The Genealogical Sorting Index and species delimitation

David J. Winter[1,2*], Steven A. Trewick[3], Jon M. Waters[2] and Hamish G. Spencer[2]

[1]The Biodesign Institute, Arizona State University, Tempe, AZ, USA.

[2]Allan Wilson Centre for Molecular Ecology & Evolution, Department of Zoology, University of Otago

[3]Ecology Group, Institute of Agriculture and Environment, Massey University, Palmerston North, New Zealand.

***Corresponding Author,** email: david.winter@gmail.com

Abstract

The Genealogical Sorting Index (*gsi*) has been widely used in species-delimitation studies, where it is usually interpreted as a measure of the degree to which each of several predefined groups of specimens display a pattern of divergent evolution in a phylogenetic tree. Here we show that the *gsi* value obtained for a given group is highly dependent on the structure of the tree outside of the group of interest. By calculating the *gsi* from simulated datasets we demonstrate this dependence undermines some of desirable properties of the statistic. We also review the use of the *gsi* delimitation studies, and show that the *gsi* has typically been used under scenarios in which it is expected to produce large and statistically significant results for samples that are not divergent from all other populations and thus should not be considered species. Our proposed solution to this problem performs better than the *gsi* in under these conditions. Nevertheless, we show that our modified approach can produce positive results for populations that are connected by substantial levels of gene flow, and are thus unlikely to represent distinct species. We stress that the properties of *gsi* made clear in this manuscript must be taken into account if the statistic is used in species-delimitation studies. More generally, we argue that the results of genetic species-delimitation methods need to be interpreted in the light the biological and ecological setting of a study, and not treated as the final test applied to hypotheses generated by other data.

Keywords

gsi, pairwise-*gsi*, population genetics, population structure, species delimitation

Introduction

Genetic sequence data and phylogenetic methods are increasingly being used to aid in the discovery and delimitation of species (reviewed in Fujita *et al.* 2012). The widespread application of such data and analyses to alpha taxonomy has confirmed that evolutionarily distinct species will not necessarily fall into reciprocally monophyletic groups in phylogenies estimated from DNA sequences.

Indeed, species can remain paraphyletic with respect to their close relatives in gene trees even millions of years after they begin to diverge (Tajima 1983; Hudson & Coyne 2002) .

A number of methods have been developed with the objective of delimiting species using such unsorted gene trees (Knowles & Carstens 2007; O'Meara 2010; Yang & Rannala 2010, 2014; Ence & Carstens 2011; Zhang *et al.* 2011). The increasingly popular use of these methods in empirical species-delimitation studies has inspired a number of methodological papers exploring their statistical properties. These theoretical investigations have shown the methods to be powerful and accurate when their underlying assumptions are met, but it has become clear that violations of these assumptions can generate misleading results (Reid & Carstens 2012; Carstens *et al.* 2013; Edwards & Knowles 2014; Olave *et al.* 2014). Thus, species delimitation methods are most useful when their statistical properties are understood, and studies can be designed and interpreted in the light of these properties.

Although not exclusively designed for species-delimitation studies, the *gsi* of Cummings et al. (2008) has been widely used in this context (see references in Table 1). This statistic is a measure of the degree to which a pre-defined group of leaves in a phylogenetic tree falls into an exclusive region in that tree. The value of the statistic ranges from 0 to 1, with higher values corresponding to more phylogenetic exclusivity. In this way, the *gsi* aims to bridge the gap between monophyly and paraphyly as categorical terms, and in so doing, quantify the degree to which a lineage has become distinct as a result of evolutionary divergence. The calculation of the *gsi* is usually accompanied by an hypothesis test, in which the *gsi* for each group is compared to values calculated from trees in which the tip labels have been permuted.

When compared to other widely used species delimitation methods (Table 2) the *gsi* has many desirable properties. As well as having power to detect recently diverged lineages, the *gsi* differs from many comparable methods in not needing, as input parameters, the values of often difficult-to-estimate quantities such as the effective population size or mutation rate of the genetic sequences

under consideration. The *gsi* value obtained for a given group is also purported to be comparable to those obtained for different groups within the same tree and between those arising from different studies (Cummings *et al.* 2008).

In applying the *gsi* to empirical data, however, we have found the value of this statistic to be highly dependent on the structure of the tree outside of the group of interest. This dependence is not reflected in the way the statistic is typically applied and interpreted in species-delimitation studies, and this mismatch between the *gsi* as it is used and goals of species delimitation undermines the many advantages of the statistic.

The gsi *measures exclusivity relative to the entire tree*

An example serves to illustrate the dependence of the *gsi* obtained for a particular group on the over-all structure of the phylogeny from which it is calculated. Take the tree presented in Figure 1. To calculate *gsi* for the group "a" in this tree we first need to calculate the intermediate statistic *gs*, which is defined as

$$gs = \frac{n}{\sum_{u=1}^{U}(d_u - 2)} \quad (1)$$

where d_u is the degree (i.e. the number of connections) of the node *u,* which is one of *U* total nodes in the smallest sub-tree uniting all members of a group and *n* is the minimum number of nodes that could be used to unite this group, which is one less than the number of leaves. In the case of Figure 1, *n* is 3, but all 7 nodes in the tree are needed to create a sub-tree uniting group "a". As the tree is fully dichotomous, the degree of each node is 3. Thus, *gs* is 3 / (7 × (3-2)) = 3/7. To obtain *gsi* for group "a" the observed value of *gs* is normalised using the maximum and minimum obtainable value for the statistic given the size of the group and the number of nodes in the tree:

$$gsi = \frac{observed gs - \min(gs)}{\max(gs) - \min(gs)} \quad (2)$$

Here max(*gs*) is 1 (the case in which a group is united by the minimum number of nodes, i.e. monophyly) and min(*gs*) is given by the equation

$$\min(gs) = \frac{n}{\sum_{i=1}^{I}(d_i - 2)} \qquad (3)$$

where d_i is the degree of node *i*, one of *I* nodes in the entire tree. Because the smallest sub-tree uniting the group "a" in Figure 1 is the entire tree, min(*gs*) for this group is equal to the observed *gs* value. Thus, the numerator in equation (2) is 3/7 – 3/7 = 0 and so the value of *gsi* is also 0. This result is desirable, as the tree presented in Figure 1 has each group arranged in the least exclusive fashion possible. Nevertheless, defining min(*gs*) in this way means the value of *gsi* is partially dependent on the degree to which other groups in the tree fall into exclusive regions.

Consider now the tree presented in Figure 2, which could be obtained from genetic data underlying the topology illustrated in Figure 1 by simply adding further data from two distantly related taxa. Because the clade containing the "a" and "b" groups is unchanged the observed value of *gs* for "a" is still 7. The addition of the two groups "c" and "d" to the tree , however, has added 8 new dichotomous nodes. Thus the value of min(*gs*) is now 3/(15 × (3-2)) = 3/15 and, following equation (2), *gsi* is equal to [3/7 – 3/15]/[1 – 3/15] ≈ 0.29. This difference arises from the inclusion of min(*gs*) in the calculation of *gsi*, which makes a *gsi* value obtained for a group a reflection of that group's exclusivity relative to the entire tree. This property is not desirable for a species-delimitation statistic, as it means large *gsi* values can be obtained for groups that do not represent a population that is divergent from all other samples in a given analysis. Moreover, it also compromises comparisons between groups within one study, or between values obtained from different studies.

