Genetica (2008) 132:219–225
DOI 10.1007/s10709-007-9167-0

Does organismal pedigree impact the magnitude of topological congruence among gene trees for unlinked loci?

Chih-Horng Kuo · John C. Avise

Received: 31 July 2006 / Accepted: 19 June 2007 / Published online: 17 July 2007

Abstract One of the fundamental assumptions in the multi-locus approach to phylogeographic studies is that unlinked loci have independent genealogies. For this reason, congruence among gene trees from unlinked loci is normally interpreted as support for the existence of external forces that may have concordantly shaped the topology of multiple gene trees. However, it is also important to address and quantify the possibility that gene trees within a given species are all inherently constrained to some degree by their shared organismal pedigree, and thus in this strict sense are not entirely independent. Here we demonstrate by computer simulations that gene trees from a shared pedigree tend to display higher topological concordance than do gene trees from independent pedigrees with the same demographic parameters, but we also show that these constraining effects are normally minor in comparison to the much higher degree of topological concordance that can routinely emerge from external phylogeographic shaping forces. The topology-constraining effect of a shared pedigree decreases as effective population size increases, and becomes almost negligible in a random mating population of more than 1,000 individuals. Moreover, statistical detection of the pedigree effect requires a relatively large number of unlinked loci that far exceed what is typically used in current phylogeographic studies. Thus, with the possible exception of extremely small populations, multiple unlinked genes within a pedigree can indeed be assumed, for most practical purposes, to have independent genealogical histories.

Keywords Dispersal · Gene flow · Genealogy · Phylogeny · Phylogeography · Vicariance

C.-H. Kuo (✉)
Department of Genetics, University of Georgia, Athens, GA 30602, USA
e-mail: chkuo@uga.edu

J. C. Avise
Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA

Introduction

Lineage sorting and gene coalescence are inherently rather stochastic processes, so genealogies of physically unlinked genes within a sexually reproducing species are expected to exhibit considerable topological incongruence unless external forces somehow act on populations to shape multiple gene trees concordantly. Examples of such external forces include limited dispersal, vicariant events, and other biological and phylogeographic factors that promote strong population genetic structure (Avise 2000). However, it is also true that all gene trees within a species are confined within one underlying organismal pedigree and, thus, in that sense are not fully independent. In fact, there are 2^G possible gender-defined transmission routes for autosomal alleles traced back through G generations, but only four of them (F → F → F → F → …; F → M → F → M → …; M → F → M → F → …; and M → M → M → M → …; where F and M are males and females) are completely non-coincident in every generation (Wollenberg and Avise 1998).

Based on this observation, it seems at least theoretically possible that the underlying organismal pedigree could materially constrain the distribution of gene tree topologies even in the absence of other external forces. In other words, multiple gene trees sampled from a single organismal pedigree might show a higher level of topological concordance than gene trees sampled from independent pedigrees with

comparable sets of demographic variables. One previous study found that organismal pedigree does not greatly affect the distribution of mean coalescent times in a random mating population (Ball et al. 1990). However, vastly different gene tree topologies can produce similar mean pairwise coalescent times, so measuring topological concordance across gene trees in random mating and structured populations should be another key aspect of evaluating the possible constraining impact of organismal pedigree.

In this study we directly assess gene tree topologies, via computer simulations, to address the following questions. Do gene trees for unlinked loci sampled from a shared pedigree show higher genealogical concordance than do truly independent trees from separate pedigrees of same demography? If so, how strong is the effect compared to other external factors such as the level of gamete dispersal? How many gene trees must be sampled from organismal pedigrees to detect such effects? Answers to these questions could be important to phylogeographical studies, especially if non-independence among gene trees would turn out to impede our ability to infer the evolutionary histories of organisms. While several sources of the non-independence among gene trees (e.g., physical linkage and functional constraint) have been subject to intensive studies, the effect of shared organismal pedigree has not previously been fully characterized.

