burbrink F.T., McGuire J.A., Moritz C. 2012. Coalescent-based
673 species delimitation in an integrative taxonomy. Trends Ecol. Evol. 27:480–8.

674 Hambäck P.A., Weingartner E., Ericson L., Fors L., Cassel-Lundhagen A., Stenberg J.A.,
675 Bergsten J. 2013. Bayesian species delimitation reveals generalist and specialist parasitic

676 wasps on Galerucella beetles (Chrysomelidae): sorting by herbivore or plant host. BMC
677 Evol. Biol. 13:92.

678 Hebert P.D.N., Cywinska A., Ball S.L., DeWaard J.R. 2003. Biological identifications
679 through DNA barcodes. Proc. R. Soc. B Biol. Sci. 270:313–321.

680 Jolley K.A., Chan M.-S., Maiden M.C.J. 2004. mlstdbNet - distributed multi-locus sequence
681 typing (MLST) databases. BMC Bioinformatics. 5:86.

682 Koufopanou V., Burt A., Taylor J.W. 1997. Concordance of gene genealogies reveals
683 reproductive isolation in the pathogenic fungus Coccidioides immitis. Proc. Natl. Acad.
684 Sci. U. S. A. 94:5478–82.

685 Kubatko L.S., Gibbs H.L., Bloomquist E.W. 2011. Inferring species-level phylogenies and
686 taxonomic distinctiveness using multilocus data in Sistrurus rattlesnakes. Syst. Biol.
687 60:393–409.

688 Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. PartitionFinder: Combined Selection of
689 Partitioning Schemes and Substitution Models for Phylogenetic Analyses. Mol. Biol.
690 Evol. 29:1695–1701.

691 Leaché A.D., Fujita M.K. 2010. Bayesian species delimitation in West African forest geckos
692 (Hemidactylus fasciatus). Proc. Biol. Sci. 277:3071–7.

693 Lemmon A.R., Emme S. a, Lemmon E.M. 2012. Anchored hybrid enrichment for massively
694 high-throughput phylogenomics. Syst. Biol. 61:727–44.

695 Li C., Lu G., Ortí G. 2008. Optimal data partitioning and a test case for ray-finned fishes
696 (Actinopterygii) based on ten nuclear loci. Syst. Biol. 57:519–39.

697 Liu L., Yu L., Edwards S. V. 2010. A maximum pseudo-likelihood approach for estimating
698 species trees under the coalescent model. BMC Evol. Biol. 10:302.

699 Maiden M.C.J., Bygraves J.A., Feil E., Morelli G., Russell J.E., Urwin R., Zhang Q., Zhou J.,
700 Zurth K., Caugant D.A., Feavers I.M., Achtman M., Spratt B.G. 1998. Multilocus
701 sequence typing: A portable approach to the identification of clones within populations
702 of pathogenic microorganisms. Proc. Natl. Acad. Sci. 95:3140–3145.

703 Maiden M.C.J. 2006. Multilocus sequence typing of bacteria. Annu. Rev. Microbiol. 60:561–
704 88.

705 Millanes A.M., Truong C., Westberg M., Diederich P., Wedin M. 2014. Host Switching
706 Promotes Diversity in Host-Specialized Mycoparasitic Fungi: Uncoupled Evolution in
707 the Biatoropsis-Usnea System. Evolution (N. Y).:1–18.

708 Mirarab S., Reaz R., Bayzid M.S., Zimmermann T., Swenson M.S., Warnow T. 2014.
709 ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics.
710 30:i541–8.

711 O'Meara B.C. 2010. New Heuristic Methods for Joint Species Delimitation and Species Tree
712 Inference. Syst. Biol. 59:59–73.

713 Pamilo P., Nei M. 1988. Relationships between Gene Trees and Species Trees. Mol. Biol.
714 Evol. 5:568–583.

715 Papke R.T., Zhaxybayeva O., Feil E.J., Sommerfeld K., Muise D., Doolittle W.F. 2007.
716 Searching for species in haloarchaea. Proc. Natl. Acad. Sci. U. S. A. 104:14092–14097.

717 Paradis E., Claude J., Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R
718 language. Bioinformatics. 20:289–290.

719 Paradis E. 2010. pegas: an R package for population genetics with an integrated–modular
720 approach. Bioinformatics. 26:419–420.

721 Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S.,
722 Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the dna
723 taxonomy of undescribed insects. Syst. Biol. 55:595–609.

724 Puillandre N., Lambert A., Brouillet S., Achaz G. 2012. ABGD, Automatic Barcode Gap
725 Discovery for primary species delimitation. Mol. Ecol. 21:1864–77.

726 Rambaut A., Drummond A.J. 2007. Tracer v1.4. .

727 Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of
728 DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–8.

729 Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral
730 population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.

731 Rosenberg N.A. 2007. Statistical tests for taxonomic distinctiveness from observations of
732 monophyly. Evolution (N. Y). 61:317–23.

733 Satler J.D., Carstens B.C., Hedin M. 2013. Multilocus species delimitation in a complex of
734 morphologically conserved trapdoor spiders (mygalomorphae, antrodiaetidae, aliatypus).
735 Syst. Biol. 62:805–23.

736 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
737 large phylogenies. Bioinformatics. 30:1312–3.

738 Taylor J.W., Jacobson D.J., Kroken S., Kasuga T., Geiser D.M., Hibbett D.S., Fisher M.C.
739 2000. Phylogenetic species recognition and species concepts in fungi. Fungal Genet. Biol.
740 31:21–32.

741 Vialle A., Feau N., Frey P., Bernier L., Hamelin R.C. 2013. Phylogenetic species recognition
742 reveals host-specific lineages among poplar rust fungi. Mol. Phylogenet. Evol. 66:628–
743 44.

744 Vos M., Didelot X. 2009. A comparison of homologous recombination rates in bacteria and
745 archaea. ISME J. 3:199–208.

746 Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data.
747 Proc. Natl. Acad. Sci. U. S. A. 107:9264–9269.

748 Yang Z., Rannala B. 2014. Unguided Species Delimitation Using DNA Sequence Data from
749 Multiple Loci. Mol. Biol. Evol. 31:3125–3135.

750 Zaykin D. V, Pudovkin A., Weir B.S. 2008. Correlation-based inference for linkage
751 disequilibrium with multiple alleles. Genetics. 180:533–45.

752 Zhang C., Rannala B., Yang Z. 2014. Bayesian Species Delimitation Can Be Robust to
753 Guide-Tree Inference Errors. Syst. Biol. 63:993–1004.

754 Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method
755 with applications to phylogenetic placements. Bioinformatics. 29:2869–2876.

756 Zwickl D.J., Stein J.C., Wing R.A., Ware D., Sanderson M.J. 2014. Disentangling
757 methodological and biological sources of gene tree discordance on oryza (poaceae)
758 chromosome 3. Syst. Biol. 63:645–59.

759

760 CAPTIONS OF FIGURES

761 Figure 1. An illustration of how the dynamic programming algorithm finds the optimal
762 delimitation. Below each node in the guide tree, two alternative delimitations are compared
763 (horizontal arrows) and the better one is chosen (dotted squares). The best delimitation below
764 one node is inserted into the comparison at the higher level successively to yield the final
765 optimal delimitation.

766

767 Figure 2. Schematic representations of alternative hypotheses of delimitation for the three-
768 species simulations: a) Correct hypothesis, b) Under-split and c) Over-split.

769

770 Figure 3. Relationships between false negative rate and the number of loci used for
771 delimitation in the three-species simulations that simulated different effective population sizes
772 within species relative to the divergence time between species.

773

774 Figure 4. Relationships between the number of exact matches and the number of loci used in
775 the 10-species simulations. A) Both guide trees and gene trees are known, B) Guide trees are
776 estimated but gene trees are known and C) Guide trees and gene trees are estimated from
777 DNA sequences.

778

779 Figure 5. Relationships between the number of estimated species and the number of loci used
780 in the 10-species simulations. A) Both guide trees and gene trees are known, B) Guide trees

are estimated but gene trees are known and C) Guide trees and gene trees are estimated from DNA sequences.

Figure 6. The number of species estimated when randomly re-sampling loci in the empirical data sets. Left) Rattlesnakes. Right) *Bacillus* complex.

Figure 7. Results of delimitation with 100 sets of gene trees and guide trees sampled from MCMC runs. Trees from ClonalFrame MCMC were used as guide trees. Left) The 50% majority consensus tree built with ClonalFrame . Right) The frequency that each pair of isolates was grouped by tr2.