A similar issue affects the hypothesis test that is often performed alongside the *gsi*. Statistical significance, or P-values, arising from this test are usually reported for each of several putative species under consideration in a single analysis. In practice, these P-values are interpreted as the results of independent tests that each group being considered is divergent from all other groups. In

fact, as Cummings et al. (2008) make clear, because the test is performed by permuting group assignments across the entire tree, the null hypothesis being tested is that all individuals included in the tree come from a single panmictic population. It is seldom the case that all individuals considered in a species-delimitation study could plausibly have come from a randomly mating population. Thus, statistically significant results may simply represent the rejection of an implausible null hypothesis.

The gsi *and species delimitation*

The problems discussed above are most likely to affect interpretation of the *gsi* when the statistic is calculated for a large number of groups, especially when those groups are likely to be divergent from at least some others under consideration. To determine how often the *gsi* has been used in such contexts, we performed a literature review (Table 1). We identified papers recorded as citing Cummings et al. (2008) in Web Of Science and Google Scholar. For each study we recorded the context in which the *gsi* was used, the largest number of groups considered in a single analysis and the criteria by which those groups were determined. The results of this analysis show the *gsi* has mainly been used in the context of species-delimitation (54 of 78 empirical studies) and that these studies have often applied the statistic to several groups (mean = 9.9, median = 6) for which there is *a priori* evidence for evolutionary divergence. The basis of the group assignment is frequently an existing taxonomic distinction, or a preliminary phylogenetic or clustering analysis performed on data from which the *gsi* was calculated. Worryingly, these circumstances are exactly those in which (as we show above) use of the *gsi* can be misleading. We did not find any papers in which the plausibility of the null hypothesis was considered in discussing the statistical significance of results.

To investigate how large the effect of including multiple divergent groups in the calculation of *gsi* is likely to be in practice, we calculated the statistic from simulated datasets. We used the program ms (Hudson 2002) to simulate gene trees arising from neutral evolution under the demographic history depicted in Figure 3, that is, four divergent populations with two (“a” and “b”)

diverging at a time point *t* which was varied among simulations. We performed 500 simulations for each value of *t* between 0 and 1 N_e generations in 0.05 N_e increments, sampling 10 individuals per population in each simulation. For each simulation, we calculated the *gsi* for group "a" twice, first considering all populations in the simulation (the "four-population tree") then after discarding individuals from the divergent populations "c" and "d" (the "two-population tree").

As expected, the mean *gsi* value obtained for the group "a" tracks the divergence of this population from population "b" in all simulations (Fig 4). However, the *gsi* value calculated from the four-population tree is substantially larger than the value obtained from the two-population tree, even though the same individuals make up the "a" population in each case. The difference is especially pronounced early in the divergence process, indeed, the expected value of *gsi* in the four-population case is high (0.40) even when $t = 0\ N_e$ (i.e. when populations "a" and "b" are panmictic with respect to each other). For every simulation, including those in which the "a" and "b" populations were panmictic, the calculation based on the four population-tree produced a significant result. By contrast, calculations based on the two-population tree produced a near-uniform distribution of P-values under panmixia and became increasingly likely to return significant results as the populations diverged.

These simulations confirm that both the value and the significance obtained for the *gsi* of a given group is partially dependent on the degree to which other groups fall into exclusive regions of the phylogeny being considered. As we note above, this characteristic is not desirable in a statistic purported to relate only to the group under consideration, as it makes comparison of *gsi* values obtained from different trees problematic. Significant results can readily be obtained from populations that are not genealogically divergent from all others groups in an analysis.

Aligning the gsi with species delimitation as it is practiced

As noted above, the *gsi* has many desirable properties (Table 2). Unlike many species delimitation methods, the *gsi* can be calculated without the need for often difficult-to-estimate

158 population-genetic parameters as input. Additionally, the relative simplicity of the *gsi* means the
159 statistic can be applied to large datasets. Unlike the GMYC (Pons et al., 2006), the *gsi* can be applied
160 to unsorted gene trees and the *gsi* can be used to test the validity of proposed species suggested by
161 morphological or other data. Given these unique properties of the *gsi,* we do no propose that
162 empiricists discarding the statistic entirely. Rather, the properties described here should be carefully
163 considered before the statistic before is applied to datasets.

164 There are likely to be multiple ways to reasonably incorporate the *gsi* in particular species
165 delimitation studies; here we propose a general solution that retains the *gsi*'s simplicity but removes
166 its dependence on the overall structure of the tree from which it is calculated. Our proposed
167 statistic, the "mean pairwise gsi" or *pwgsi* is calculated for a pair of groups, after all tips other than
168 those in the groups of interest have been dropped from the phylogenetic tree under consideration.
169 This approach can be applied to all putative species under consideration in a given study, or only to a
170 subset that are of particular interest.

171 For example, to analyse the tree and group-assignments depicted in Figure 2 we first produce
172 trees representing the possible pairwise comparisons of groups (Fig 5). For each tree, the *pwgsi* is
173 simply the mean of the *gsi* values obtained for these two groups. Thus, in the case of Figure 5 the
174 *pwgsi* for the "a:b" comparison is 0 and all other comparisons have *pwgsi* of 1. This approach
175 requires at most $\binom{n}{2}$ values to be calculated, where n is the number of groups being considered.
176 Thus, the pairwise approach is not subject to the computation limitations of methods that consider
177 all possible partitions of a group-assignment (O'Meara 2010; Ence & Carstens 2011), and can be
178 applied to datasets in which a relatively large number of groups are being considered. It also
179 relatively easy to calculate the *pwgsi* with the existing GENEALOGICALSORTING software, as we
180 demonstrate in Supplementary Text 1.

181 The pairwise approach aligns well with the goals of species delimitation studies, as the *pwgsi*
182 quantifies each population's exclusivity relative to all other populations. Moreover, the pattern of

183 *pwgsi* values resulting from a single analysis can identify groups that are not divergent with respect
184 to each other, but are divergent from all other groups and thus might be considered part of a single
185 divergent population in subsequent analyses (as is the case with "a" and "b" in example discussed
186 above). This approach uses the same procedure as in the calculation of the two-population scores in
187 Fig 4. We can infer from Fig 4, therefore, that the *pwgsi* tracks lineage divergence and a permutation
188 test applied to a particular between-population comparison has a strong power to detect divergent
189 groups.