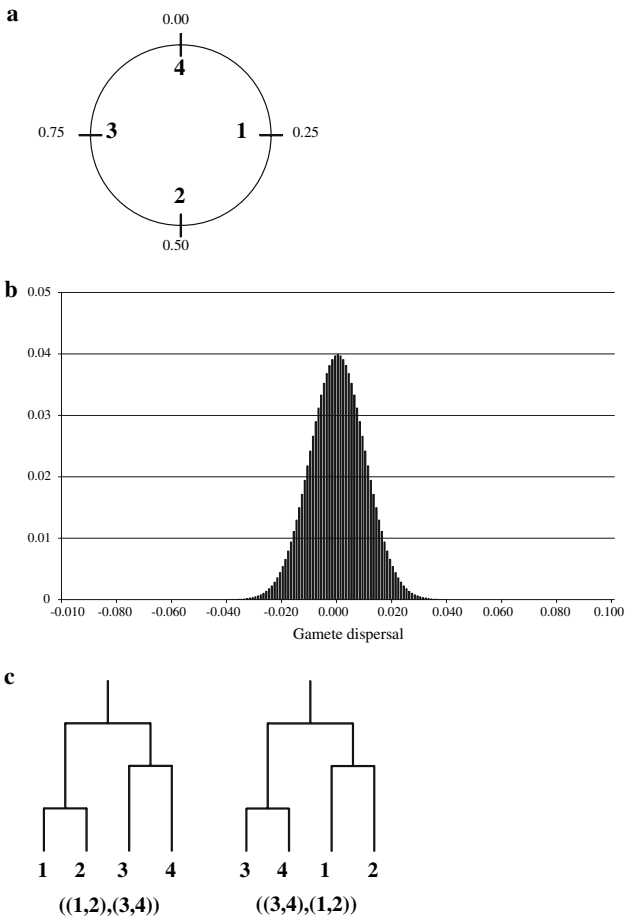


Fig. 1 The simulation model. (**a**) the one-dimensional ring habitat (numbers inside the circle represent alleles sampled for generating gene trees, and numbers outside the circle indicate the position of each allele); (**b**) a sample probability distribution of gamete dispersal when $\sigma_{disp} = 0.01$ (a positive number corresponds to a counter-clockwise movement and a negative number corresponds to a clockwise movement); (**c**) two tree topologies that are considered non-equivalent in this model because of their differences in branching order