CAPTIONS OF SUPPLEMENTARY FIGURES

Figure S1. a) Frequency of false positives for three types of model comparison procedures. The count of each topology was drawn from a trinomial distribution with equal rate, which simulates samples from a single species, and model comparisons using AIC and the Bayesian modelling were conducted. AICs were calculated using the likelihood of eq.(1) for the single-species case and eq.(2) for the three-species case. Posterior probabilities were calculated with eq.(6) and eq.(7). The numbers of trials where the three-species case had larger AICs or smaller posterior probabilities than the single-species case were recorded as false positives. Abbreviations of models are: AIC_MF: AIC with the most frequent topology as dominant topology. AIC_Random: AIC with randomly chosen triplet as dominant topology. Bayes: Bayesian model comparison with six models.

803

804

805 Figure S2. Relationships between the required time for delimitation and the number of loci.

806

807 Figure S3. Average number of non-monophyletic species for each setting in the 10-species
808 simulations.

809

810 Figure S4. A rooted triple consensus tree inferred from 18 nuclear loci of *Sistrurus* snakes.
811 Polytomies were randomly resolved. Bars in dark grey indicate species delimited by the tr2.
812 Bars in light grey show an alternative delimitation that appeared in the repeated delimitations.
813 The numbers on nodes represent the average difference of log posterior probability scores
814 between the null and alternative models defined by the nodes. A positive value indicates that
815 the delimitation B (separate species) is preferred over the delimitation W (same species).
816 Nodes indicated by asterisks are the most recent common ancestor nodes for species groups.

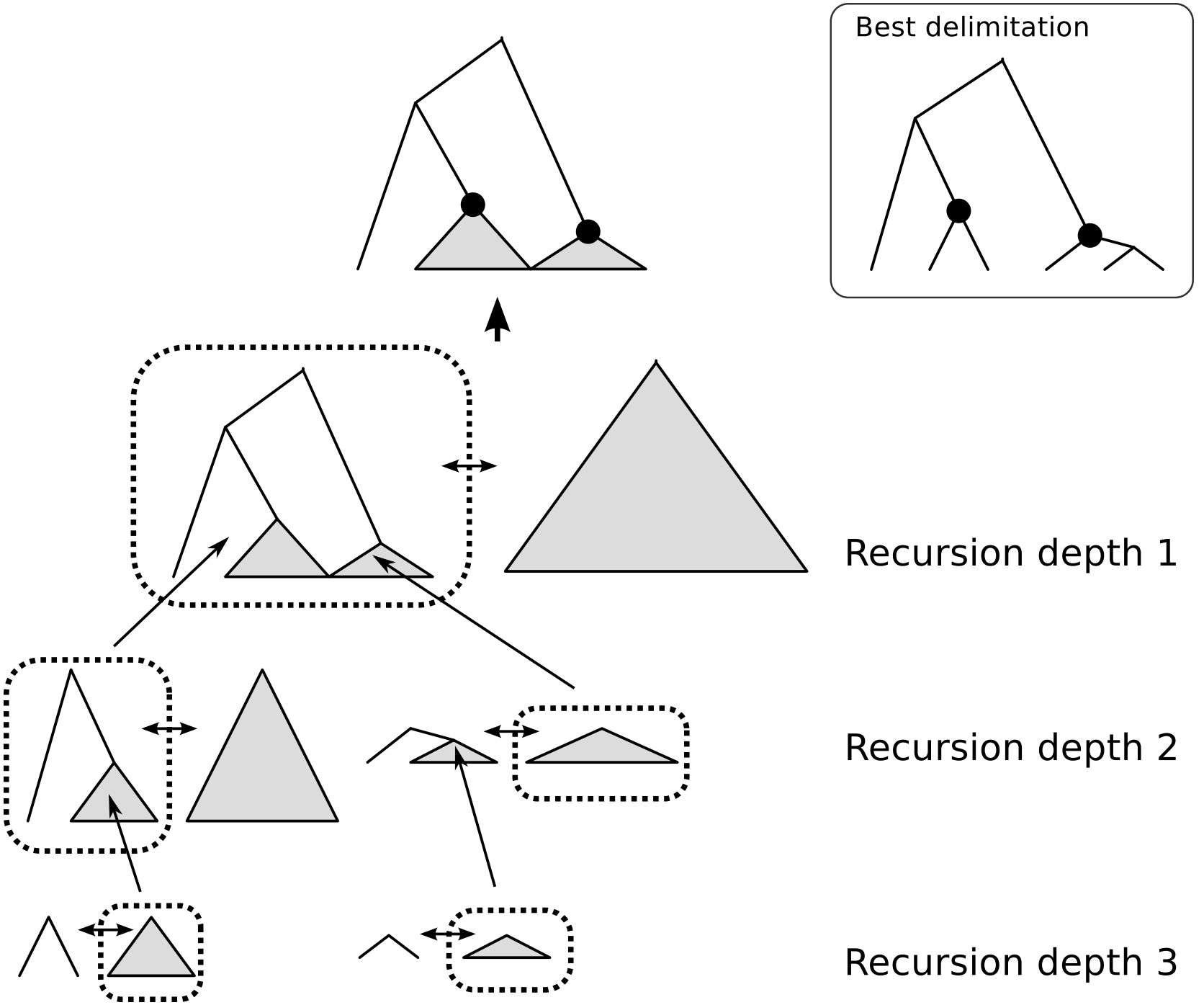
817

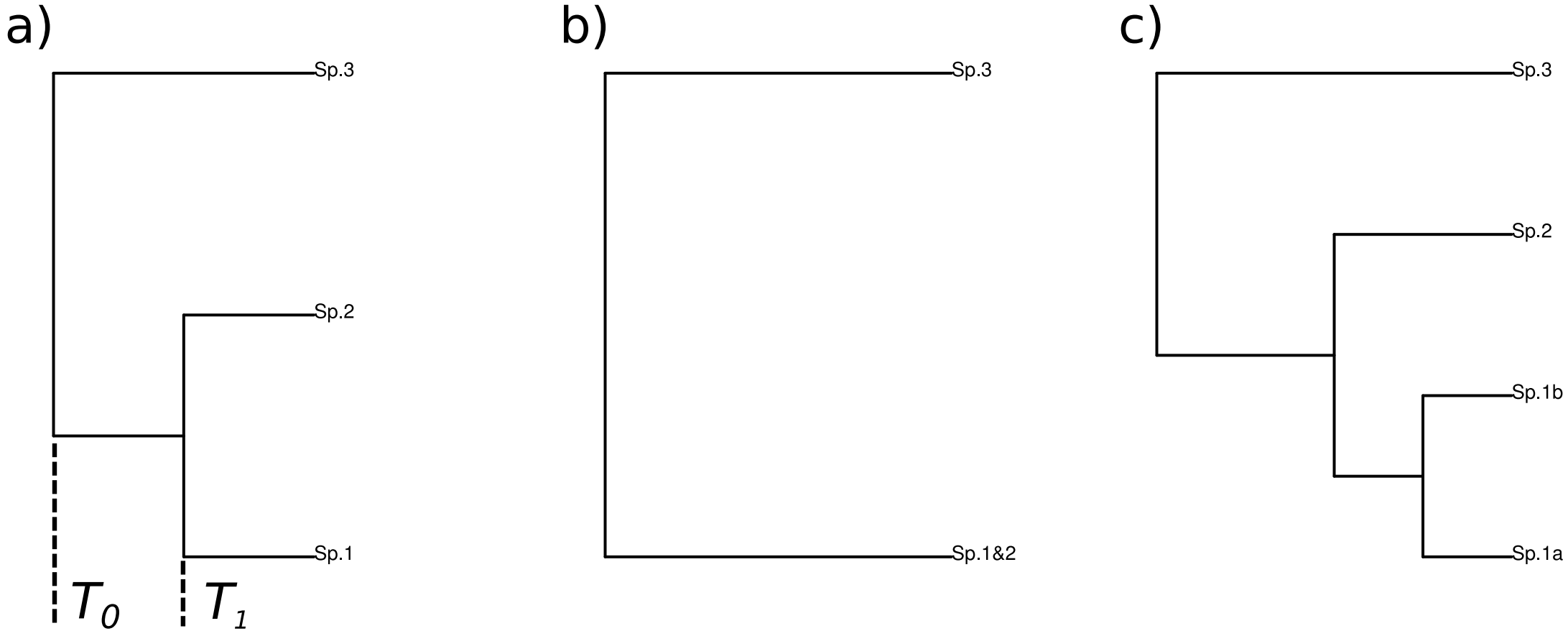
818 Figure S5. Results of delimitation of *Bacillus* with 100 sets of gene trees sampled from
819 MCMC runs using rooted triple consensus as guide trees. Left) The rooted triple consensus
820 built with seven MCC trees . Right) Frequency that each isolate is grouped by the tr2.