190 pwgsi *and population structure*

191 The high power of the *pwgsi* to detect an exclusive distribution of tips on a phylogeny may seem
192 to make it an ideal statistic for species delimitation in the presence of incomplete lineage sorting.
193 Care needs to be taken, however, in the interpretation of these results. The short period of
194 divergence required to obtain significant results means that such results can be obtained even for
195 what may turn out to be transient isolation between populations. Moreover, speciation is not the
196 only one way in which a non-random distribution of groups might occur on a phylogeny. Specifically,
197 sub-populations within a population with some degree of genetic structuring may be expected to fall
198 into partially exclusive regions of a gene tree. To investigate the degree to which population
199 structure affects the value of the *pw-gsi*, we simulated neutral gene trees under the scenario
200 depicted in Figure 6. In this case, three populations diverged instantaneously at a time point that
201 was held constant at 5 N_e generations. Two of the descendent populations ("a" and "b") were united
202 by ongoing and constant gene flow due to migration at a rate $4N_em$ with values {1, 2, 5, 25, 100}*.
203 We performed 1500 simulations for each migration-rate value and sampled.

* Note, the inclusion of population "c" in this design illustrates the importance of the *pw-gsi* approach to quantifying lineage divergence. As this simulation proceeds population "c" is expected to become increasingly exclusive in gene trees arising from this history; thus the *gsi* values of "a" and "b" will increase over the course of the simulation, even when these populations are panmictic with respect to each other.

204 The *pwgsi* value increased for the "a:b" comparison as the number of migrants exchanged
205 between these populations decreased (and thus the populations became more structured) (Fig 7).
206 We also investigated the power of the *pwgsi* to detect population structure in these simulations by
207 performing 10^4 group-label permutations per simulation. Significant results were obtained even with
208 relatively limited population structure (Fig . 7). For example, when $4N_em = 100$, a value giving a
209 negligible expected F_{ST} of < 0.01 (Wright 1949), more than 10% of simulations produced a result with
210 a P-value < 0.05.

211 Although speciation with gene flow is certainly possible (Emelianov *et al.* 2004; Davison *et al.*
212 2005; Niemiller *et al.* 2008; de León *et al.* 2010) and perhaps even common (Nosil 2008), it is
213 generally accepted even a small number of successful migrants are enough to prevent speciation in
214 the absence of selection (Slatkin 1995; Gavrilets 2000). Speciation is only possible with greater rates
215 of migration when very strong divergent selection is acting (Felsenstein 1981; Kirkpatrick & Ravigné
216 2002) . Clearly then, the results of the *pwgsi* cannot be treated as unambiguous evidence that the
217 groups being considered are different species. Rather, researchers need to consider it in the design
218 of their studies and the interpretation of results. In particular, the *pwgsi* may be a poor choice of
219 statistic if a putative species is known to have a distinct geographic distribution with respect to
220 others to which it is being compared (or if the population samples being analysed come from
221 different regions).

222 By contrast, our finding that the *pwgsi* measures population structure may make it a useful
223 statistic for within-species phylogeographic studies, complementing the AMOVA approach (Excoffier
224 et al. 1992) that is currently widely used for sequence data in this context. Indeed, the *gsi* has
225 already been used in this context (Chen & Hare 2011; Gustafsson & Olsson 2012).

226 ## *Conclusions*

Genetic sequences are a potentially powerful source of data for the discovery and delimitation of species. The results reported above, however, emphasise the care that needs to be taken in interpreting the results of DNA-based species-delimitation methods. We have shown that a naïve interpretation of *gsi*, a statistic that has been widely used in species-delimitation studies, can lead to erroneous conclusions. Although the *gsi* remains powerful approach to species when applied carefully (as in the *pwgsi* described here) it is still possible to obtain large *gsi* values and statistically significant results from populations connected by substantial gene flow.

Thus, we suggest that this statistic and other species-delimitation methods should be used as part of a genuinely integrative approach to taxonomy. In particular, the phylogenetic and population-genetic signals measured by species-delimitations methods should considered within the biological and ecological setting of a study, rather than as final arbiters of species' status applied to hypotheses generated by other data.

Acknowledgements

We thank three anonymous reviewers, Steffen Klaere, Giulio Dalla Riva, and Michael Cummings, whose comments on this manuscript greatly improved it. SK in particular should be credited with the insight described in the footnote on pgX.

References

Almendra AL, Rogers DS, González-Cózatl FX (2014) Molecular phylogenetics of the *Handleyomys chapmani* complex in Mesoamerica. *Journal of Mammalogy*, **95**, 26–40.

Ardila NE, Giribet G, Sánchez JA (2012) A time-calibrated molecular phylogeny of the precious corals: reconciling discrepancies in the taxonomic classification and insights into their evolutionary history. *BMC Evolutionary Biology*, **12**, 246.

Ashalakshmi NC, Nag KSC, Karanth KP (2014) Molecules support morphology: species status of South Indian populations of the widely distributed Hanuman langur. *Conservation Genetics*, **16**, 43–58.

Bagley JC, Alda F, Breitman MF *et al.* (2015) Assessing species boundaries using multilocus species delimitation in a morphologically conserved group of neotropical freshwater fishes, the *Poecilia sphenops* species complex (Poeciliidae) *PLoS ONE*, **10**, e0121139.

Bon M-C, Hoelmer KA, Pickett CH *et al.* (2015) Populations of *Bactrocera oleae* (Diptera: Tephritidae) and Its Parasitoids in Himalayan Asia. *Annals of the Entomological Society of America*, sav114.

Boykin LM, Schutze MK, Krosch MN *et al.* (2014) Multi-gene phylogenetic analysis of south-east Asian pest members of the Bactrocera dorsalis species complex (Diptera: Tephritidae) does not support current taxonomy. *Journal of Applied Entomology*, **138**, 235–253.

Camargo A, Morando M, Avila LJ, Sites JW Jr (2012) Species delimitation with ABC and other coalescent-based methods: a test of accuracy with simulations and an empirical example with lizards of the Liolaemus darwinii complex (Squamata: Liolaemidae). *Evolution*, **66**, 2834–2849.

Carstens BC, Pelletier TA, Reid NM, Satler JD (2013) How to fail at species delimitation. *Molecular Ecology*, **22**, 4369–4383.

Cesar S. Herrera CL (2013) Revision of the genus Corallomycetella with *Corallonectria gen. nov*. for *C. jatrophae* (Nectriaceae, Hypocreales). *Mycosystema*, **32**, 518–544.

269 Chen G, Hare MP (2011) Cryptic diversity and comparative phylogeography of the estuarine copepod
270 *Acartia tonsa* on the US Atlantic coast. *Molecular Ecology*, **20**, 2425–2441.
271 Corcoran P, Dettman JR, Sun Y *et al.* (2014) A global multilocus analysis of the model fungus
272 *Neurospora* reveals a single recent origin of a novel genetic system. *Molecular Phylogenetics*
273 *and Evolution*, **78**, 136–147.
274 Costanzo KS, Taylor DJ (2010) Rapid ecological isolation and intermediate genetic divergence in
275 lacustrine cyclic parthenogens. *BMC Evolutionary Biology*, **10**, 166.
276 Cummings MP, Neel MC, Shaw KL (2008) A genealogical approach to quantifying lineage divergence.
277 *Evolution*, **62**, 2411–2422.
278 Davison A, Chiba S, Barton NH, Clarke B (2005) Speciation and Gene Flow between Snails of Opposite
279 Chirality. *PLoS Bioogyl*, **3**, e282.
280 Derkarabetian S, Hedin M (2014) Integrative taxonomy and species delimitation in harvestmen: a
281 revision of the western North American genus *Sclerobunus* (Opiliones: Laniatores:
282 Travunioidea) (W Arthofer, Ed,). *PLoS ONE*, **9**, e104982.
283 Donald KM, Preston J, Williams ST *et al.* (2012) Phylogenetic relationships elucidate colonization
284 patterns in the intertidal grazers *Osilinus* Philippi, 1847 and *Phorcus* Risso, 1826
285 (Gastropoda: Trochidae) in the northeastern Atlantic Ocean and Mediterranean Sea.
286 *Molecular Phylogenetics and Evolution*, **62**, 35–45.
287 Doyle VP, Oudemans PV, Rehner SA, Litt A (2013) Habitat and host indicate lineage identity in
288 *Colletotrichum gloeosporioides s.l.* from wild and agricultural landscapes in North America.
289 *PLoS ONE*, **8**, e62394.
290 Edwards DL, Knowles LL (2014) Species detection and individual assignment in species delimitation:
291 can integrative data increase efficacy? *Proceedings of the Royal Society of London B:*
292 *Biological Sciences*, **281**, 20132765.