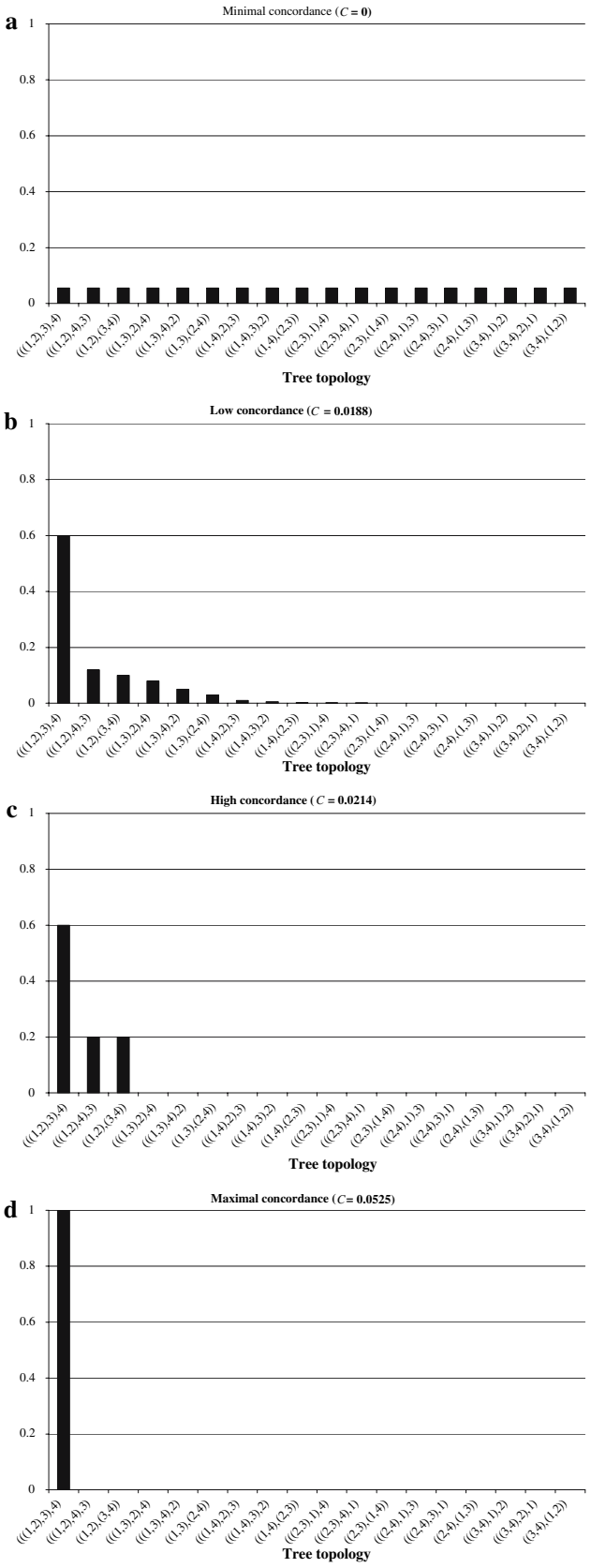
Methods

Model description

We adapted the simulation model from Kuo and Avise (2005) to simulate the coalescent process in a one-dimensional ring-shaped habitat (Fig. 1a). As pointed out by Wilkins and Wakeley (2002), a linear non-ring habitat can produce biased patterns because mean coalescent times tend to be greater at the habitat center. We choose a ring-shaped habitat in our simulation model to ensure all possible topologies of gene genealogy are equally probable. The population is uniformly distributed and has a distribution range of [0, 1], with positions 0 and 1 being equivalent (hence the ring shape). Organisms are diploid, sexually reproducing, monoecious, capable of self-fertilization, and have discrete generations. The source code of our simulation program is written in C++ programming language and is freely available at http://chkuo.name/

The simulation process traces backward in time to generate each organismal pedigree one at a time. In each generation, two random numbers are drawn from a normal distribution with mean = 0 and standard deviation = σ_{disp} (see Fig. 1b for an example) for each individual. These two random numbers represent the dispersal of gametes from the individual's parents; a positive value represents a counter-clockwise movement of the gamete along the one-dimensional ring habitat whereas a negative value represents a clockwise movement. The physical location of a parent could be calculated by adding the random number to a progeny's position and rounding to the nearest monitored compass point on the ring. Such information in composite is regarded as the genealogical history (i.e., the pedigree) of the population and is used for generating multiple gene trees.

After a pedigree of the population is determined, gene trees could be generated by tracing the transmission history of alleles backward in time. Alleles have an equal probability of tracing backward in any generation by either going through the male or the female parent. When two alleles trace back to the same diploid individual in the previous



◀**Fig. 2** Hypothetical examples of different levels of genealogical concordance (*C*). (**a**) minimal concordance when all gene tree topologies have the same frequency; (**b**) low concordance when one topology is strongly supported and several other topologies are weakly supported; (**c**) high concordance when one topology is strongly supported and few other topologies are weakly supported; (**d**) maximum concordance when all gene trees share one topology

generation, the probability of a coalescent event (i.e., when the two alleles are identical by descent) is 0.5. To generate each gene tree, we sampled a set of alleles along this one-dimensional habitat and traced backward in time until all allelic lineages coalesced in one shared ancestor. Topology and branch lengths (measured in generations) were recorded for each gene tree, and results were tabulated into frequency distributions of various outcomes.