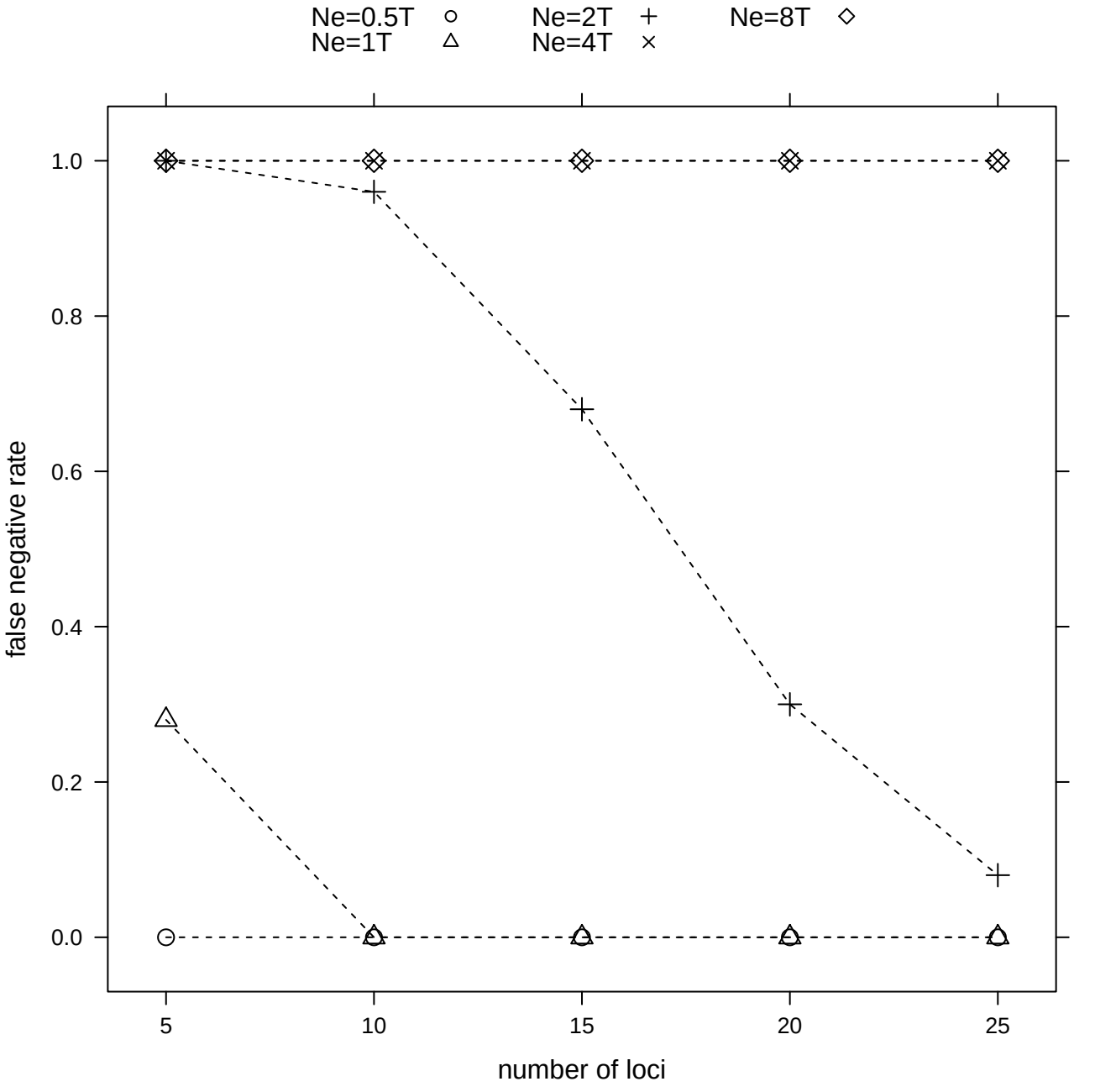
821

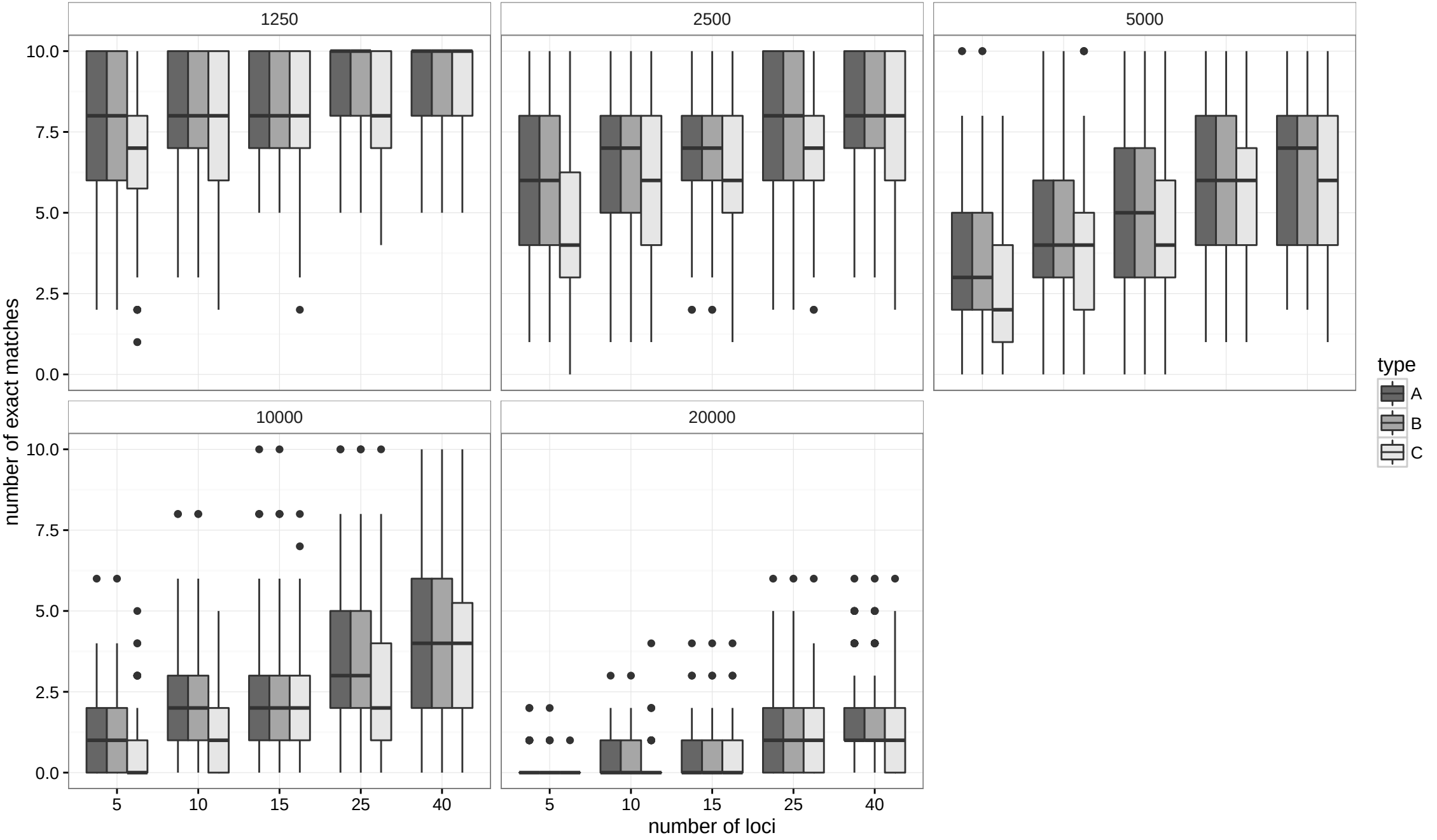
822 Figure S6. Distribution of linkage disequilibrium measured by R^2. Dotted grey lines indicate
823 the border of loci and each grid cell represents the positions of each locus, from left hand side,
824 glpF, gmk, ilv, pta, purH, pycA and tipD.