293 Egea E, David B, Choné T *et al.* (2016) Morphological and genetic analyses reveal a cryptic species
294 complex in the echinoid *Echinocardium cordatum* and rule out a stabilizing selection
295 explanation. *Molecular Phylogenetics and Evolution*, **94**, 207–220.
296 Emelianov I, Marec F, Mallet J (2004) Genomic evidence for divergence with gene flow in host races
297 of the larch budmoth. *Proceedings of the Royal Society B: Biological Sciences*, **271**, 97–105.
298 Ence DD, Carstens BC (2011) SpedeSTEM: a rapid and accurate method for species delimitation.
299 *Molecular Ecology Resources*, **11**, 473–480.
300 Escobar D, Zea S, Sánchez JA (2012) Phylogenetic relationships among the Caribbean members of the
301 *Cliona viridis* complex (Porifera, Demospongiae, Hadromerida) using nuclear and
302 mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, **64**, 271–284.
303 Esposito LA, Bloom T, Caicedo-Quiroga L *et al.* (2015) Islands within islands: diversification of tailless
304 whip spiders (Amblypygi, Phrynus) in Caribbean caves. *Molecular Phylogenetics and*
305 *Evolution*, **93**, 107–117.
306 Faustová M, Sacherová V, Sheets HD, Svensson J-E, Taylor DJ (2010) Coexisting Cyclic Parthenogens
307 Comprise a Holocene Species Flock in *Eubosmina*. *PLoS ONE*, **5**, e11623.
308 Felsenstein J (1981) Skepticism towards Santa Rosalia, or why are there so few kinds of animals?
309 *Evolution*, 124–138.
310 Fernández-Mazuecos M, Vargas P (2014) Quaternary radiation of bifid toadflaxes (*Linaria* sect.
311 *Versicolores*) in the Iberian Peninsula: low taxonomic signal but high geographic structure of
312 plastid DNA lineages. *Plant Systematics and Evolution*, **301**, 1411–1423.
313 Fourie A, Wingfield MJ, Wingfield BD, Barnes I (2014) Molecular markers delimit cryptic species in
314 *Ceratocystis sensu stricto*. *Mycological Progress*, **14**, 1020.
315 Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C (2012) Coalescent-based species
316 delimitation in an integrative taxonomy. *Trends in ecology & evolution*, **27**, 480–488.

317 Fusinatto LA, Alexandrino J, Haddad CFB *et al.* (2013) Cryptic Genetic diversity is paramount in small-
318 bodied amphibians of the genus *Euparkerella* (Anura: Craugastoridae) endemic to the
319 Brazilian Atlantic forest (D Fontaneto, Ed,). *PLoS ONE*, **8**, e79504.
320 Gavrilets S (2000) Waiting time to parapatric speciation. *Proceedings of the Royal Society B:*
321 *Biological Sciences*, **267**, 2483–2492.
322 Gazis R, Rehner S, Chaverri P (2011) Species delimitation in fungal endophyte diversity studies and
323 its implications in ecological and biogeographic inferences. *Molecular Ecology*, **20**, 3001–
324 3013.
325 Gehesquière B, Crouch JA, Marra RE *et al.* (In Press) Characterization and taxonomic reassessment of
326 the box blight pathogen *Calonectria pseudonaviculata* , introducing *Calonectria henricotiae*
327 *sp. nov*. *Plant Pathology*.
328 Groeneveld LF, Blanco MB, Raharison J-L *et al.* (2010) MtDNA and nDNA corroborate existence of
329 sympatric dwarf lemur species at Tsinjoarivo, eastern Madagascar. *Molecular Phylogenetics*
330 *and Evolution*, **55**, 833–845.
331 Gustafsson DR, Olsson U (2012) Flyway homogenisation or differentiation? Insights from the
332 phylogeny of the sandpiper (Charadriiformes: Scolopacidae: Calidrinae) wing louse genus
333 *Lunaceps* (Phthiraptera: Ischnocera). *International Journal for Parasitology*, **42**, 93–102.
334 Hendrixson BE, DeRussy BM, Hamilton CA, Bond JE (2013) An exploration of species boundaries in
335 turret-building tarantulas of the Mojave Desert (Araneae, Mygalomorphae, Theraphosidae,
336 Aphonopelma). *Molecular Phylogenetics and Evolution*, **66**, 327–340.
337 Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation.
338 *Bioinformatics*, **18**, 337.
339 Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept.
340 *Evolution; international journal of organic evolution*, **56**, 1557–1565.
341 Hu H, Al-Shehbaz IA, Sun Y *et al.* (2015) Species delimitation in Orychophragmus (Brassicaceae)
342 based on chloroplast and nuclear DNA barcodes. *Taxon*, **64**, 714–726.

343 Jones G, Aydin Z, Oxelman B (2015) DISSECT: an assignment-free Bayesian discovery method for
344 species delimitation under the multispecies coalescent. *Bioinformatics*, **31**, 991–998.
345 Keith R, Hedin M (2012) Extreme mitochondrial population subdivision in southern Appalachian
346 paleoendemic spiders (Araneae: Hypochilidae: *Hypochilus*), with implications for species
347 delimitation. *Journal of Arachnology*, **40**, 167–181.
348 Kirkpatrick M, Ravigné V (2002) Speciation by natural and sexual selection: models and experiments.
349 *The American Naturalist*, **159**, S22–S35.
350 Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. *Systematic*
351 *Biology*, **56**, 887–895.
352 Koopman MM, Baum DA (2010) Isolating nuclear genes and identifying lineages without monophyly:
353 an example of closely related species from southern Mmadagascar. *International Journal of*
354 *Plant Sciences*, **171**, 761–771.
355 De León LF, Bermingham E, Podos J, Hendry AP (2010) Divergence with gene flow as facilitated by
356 ecological differences: within-island variation in Darwin's finches. *Philosophical Transactions*
357 *of the Royal Society B: Biological Sciences*, **365**, 1041 –1052.
358 Levsen ND, Tiffin P, Olson MS (2012) Pleistocene Speciation in the Genus Populus (Salicaceae).
359 *Systematic Biology*, **61**, 401–412.
360 Lopes JS, Balding D, Beaumont MA (2009) PopABC: a program to infer historical demographic
361 parameters. *Bioinformatics*, **25**, 2747–2749.
362 Lu B, Bi K, Fu J (2014) A phylogeographic evaluation of the *Amolops mantzorum* species group:
363 Cryptic species and plateau uplift. *Molecular Phylogenetics and Evolution*, **73**, 40–52.
364 Martinsson S, Rhodén C, Erséus C (In press) Barcoding gap, but no support for cryptic speciation in
365 the earthworm *Aporrectodea longa* (Clitellata: Lumbricidae). *Mitochondrial DNA*, 1–9.
366 Medina R, Lara F, Goffinet B, Garilleti R, Mazimpaka V (2012) Integrative taxonomy successfully
367 resolves the pseudo-cryptic complex of the disjunct epiphytic moss *Orthotrichum consimile*
368 *s.l.* (Orthotrichaceae). *Taxon*, **61**, 1180–1198.