For a set of four alleles, there are 18 possible topologies when branching order is considered. In this situation, ((1,2),(3,4)) is not equivalent to ((3,4),(1,2)) (see Fig. 1c). This definition of tree topology is different from conventional definition where ((1,2),(3,4)) is equivalent to ((3,4),(1,2)) and 15 different topologies are possible for a bifurcating rooted gene tree with four labeled terminal nodes (Felsenstein, 1978). Similarly, there are 180 possible topologies for a tree of five alleles when branching order is considered (as opposed to 105 possible topologies when branching order is not considered). We take branching order into account when recording tree topology such that all topologies are equally probable in a random mating population.

From the output of each simulation, we used the variance in the frequencies of all possible tree topologies (Fig. 2) as our measurement of genealogical concordance (*C*). Thus, genealogical concordance is the lowest ($C = 0$) when all gene tree topologies are equally frequent (Fig. 2 a), and it is maximized (at $C = 0.0525$ when four alleles are sampled) when all gene trees have the same topology (Fig. 2d). We choose *C* (rather than the frequency of gene trees consistent with the best supported topology) as our primary measurement of genealogical concordance because it is more informative. For example, the best supported topologies in Fig. 2b and 2c were in each case displayed by 60% of all gene trees sampled, but these two cases clearly show different levels of genealogical concordance that can be distinguished by examining *C*.

Simulation settings and data analysis

Our simulation program includes three sampling schemes. The first sampling scheme starts with four equally spaced alleles (at positions 0, 0.25, 0.5, and 0.75) sampled from the current generation. The second sampling scheme starts

with four unequally spaced alleles at positions 0, 0,125, 0,25, and 0.5. The third sampling scheme starts with five equally spaced alleles at positions 0, 0.2, 0.4, 0.6, and 0.8.

To test the effects of organismal pedigree on gene tree topologies under different gene flow regimes, we used the following simulation settings for gamete dispersal: σ_{disp} = 0.001, 0.01, 0.1, and ∞. When $\sigma_{disp} = \infty$ (equivalent to a random mating population model), a gamete has an equal probability of dispersing to any possible location on the ring habitat. For each level of gamete dispersal, we generated 1,800 gene trees either from the same organismal pedigree or from 1,800 independent "replicate" pedigrees. For each set of 1,800 gene trees, we recorded the frequency distribution of tree topologies and used the variance of that distribution (C) as a measure of genealogical concordance.

To investigate sample sizes required to detect the pedigree effect, we sampled different numbers of gene trees from one pedigree and compared the resulting levels of genealogical concordance to those from the same numbers of independent pedigrees. The numbers of gene trees examined were 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1,024. For this part of simulations, we sampled four equally spaced alleles from the population. Two extreme levels of gamete dispersal (i.e., $\sigma_{disp} = 0.001$ for low gamete dispersal and $\sigma_{disp} = \infty$ for random mating) were examined and the population size was set to N = 1,000 individuals (equivalent to 2,000 alleles) in each generation.

To test the hypothesis that a shared pedigree has a higher degree of constraining effect on genealogical concordance in smaller populations, we varied the population size setting under two different levels of gamete dispersal (i.e., σ_{disp} = 0.01 for limited gamete dispersal and $\sigma_{disp} = \infty$ for random mating). In all simulations we performed 1,000 iterations under each parameter combination and then used a one-tailed t-test to assess if multiple gene trees from the same organismal pedigree showed significantly higher topological concordance than gene trees from separate pedigrees.

Results and discussion

The notion that gene trees within an organismal pedigree can show topological disparities is not new (Avise and Wollenberg 1997; Maddison 1997; Hudson and Coyne 2002; Degnan and Salter 2005). However, the extent if any to which a pedigree constrains such disparities (compared to outcomes from independent pedigrees with the same demography) is less well understood. Here we have addressed this issue using computer simulations.

Effects of organismal pedigree and gamete dispersal

Under all levels of gamete dispersal, gene trees sampled from the same organismal pedigree consistently showed a significantly higher level of genealogical concordance (C) than did gene trees sampled from independent pedigrees (Fig. 3). Furthermore, the effect of pedigree on C increased as the level of gamete dispersal decreased (see P-values in Fig. 3).