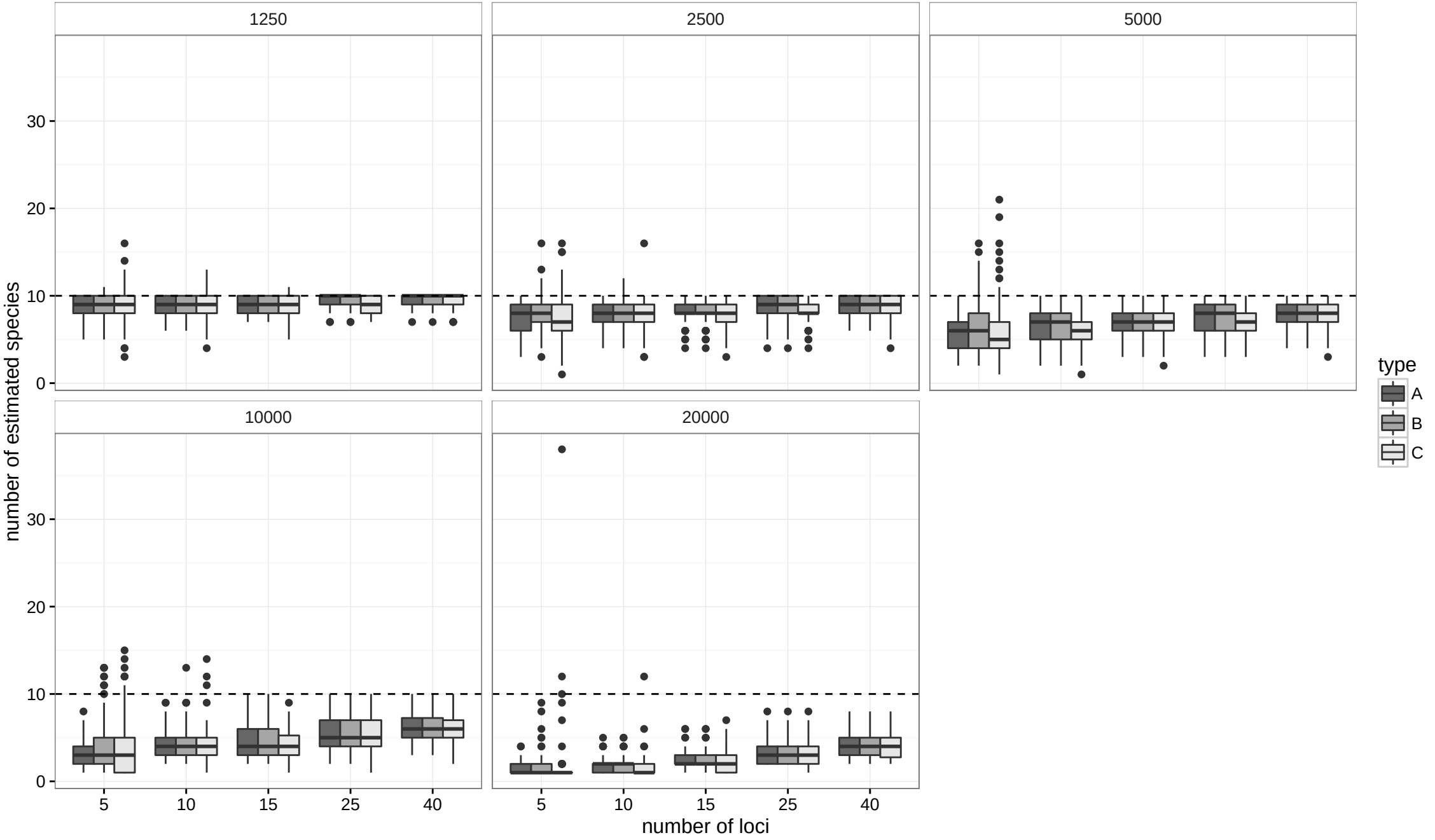
Figure S7. Relationship between the number of exact matches and total sample size (loci X individual samples). Dark grey: trials with sample size=50. Light grey: trials with sample size = 100.