369 Niemiller ML, Fitzpatrick BM, Miller BT (2008) Recent divergence with gene flow in Tennessee cave
370 salamanders (Plethodontidae: Gyrinophilus) inferred from gene genealogies. *Molecular*
371 *Ecology*, **17**, 2258–2275.
372 Niemiller ML, McCandless JR, Reynolds RG *et al.* (2013) Effects of climatic and geological processes
373 during the pleistocene on the evolutionary history of the northern cavefish, *Amblyopsis*
374 *spelaea* (Teleostei: Amblyopsidae). *Evolution*, **67**, 1011–1025.
375 Niemiller ML, Near TJ, Fitzpatrick BM (2012) Delimiting species using multilocus data: diagnosing
376 cryptic diversity in the southern cavefish, *Typhlichthys subterraneus* (Teleostei:
377 Amblyopsidae). *Evolution*, **66**, 846–866.
378 Nosil P (2008) Speciation with gene flow could be common. *Molecular Ecology*, **17**, 2103–2106.
379 Olave M, Solà E, Knowles LL (2014) Upstream analyses create problems with dna-based species
380 delimitation. *Systematic Biology*, **63**, 263–271.
381 O'Meara BC (2010) New heuristic methods for joint species delimitation and species tree inference.
382 *Systematic Biology*, **59**, 59–73.
383 Parmakelis A, Kotsakiozi P, Stathi I, Poulikarakou S, Fet V (2013) Hidden diversity of *Euscorpius*
384 (Scorpiones: Euscorpiidae) in Greece revealed by multilocus species-delimitation
385 approaches. *Biological journal of the Linnean Society*, **110**, 728–748.
386 Pažoutová S, Pešicová K, Chudíčková M, utka PŠ, Kolařík M (2015) Delimitation of cryptic species
387 inside *Claviceps purpurea*. *Fungal Biology*, **119**, 7–26.
388 Pérez G, Burgess TI, Slippers B *et al.* (2013) *Teratosphaeria pseudonubilosa sp. nov*., a serious
389 Eucalyptus leaf pathogen in the *Teratosphaeria nubilosa* species complex. *Australasian Plant*
390 *Pathology*, **43**, 67–77.
391 Pettengill JB, Moeller DA (2012) Tempo and mode of mating system evolution between incipient
392 *Clarkia* species: temporal dynamics of mating system evolution. *Evolution*, **66**, 1210–1225.

393 Pino-Bodas R, Ahti T, Stenroos S, Martin MP, Burgaz AR (2013) Multilocus approach to species
394 recognition in the *Cladonia humilis* complex (Cladoniaceae, Ascomycota). *American Journal*
395 *of Botany*, **100**, 664–678.
396 Pons J, Barraclough TG, Gomez-Zurita J *et al.* (2006) Sequence-based species delimitation for the
397 DNA taxonomy of undescribed insects. *Systematic biology*, **55**, 595–609.
398 Prévot V, Jordaens K, Sonet G, Backeljau T (2013) Exploring species level taxonomy and species
399 delimitation methods in the facultatively self-fertilizing land snail genus *Rumina*
400 (Gastropoda: Pulmonata). *PLoS ONE*, **8**, e60736.
401 Ramirez JL, Miyaki CY, Frederick PC, Lama SND (2014) Species delimitation in the genus *Eudocimus*
402 (Threskiornithidae: Pelecaniformes): first genetic approach. *Waterbirds*, **37**, 419–425.
403 Reid NM, Carstens BC (2012) Phylogenetic estimation error can decrease the accuracy of species
404 delimitation: a Bayesian implementation of the general mixed Yule-coalescent model. *BMC*
405 *Evolutionary Biology*, **12**, 196.
406 Sakalidis ML, Hardy GESJ, Burgess TI (2011) Use of the Genealogical Sorting Index (GSI) to delineate
407 species boundaries in the *Neofusicoccum parvum–Neofusicoccum ribis* species complex.
408 *Molecular Phylogenetics and Evolution*, **60**, 333–344.
409 Salariato DL, Zuloaga FO, Al-Shehbaz IA (2012) Morphometric studies and taxonomic delimitation in
410 *Menonvillea scapigera* and related species (Cremolobeae: Brassicaceae). *Plant Systematics*
411 *and Evolution*, **298**, 1961–1976.
412 Sánchez-Ramírez S, Tulloss RE, Guzmán-Dávalos L *et al.* (In Press) In and out of refugia: historical
413 patterns of diversity and demography in the North American Caesar's mushroom species
414 complex. *Molecular Ecology*.
415 Schmidt-Lebuhn AN, de Vos JM, Keller B, Conti E (2012) Phylogenetic analysis of *Primula* section
416 *Primula* reveals rampant non-monophyly among morphologically distinct species. *Molecular*
417 *Phylogenetics and Evolution*, **65**, 23–34.

418 Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies.
419 *Genetics*, **139**, 457.
420 Su X, Wu G, Li L, Liu J (2015) Species delimitation in plants using the Qinghai–Tibet Plateau endemic
421 *Orinus* (Poaceae: Tridentinae) as an example. *Ann Bot*, **116**, 35–48.
422 Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**,
423 437–460.
424 Taole MM, Burgess TI, Gryzenhout M, Wingfield BD, Wingfield MJ (2011) DNA sequence
425 incongruence and inconsistent morphology obscure species boundaries in the
426 *Teratosphaeria suttonii* species complex. *Mycoscience*, **53**, 270–283.
427 Udayanga D, Castlebury LA, Rossman AY, Hyde KD (2014) Species limits in *Diaporthe*: molecular re-
428 assessment of *D. citri, D. cytosporella, D. foeniculina and D. rudis*. *Persoonia - Molecular*
429 *Phylogeny and Evolution of Fungi*, **32**, 83–101.
430 Valcárcel V, Vargas P (2010) Quantitative morphology and species delimitation under the general
431 lineage concept: Optimization for Hedera (Araliaceae). *American Journal of Botany*, **97**,
432 1555–1573.
433 Vigalondo B, Fernández-Mazuecos M, Vargas P, Sáez L (2015) Unmasking cryptic species:
434 morphometric and phylogenetic analyses of the Ibero-North African *Linaria incarnata*
435 complex. *Botanical Journal of the Linnean Society*, **177**, 395–417.
436 Viricel A, Rosel PE (2014) Hierarchical population structure and habitat differences in a highly mobile
437 marine species: the Atlantic spotted dolphin. *Molecular Ecology*, **23**, 5018–5035.
438 Walker DM, Castlebury LA, Rossman AY, White Jr. JF (2012) New molecular markers for fungal
439 phylogenetics: Two genes for species-level systematics in the Sordariomycetes
440 (Ascomycota). *Molecular Phylogenetics and Evolution*, **64**, 500–512.
441 Walstrom VW, Klicka J, Spellman GM (2012) Speciation in the White-breasted Nuthatch (Sitta
442 carolinensis): a multilocus perspective. *Molecular Ecology*, **21**, 907–920.