Although the effects of shared pedigree on C were statistically significant, the magnitudes of the differences in C were trivial compared to those arising from different levels

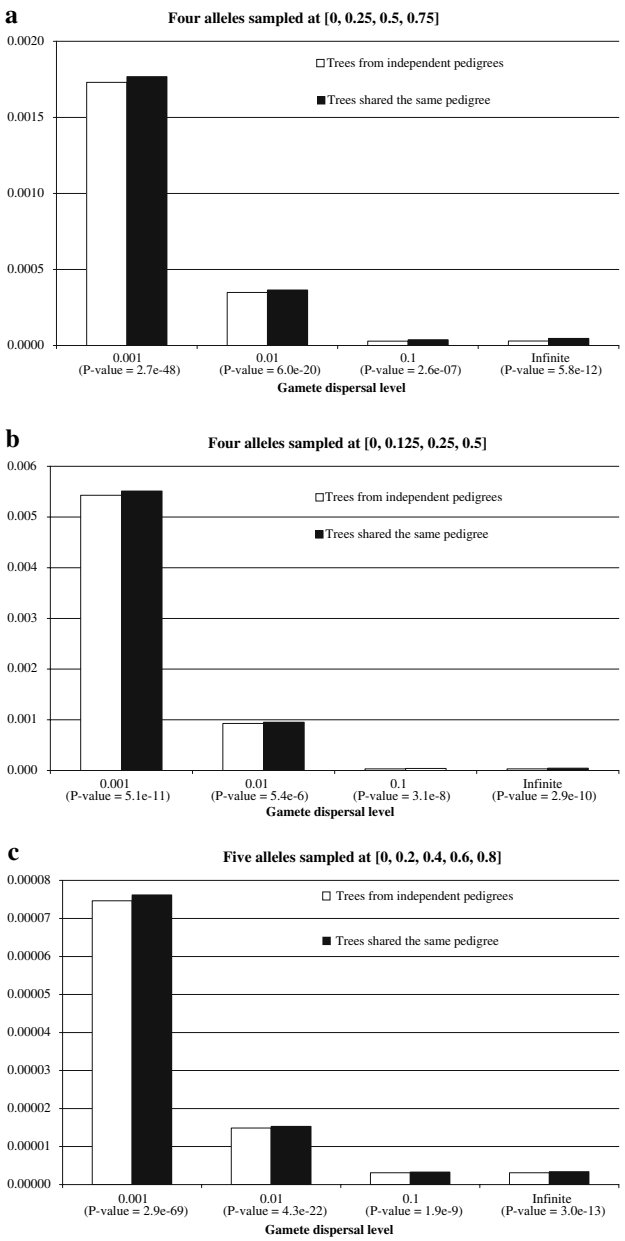


Fig. 3 Magnitudes of genealogical concordance (C) when gene trees from the same or independent pedigrees are monitored at different levels of gamete dispersal. (**a**) four alleles sampled at [0, 0.25, 0.5, 0.75]; (**b**) four alleles sampled at [0, 0.125, 0.25, 0.5]; (**c**) five alleles sampled at [0, 0.2, 0.4, 0.6, 0.8]. P-values from one-tailed t-tests are listed below the X-axis. Error bars for indicating 95% confidence intervals are not shown because they are not discernible at this scale

of gametic dispersal (Fig. 3). A previous study (Wollenberg and Avise 1998) showed that the mean pairwise coalescent time has a higher correlation under limited dispersal. Our data have extended this result and shown that the topological concordance among gene genealogies increased dramatically as gamete dispersal decreased, regardless of whether gene trees were sampled from the same or from independent pedigrees. The explanation for this phenomenon is that coalescent events involving neighboring individuals are more likely when gamete dispersal is limited. By comparing different sampling schemes (Fig. 3a, 3b), one can see that the topological concordance is higher when alleles are unequally spaced in the current generation and gamete dispersal is limited. Under this sampling scheme, allele at position 0.0125 is more likely to coalesces with allele at position 0 or 0.25 before it coalesce with allele at position 0.5.