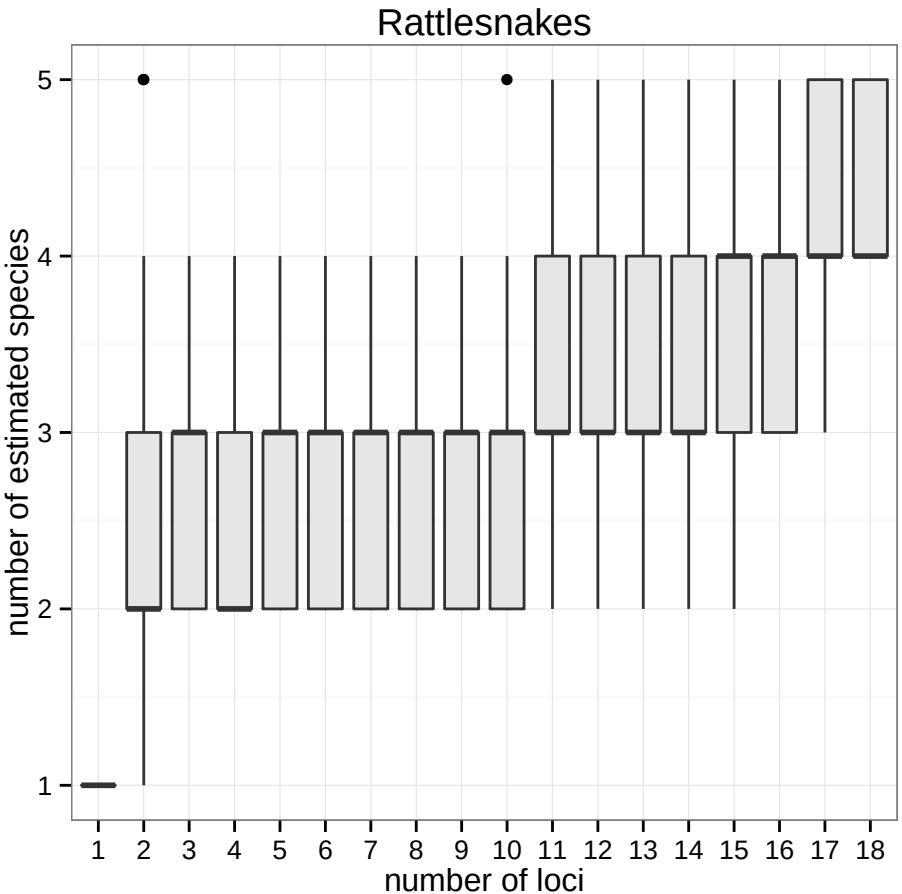

Best delimitation
Recursion depth 1
Recursion depth 2
Recursion depth 3

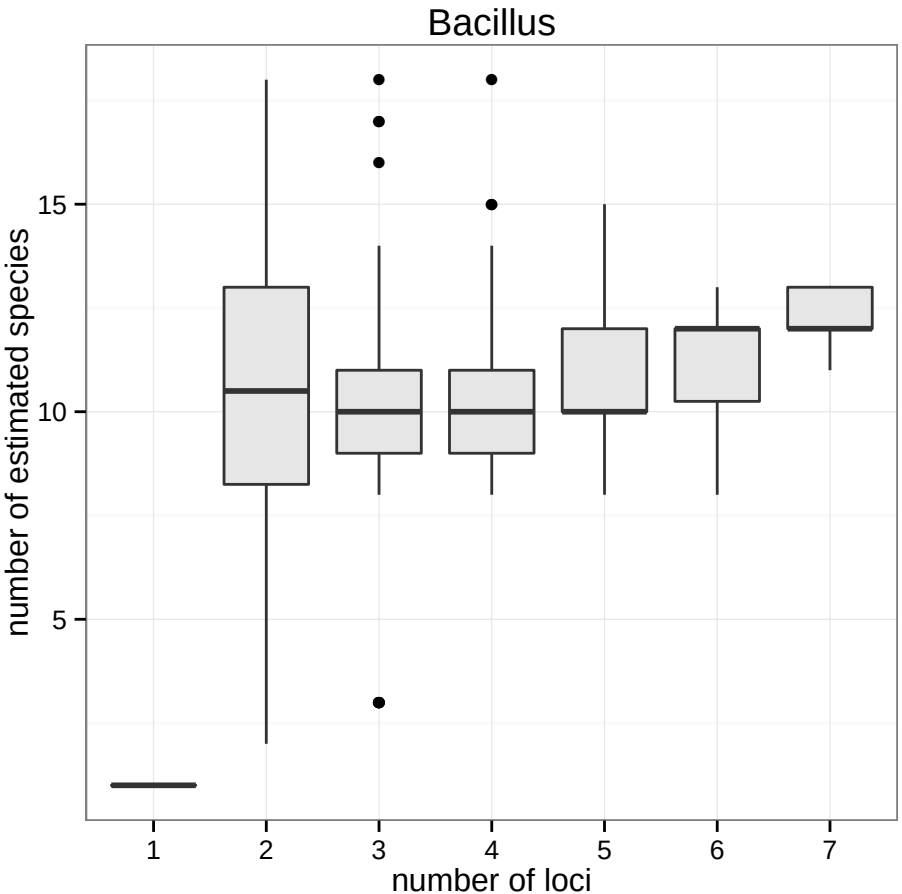

a)
Sp.3
Sp.2
Sp.1
T_0
T_1
b)
Sp.3
Sp.1&2
c)
Sp.3
Sp.2
Sp.1b
Sp.1a

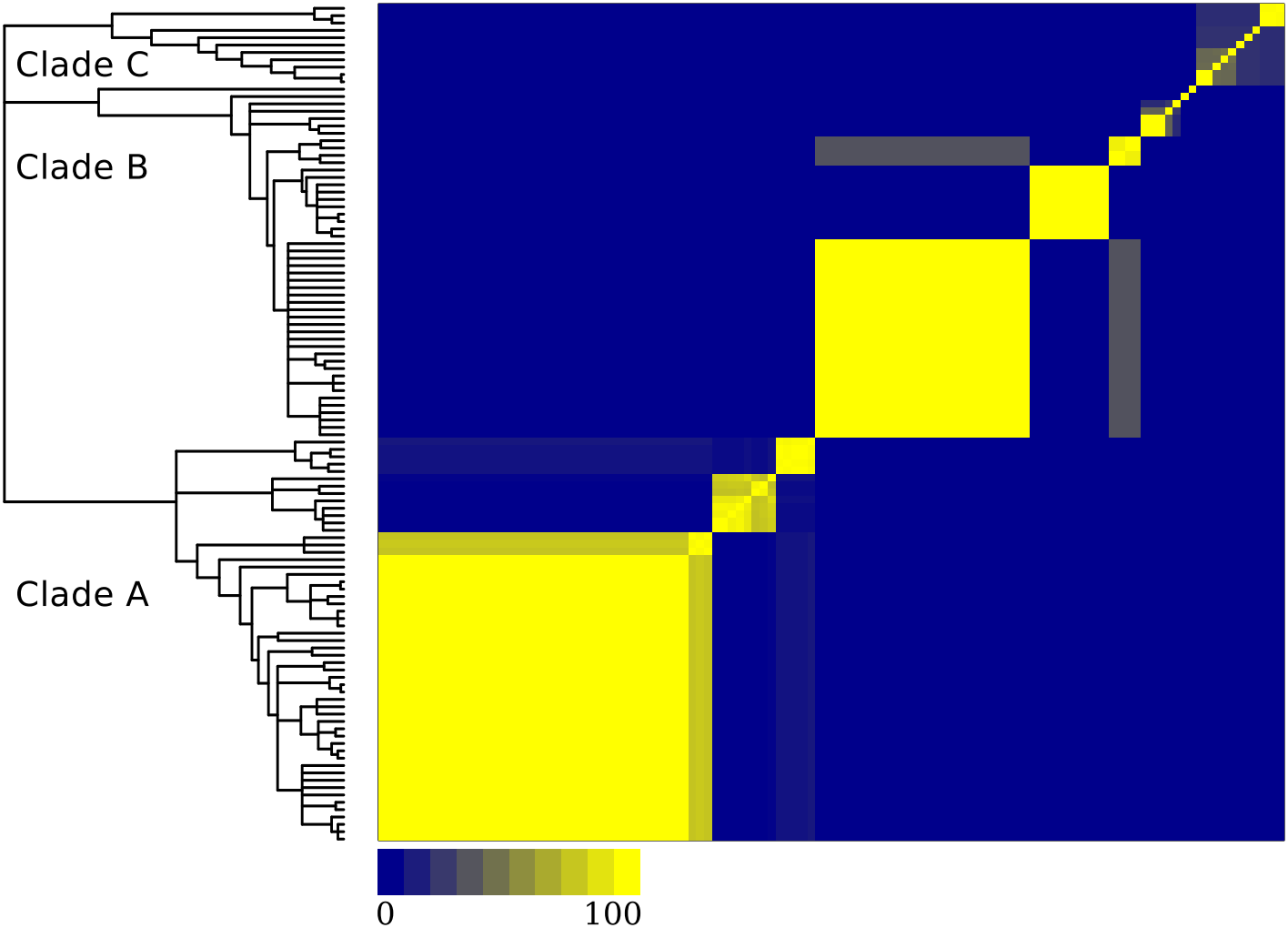

Ne=0.5T
Ne=1T
Ne=2T
Ne=4T
Ne=8T
false negative rate
number of loci
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25

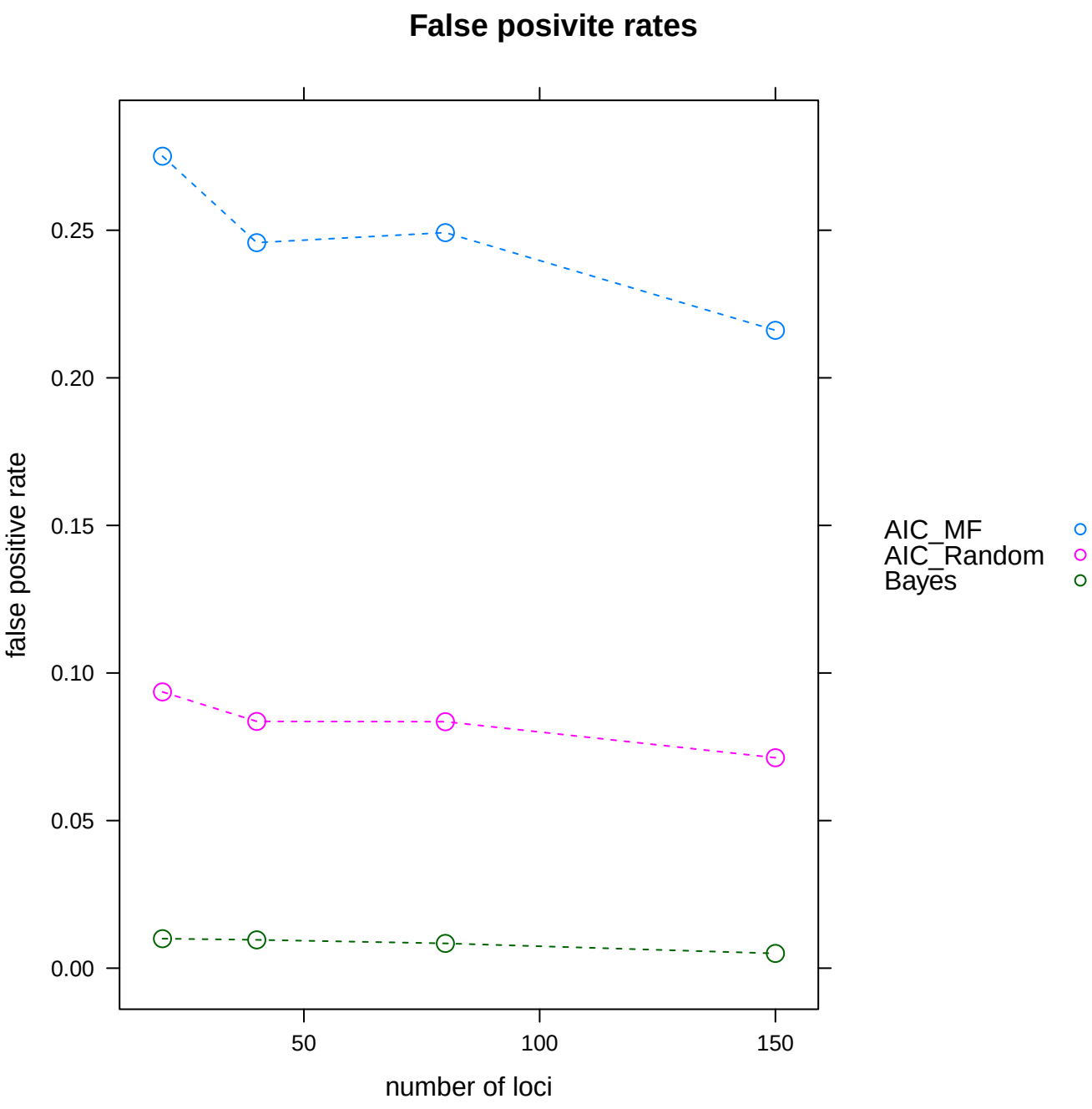

1250
2500
5000
10000
20000
number of exact matches
number of loci
10.0
7.5
5.0
2.5
0.0
5
10
15
25
40
type
A
B
C