443 Wang X, Zang R, Yin Z, Kang Z, Huang L (2014) Delimiting cryptic pathogen species causing apple
444 Valsa canker with multilocus data. *Ecology and Evolution*, **4**, 1369–1380.
445 Weisrock DW, Rasoloarison RM, Fiorentino I *et al.* (2010) Delimiting species without nuclear
446 monophyly in Madagascar's mouse lemurs. *PLoS ONE*, **5**, e9883.
447 Willyard A, Wallace LE, Wagner WL *et al.* (2011) Estimating the species tree for Hawaiian *Schiedea*
448 (Caryophyllaceae) from multiple loci in the presence of reticulate evolution. *Molecular*
449 *Phylogenetics and Evolution*, **60**, 29–48.
450 Yang Z, Rannala B (2010) Bayesian species delimitation using multilocus sequence data. *Proceedings*
451 *of the National Academy of Sciences*, **107**, 9264–9269.
452 Yang Z, Rannala B (2014) Unguided Species Delimitation Using DNA Sequence Data from Multiple
453 Loci. *Molecular Biology and Evolution*, **31**, 3125–3135.
454 Zhang C, Zhang D-X, Zhu T, Yang Z (2011) Evaluation of a Bayesian Coalescent Method of Species
455 Delimitation. *Systematic Biology*, **60**, 747–761.
456 Zhao Y, Qi Z, Ma W *et al.* (2013) Comparative phylogeography of the *Smilax hispida* group
457 (Smilacaceae) in eastern Asia and North America – Implications for allopatric speciation,
458 causes of diversity disparity, and origins of temperate elements in Mexico. *Molecular*
459 *Phylogenetics and Evolution*, **68**, 300–311.
460

461 ***Table 1***

Reference	Evidence for group assignment	n. groups
Martinsson *et al.* In press	Clades in mitochondrial gene tree	2
Sánchez-Ramírez *et al.* In Press	Existing taxonomic distinction, clades in gene trees	17
Gehesquière *et al.* In Press	Clustering of haplotypes	2
Egea *et al.* 2016	Mitochondrial clades	5
Bon *et al.* 2015	Existing taxonomic distinction	3
Hu *et al.* 2015	Clades identified in other trees	7
Esposito *et al.* 2015	Clustering of haplotypes	11
Su *et al.* 2015	Clustering of haplotypes	3
Bagley *et al.* 2015	Clades, other delimitation methods, existing taxonomic distinction	10
Vigalondo *et al.* 2015	Clades in multilocus phylogeny and existing taxonomic distinction	4
Pažoutová *et al.* 2015	Existing taxonomic distinction	4
Ramirez *et al.* 2014	Existing taxonomic distinction	2
Fourie *et al.* 2014	Existing taxonomic distinction and geographically isolated populations	17
Fernández-Mazuecos & Vargas 2014	Existing taxonomic distinction	6
Viricel & Rosel 2014	Morphology, clustering of genotypic data and geographic distribution	2
Derkarabetian & Hedin 2014	Morphology, mitochondrial clades	11
Wang *et al.* 2014	Existing taxonomic distinction, clades in gene trees	3
Ashalakshmi *et al.* 2014	Existing taxonomic distinction	4
Udayanga *et al.* 2014	Existing taxonomic distinction and congruent clades among gene trees	10
Almendra *et al.* 2014	Existing taxonomic distinction and clades in gene trees	3
Lu *et al.* 2014	Clades in a multi-locus phylogeny	8
Corcoran *et al.* 2014	Clustering of haplotypes	11
Boykin *et al.* 2014	Existing taxonomic distinction and clades in consensus tree	39
Fusinatto *et al.* 2013	Clades in a multi-locus phylogeny	6
Pérez *et al.* 2013	ITS clades	2
Parmakelis *et al.* 2013	Clades in a multi-locus phylogeny	16
Doyle *et al.* 2013	Clades supported in >=3 of 4 gene trees	16
Pino-Bodas *et al.* 2013	Existing taxonomic distinction and morphological differences	9
Prévot *et al.* 2013	Clades in COI gene tree or combined mitochondrial tree	8
Zhao *et al.* 2013	Existing taxonomic distinction	6
Cesar S. Herrera 2013	Existing taxonomic distinction and geographic distribution	4
Niemiller *et al.* 2013	Existing taxonomic distinction and geographic distribution	4
Hendrixson *et al.* 2013	Existing taxonomic distinction and clades in mitochondrial gene tree	4
Keith & Hedin 2012	Existing taxonomic distinction and geographic distribution	76
Niemiller *et al.* 2012	Species delimitation/assignment via Brownie	19
Ardila *et al.* 2012	Existing taxonomic distinction	13
Walker *et al.* 2012	Existing taxonomic distinction	11
Donald *et al.* 2012	Existing taxonomic distinction and geographic distribution	9
Walstrom *et al.* 2012	Geographic distribution	7
Schmidt-Lebuhn *et al.* 2012	Existing taxonomic distinction	6
Escobar *et al.* 2012	Existing taxonomic distinction	6
Medina *et al.* 2012	Morphological differences	4
Salariato *et al.* 2012	Existing taxonomic distinction	2

Pettengill & Moeller 2012	Existing taxonomic distinction	2
Levsen *et al.* 2012	Existing taxonomic distinction	2
Willyard *et al.* 2011	Existing taxonomic distinction	32
Gazis *et al.* 2011	Clades in consensus tree	16
Taole *et al.* 2011	Clades in ITS gene tree	14
Sakalidis *et al.* 2011	Existing taxonomic distinction and morphological differences	8
Weisrock *et al.* 2010	Existing taxonomic distinction and geographic distribution	16
Faustová *et al.* 2010	Existing taxonomic distinction and geographic distribution	8
Groeneveld *et al.* 2010	Existing taxonomic distinction and morphological differences	4
Valcárcel & Vargas 2010	Existing taxonomic distinction	4
Koopman & Baum 2010	Existing taxonomic distinction	3
Costanzo & Taylor 2010	Existing taxonomic distinction	2