When alleles sampled from a population are equally spaced in the current generation, increasing sample size from four alleles to five alleles generated qualitatively similar results (cf. Fig. 3a, 3c). Because the number of possible topologies increases with the number of alleles sampled, the level of concordance decreases as the number of alleles increases.

Effects of the number of gene trees sampled

Genealogical concordance also appeared to increase as the sampled number of gene trees decreased in random mating populations (Fig. 4). However, the high level of C observed when few gene genealogies were sampled is merely an artifact of a small sample size of loci. The most extreme example occurs when only one gene tree is examined, in which case concordance is by definition maximal. With respect to possible effects of organismal pedigree on the differences in C values between same and separate pedigrees, our simulation results became statistically significant only when large numbers of gene trees

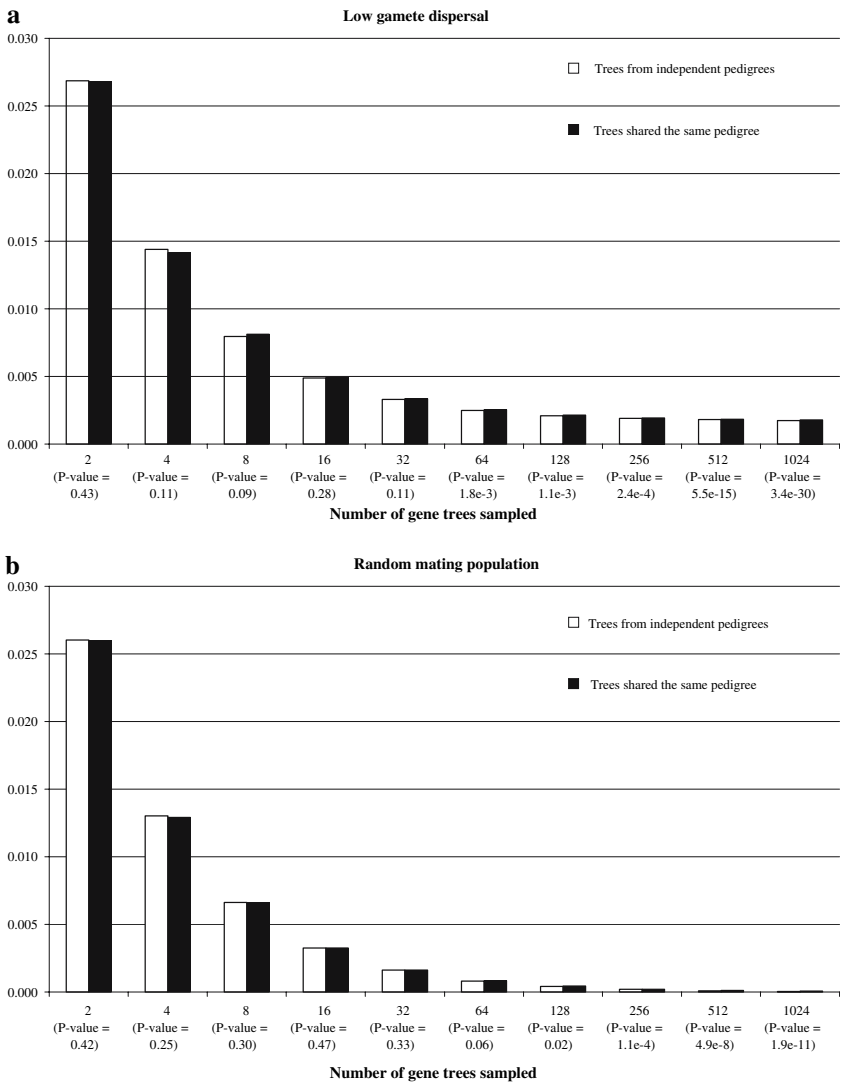


Fig. 4 Magnitudes of genealogical concordance (C) when different numbers of gene trees are sampled from the same or independent pedigrees. (**a**) low level of gamete dispersal (i.e., $\sigma_{disp} = 0.001$); (**b**) random mating population (i.e., $\sigma_{disp} = \infty$). P-values from one-tailed t-tests are listed below the X-axis. Error bars for indicating 95% confidence intervals are not shown because they are not discernible at this scale

were sampled (≥64 under low gamete dispersal and ≥128 in random mating population; see *P*-values in Fig. 4). Otherwise, the lack of statistical significance is probably attributable to stochastic noise stemming from small numbers of sampled loci.