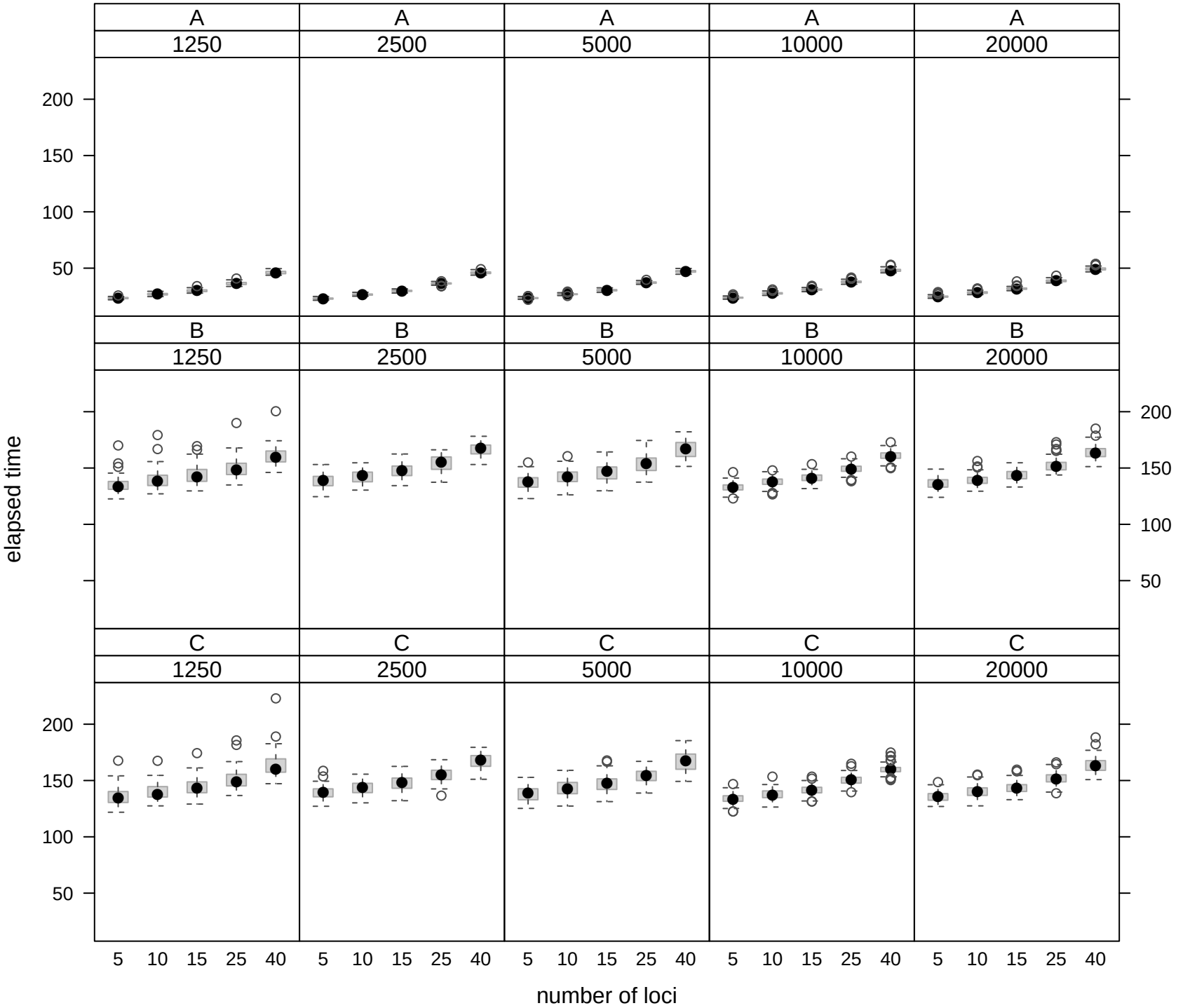

1250
2500
5000
10000
20000
number of estimated species
0
10
20
30
5
10
15
25
40
number of loci
type
A
B
C

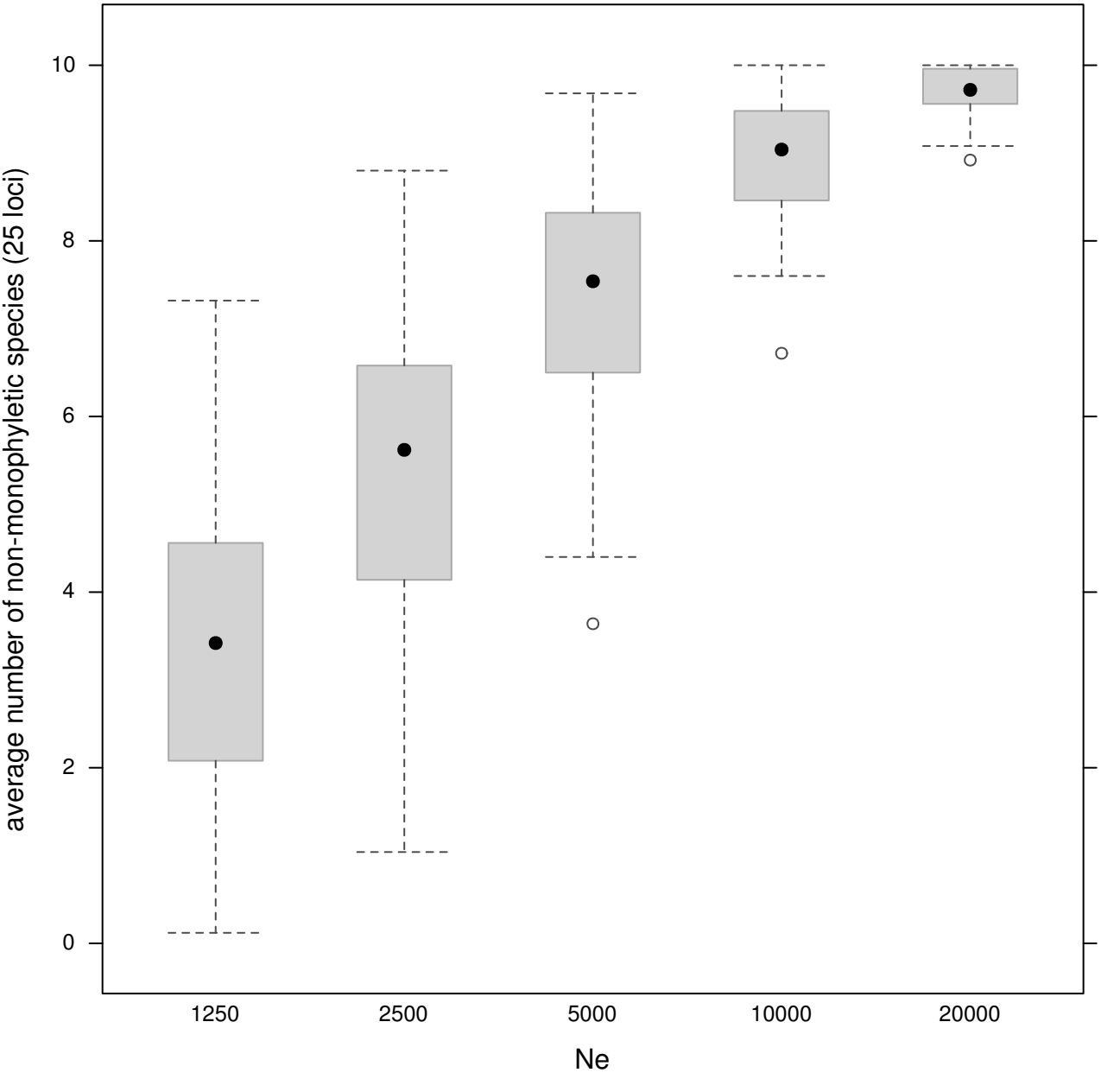

Rattlesnakes
number of estimated species
number of loci
1
2
3
4
5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