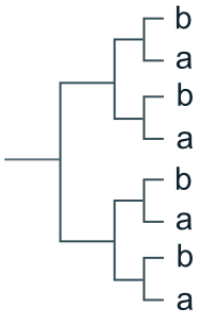
Summary of papers in which the gsi is principally used for species delimitation. "n.groups" refers to the greatest number of putative species considered in a single analysis

Table 2

Method	Primary input	Group assignment	Other parameters	Result
DISSECT (Jones *et al.* 2015)	Alignments	Inferred	$\theta, \tau, \omega, \lambda, \mu$	Posterior distribution of model-paramaters, including species-tree and species- delimitation
BPP (Yang & Rannala 2010, 2014)	Alignments	*A priori*	$\theta, \tau,$ species tree*	Posterior probability that species tree nodes represent speciation events
popABC (Lopes *et al.* 2009; Camargo *et al.* 2012)	Alignments	*A priori*	$\theta, \tau,$ species tree	Posterior probability that species tree nodes represent speciation events
SpedeSTEM (Ence & Carstens 2011)	Gene trees	*A priori*	θ	Likelihood for population-delimitation models
Brownie (O'Meara 2010)	Gene trees	Inferred	None	Joint inference of maximum likelihood species-tree and species-delimitation
GMYC (Pons *et al.* 2006)	Gene tree	Inferred	None	Likelihood for species-delimitation models
gsi [14]	Gene tree	*A priori*	None	Statistic and hypothesis-test measuring each group's phylogenetic exclusivity

"θ" refers to the population mutation rate for each locus, "τ" to the over-all height of a species tree in coalescent units. For DISCUSS, the ω parameter specifies a prior on the on the number of species present in a dataset, and λ and μ represent the speciation and extinction rates of a birth-death processes generating the underlying species tree. The description of the popABC approach to species delimitation refers specifically to the method employed by Camargo et al (Camargo *et al.* 2012)

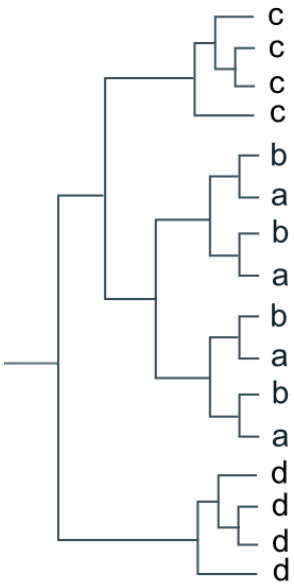
* Note the species tree is an optional input parameter for BPP



472

473 ***Figure 1***

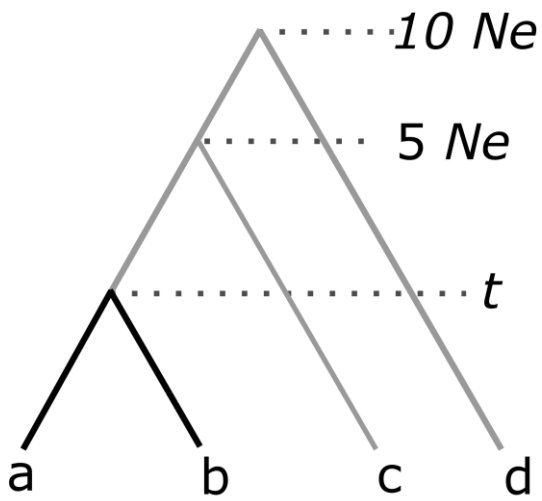
474 Hypothetical phylogenetic tree, with tips assigned to two groups “a” and “b”



475

476 ***Figure 2***

477 Hypothetical phylogenetic tree, obtained by adding two additional groups (“c” and “d”) to
478 the tree presented in Figure 2.
479

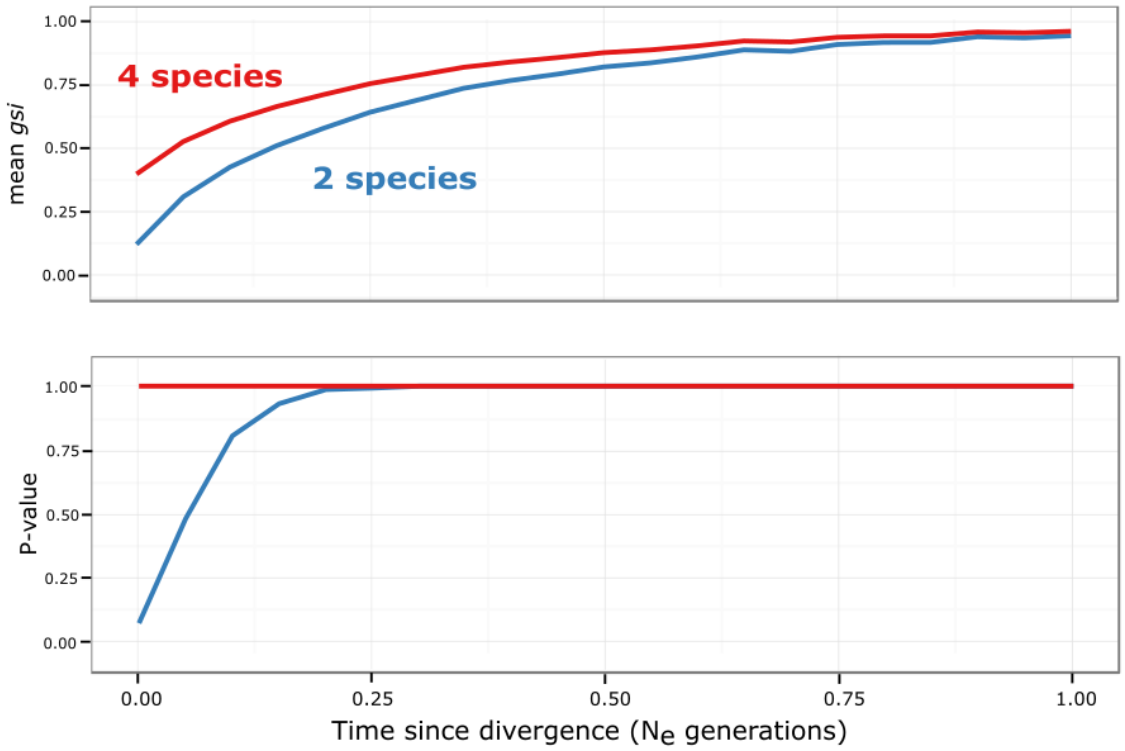


480

481 ***Figure 3***

482 Demographic history under which simulations were performed. In each simulation *gsi* was
483 calculating from a gene tree containing 10 individuals from each population (the four-population
484 tree) and from a gene tree from which tips corresponding to individuals from populations “c” and
485 “d” had been dropped (the two-population tree).