Effects of population size

Although all of the topology-constraining effects of a shared pedigree were relatively minor, our results support the hypothesis that such effects are greater in smaller populations (Fig. 5). The topology-constraining effect of a shared pedigree is most apparent when the effective population size is lower than 128. However, it is worth emphasizing that even in extremely small populations, other factors such as population structuring may remain more prominent in their impacts on topological concordance across gene trees. By comparing results shown in Fig. 5a and 5b, one can see that the level of gamete dispersal has a far greater impact of the genealogical concordance than that of population size.

Metric of genealogical concordance

For reasons discussed in the Methods, we have focused here on the variance in the frequency distribution of gene-tree topologies (*C*) as a measure of genealogical concordance. As an alternative measure of genealogical concordance, we also examined the output of each simulation to identify the highest proportion of gene trees whose topologies agreed with one another. The results (not shown) were in all cases qualitatively similar to those that emerged from the *C* measure of concordance.

One limitation of using *C* as a measurement of genealogical concordance is that this metric treats all pairs of topologies as equally different. In other words, we will obtain the same value of *C* when two topologies are both supported by 50% of the gene trees sampled regardless of

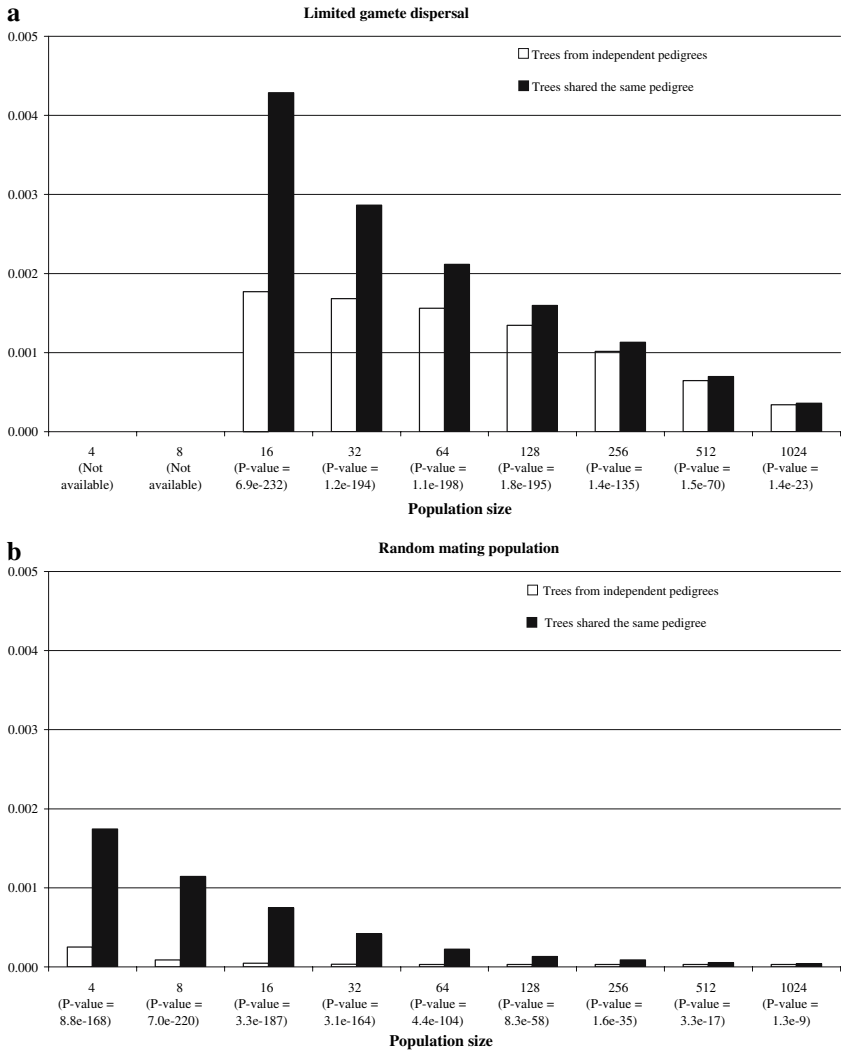


Fig. 5 Magnitudes of genealogical concordance (*C*) under different population sizes. (**a**) limited level of gamete dispersal (i.e., $\sigma_{disp} = 0.01$). The results from population size of four and eight individuals were not available because alleles sampled from different locations have a zero probability of coalescing with each other under this level of dispersal; (**b**) random mating population (i.e., $\sigma_{disp} = \infty$). *P*-values from one-tailed *t*-tests are listed below the X-axis. Error bars for indicating 95% confidence intervals are not shown because they are not discernible at this scale

similarity between these two topologies. A metric that takes topological distance among trees into account (e.g., average pairwise distance among gene trees) will be more informative in this regard. However, development of such a metric presents several challenges because the commonly used symmetric difference (Bourque 1978) does not consider branching order (such that ((1,2),(3,4)) and ((3,4),(1,2)), for example, are considered equivalent). On the other hand, distance measurements that consider branch lengths such as branch score (Kuhner and Felsenstein 1994) may introduce a high level of noise because gene trees that shared the same topology can have vastly different branch lengths. In other words, one can observe a low level of genealogical concordance even when all gene trees agree on the same topology but have different branch lengths.