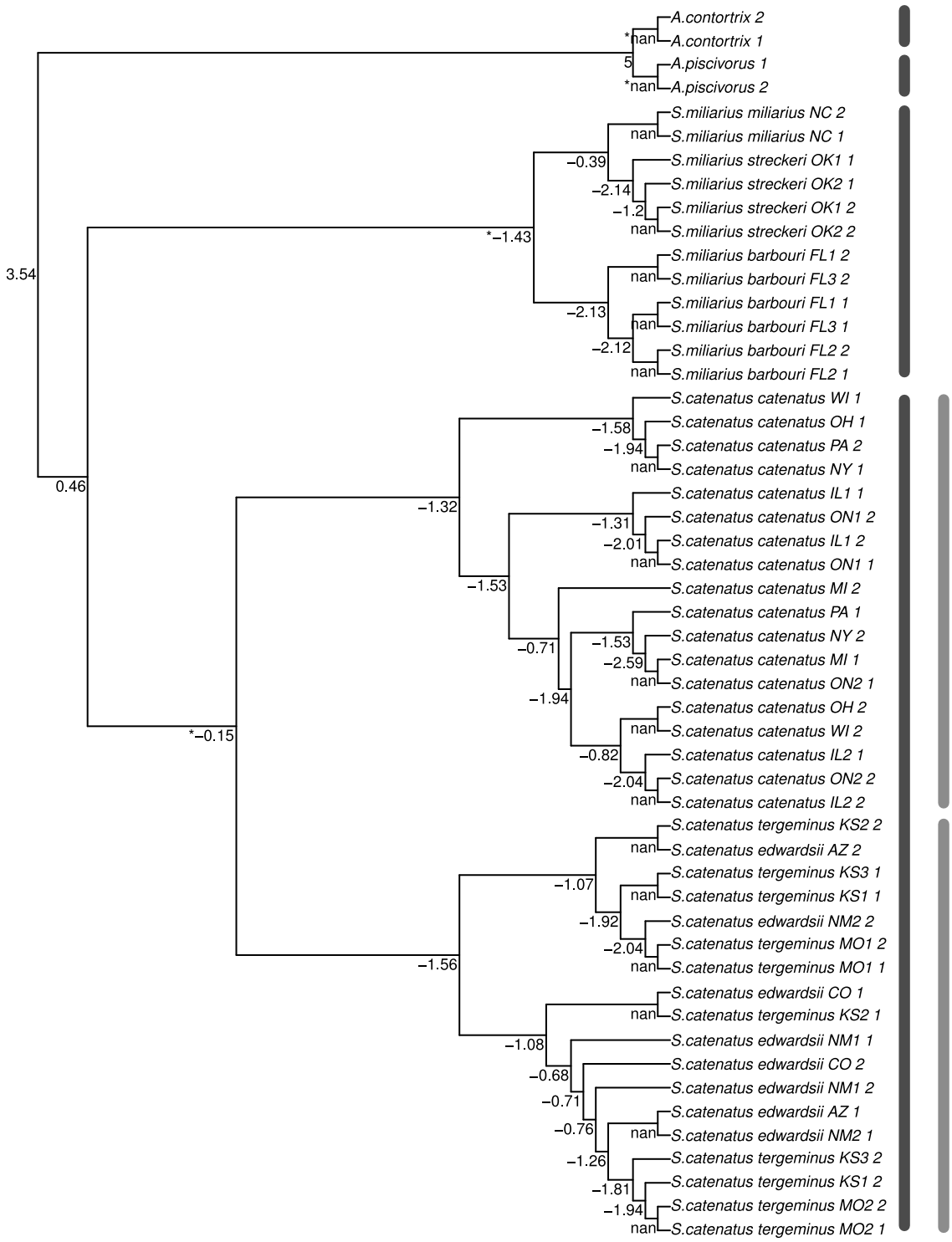

Bacillus
number of estimated species
number of loci
5
10
15
1 2 3 4 5 6 7

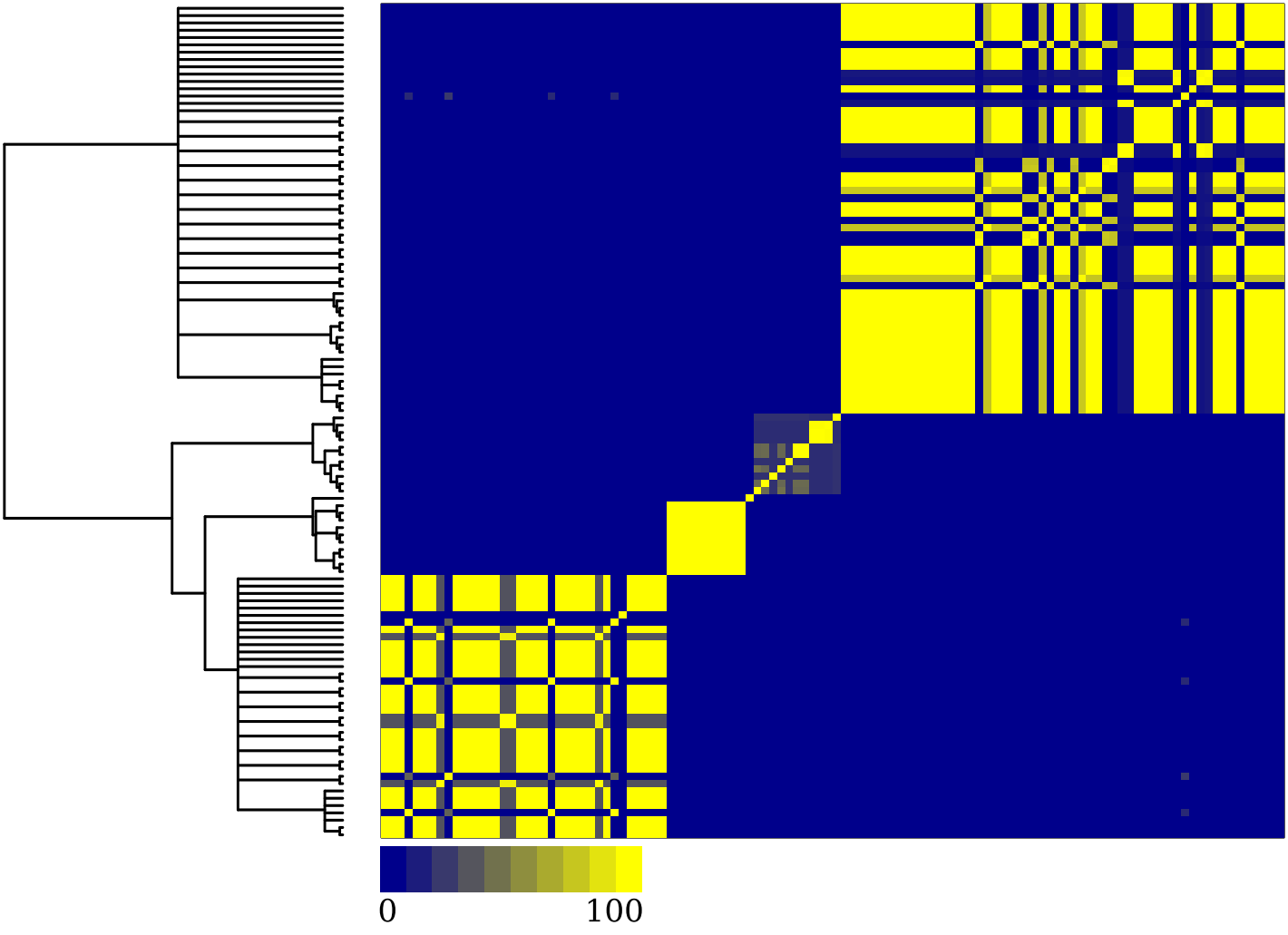

Clade C
Clade B
Clade A
0
100

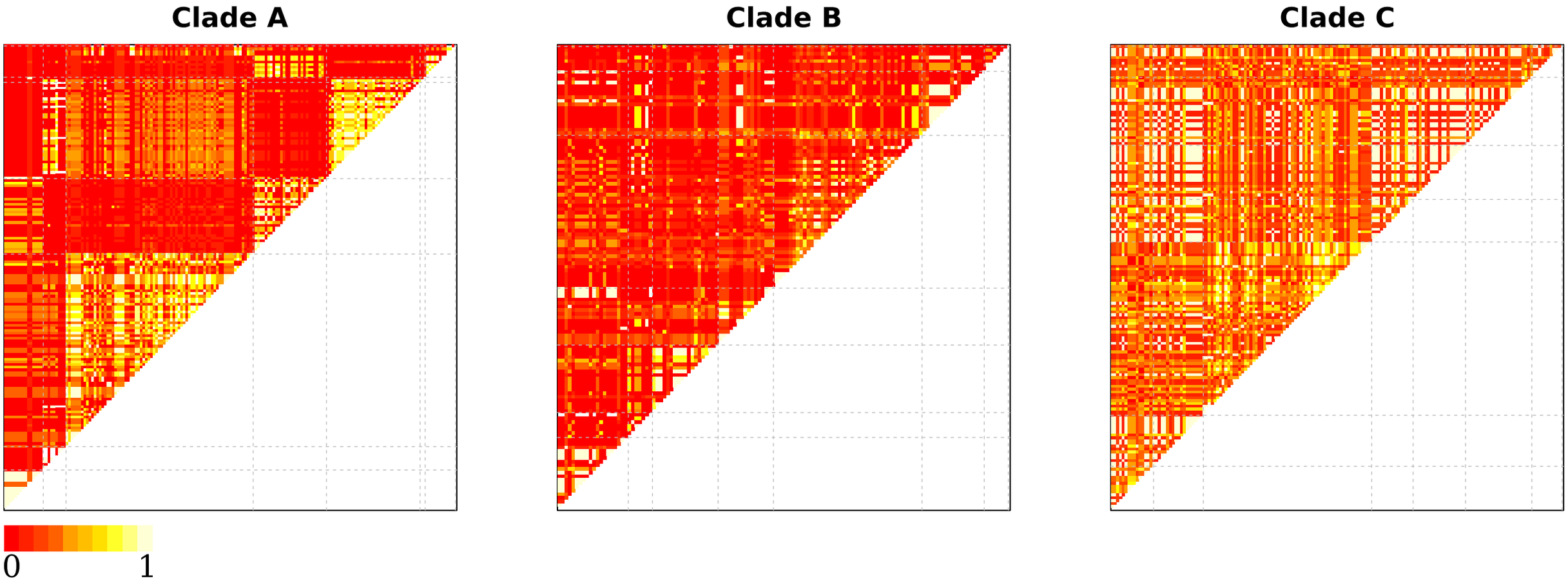

False posivite rates
false positive rate
0.00
0.05
0.10
0.15
0.20
0.25
50
100
150
number of loci
AIC_MF
AIC_Random
Bayes

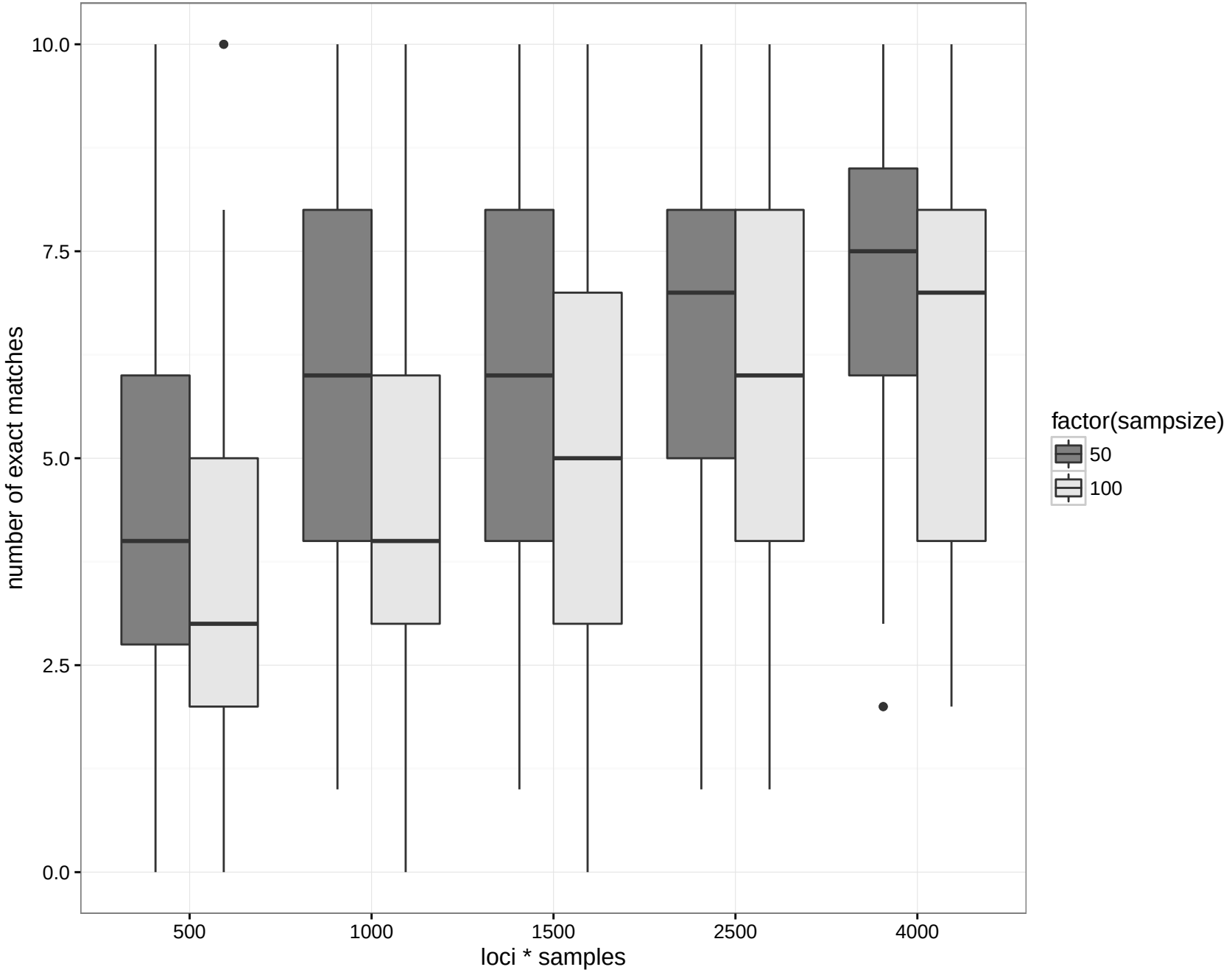

A
B
C
1250
2500
5000
10000
20000
elapsed time
number of loci
5
10
15
25
40
50
100
150
200


average number of non-monophyletic species (25 loci)
0
2
4
6
8
10
1250
2500
5000
10000
20000
Ne


A.contortrix 2
A.contortrix 1
A.piscivorus 1
A.piscivorus 2
S.miliarius miliarius NC 2
S.miliarius miliarius NC 1
S.miliarius streckeri OK1 1
S.miliarius streckeri OK2 1
S.miliarius streckeri OK1 2
S.miliarius streckeri OK2 2
S.miliarius barbouri FL1 2
S.miliarius barbouri FL3 2
S.miliarius barbouri FL1 1
S.miliarius barbouri FL3 1
S.miliarius barbouri FL2 2
S.miliarius barbouri FL2 1
S.catenatus catenatus WI 1
S.catenatus catenatus OH 1
S.catenatus catenatus PA 2
S.catenatus catenatus NY 1
S.catenatus catenatus IL1 1
S.catenatus catenatus ON1 2
S.catenatus catenatus IL1 2
S.catenatus catenatus ON1 1