486

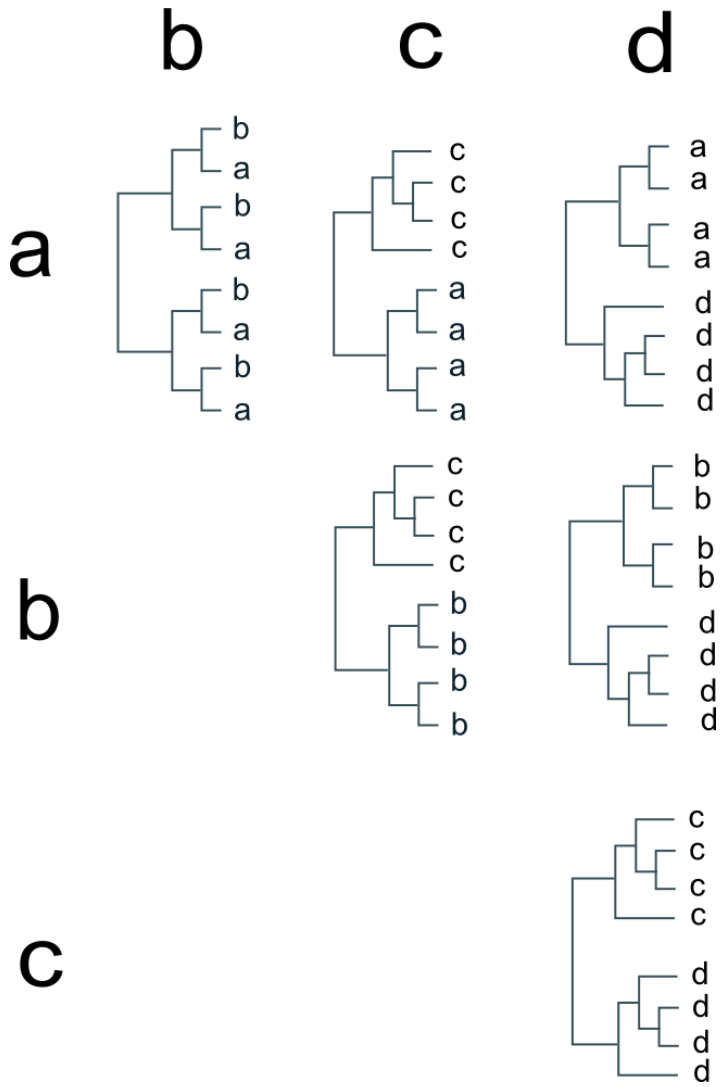


487

488 ***Figure 4***

489 Above: The *gsi* value obtained for a group depends on the overall structure of the tree from
490 which it is calculated. The red line represents the mean *gsi* value calculated for group "a" in the
491 simulation depicted in Figure 3 when the all four populations are included in the calculation. The
492 blue line is the mean *gsi* calculated for the same population when the divergent populations "c" and
493 "d" are first discarded.

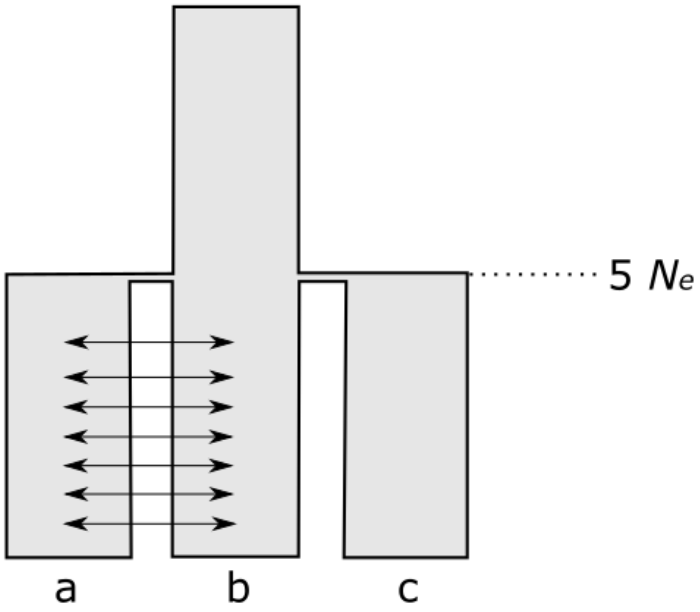
494 Below: The hypothesis test usually performed alongside the calculation of *gsi* readily produces
495 significant results for populations that are not divergent from all other populations. The red line
496 represents the proportion of simulations in which population "a" was found to have a significant
497 pattern of exclusivity in the four-population tree. The blue line represents the same proportion
498 calculated from the two-population tree.



499

500 ***Figure 5***

501 Trees used for pairwise comparison of all groups in Figure 2.



502

503 ***Figure 6***

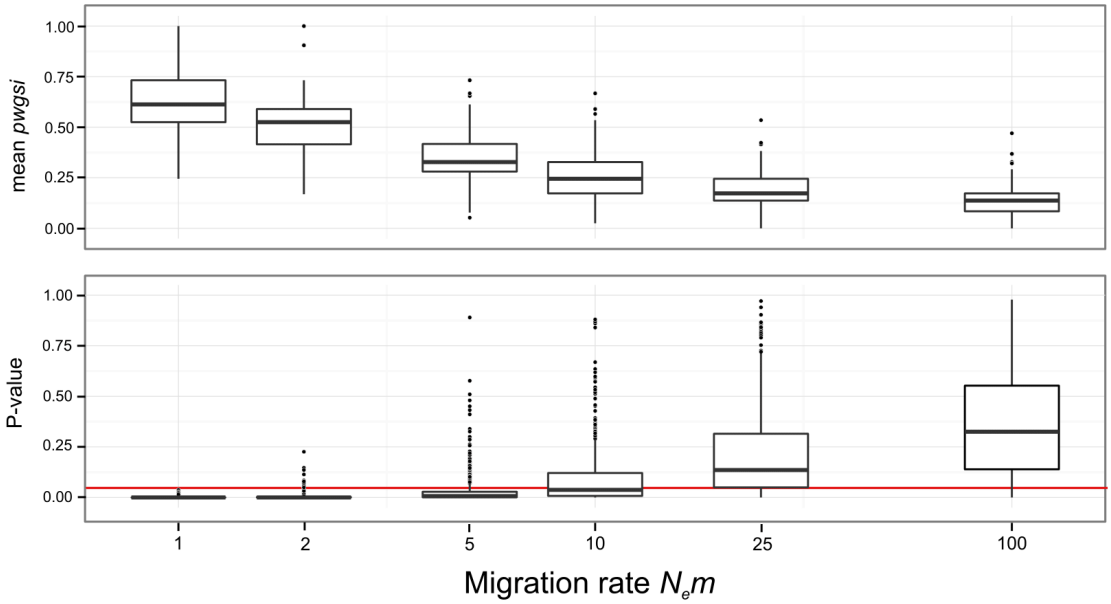
504 Demographic history under which population structure simulations were performed. Arrows

505 represent ongoing gene flow between populations “a” and “b” after their divergence at 5 N_e

506 generations.

507

508



509

510 ***Figure 7***

511 The *pwgsi* measures population structure as well as lineage divergence.

512 Above: In the top panel, boxes represent distributions of *pwgsi* for the groups “a” and “b”,

513 which were sampled from populations experiencing gene flow at a constant rate which varies along

514 the x-axis (large values of N_em correspond to more migration and thus less population-structure).

515 Below:. Boxes represent distributions of P-values obtained for the “a:b” comparison; red line the

516 drawn at $P = 0.05$.Note, note x-axis is on a log-10 scale.

517