Conclusions

Our results demonstrate that although organismal pedigree can constrain the topological distribution of gene trees in both structured and random mating populations even in the absence of physical linkage between genes, such effects are normally relatively minor compared to other (external) phylogeographic shaping forces (such as limited gametic dispersal). Furthermore, the number of loci that is necessary to detect such topological constraint is much higher than what is typically used in phylogeographical studies. For these reasons, the fundamental assumption of coalescent theory that physically unlinked genes have independent genealogical histories appears to hold for most practical purposes.

Acknowledgements We thank J. Kissinger, the Institute of Bioinformatics, and the Research Computing Center at University of Georgia for providing computation resources. Comments from anonymous reviewers and editor greatly improved this manuscript. C.-H.K. was supported by a National Institute of Health Training Grant (GM07103) and the Alton Graduate Research Fellowship to the Department of Genetics at University of Georgia, and J.C.A. was supported by funds from the University of California at Irvine.

References

Avise JC (2000) Phylogeography: the history and formation of species. Harvard University Press, Cambridge, MA

Avise JC, Wollenberg K (1997) Phylogenetics and the origin of species. Proc Natl Acad Sci USA 94:7748–7755

Ball RM, Neigel JE, Avise JC (1990) Gene genealogies within the organismal pedigrees of random-mating populations. Evolution 44:360–370

Bourque M (1978) Arbres de Steiner et reseaux dont certains sommets sont a localization variable. Ph.D. Dissertation, Universite de Montreal, Quebec

Degnan JH, Salter LA (2005). Gene tree distributions under the coalescent process. Evolution 59:24–37

Felsenstein J (1978) The number of evolutionary trees. Syst Zool 27:27–33

Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. Evolution 56:1557–1565

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468

Kuo CH, Avise JC (2005) Phylogeographic breaks in low-dispersal species: the emergence of concordance across gene trees. Genetica 124:179–186

Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536

Wilkins JF, Wakeley J (2002) The coalescent in a continuous, finite, linear population. Genetics 161:873–888

Wollenberg K, Avise JC (1998) Sampling properties of genealogical pathways underlying population pedigrees. Evolution 52:957–966