S.catenatus catenatus MI 2
S.catenatus catenatus PA 1
S.catenatus catenatus NY 2
S.catenatus catenatus MI 1
S.catenatus catenatus ON2 1
S.catenatus catenatus OH 2
S.catenatus catenatus WI 2
S.catenatus catenatus IL2 1
S.catenatus catenatus ON2 2
S.catenatus catenatus IL2 2
S.catenatus tergeminus KS2 2
S.catenatus edwardsii AZ 2
S.catenatus tergeminus KS3 1
S.catenatus tergeminus KS1 1
S.catenatus edwardsii NM2 2
S.catenatus tergeminus MO1 2
S.catenatus tergeminus MO1 1
S.catenatus edwardsii CO 1
S.catenatus tergeminus KS2 1
S.catenatus edwardsii NM1 1
S.catenatus edwardsii CO 2
S.catenatus edwardsii NM1 2
S.catenatus edwardsii AZ 1
S.catenatus edwardsii NM2 1
S.catenatus tergeminus KS3 2
S.catenatus tergeminus KS1 2
S.catenatus tergeminus MO2 2
S.catenatus tergeminus MO2 1
*nan
5
*nan
nan
−0.39
−2.14
−1.2
nan
*−1.43
3.54
nan
−2.13
nan
−2.12
nan
−1.58
−1.94
nan
0.46
−1.32
−1.31
−2.01
nan
−1.53
−0.71
−1.53
−2.59
nan
−1.94
nan
−0.82
−2.04
nan
*−0.15
nan
−1.07
nan
−1.92
−2.04
nan
−1.56
nan
−1.08
−0.68
−0.71
−0.76
nan
−1.26
−1.81
−1.94
nan


0
100


Clade A
Clade B
Clade C
0
1


10.0
7.5
5.0
2.5
0.0
number of exact matches
500
1000
1500
2500
4000
loci * samples
factor(sampsize)
50
100