Received: 1 September 2016 | Revised: 13 July 2017 | Accepted: 24 July 2017

DOI: 10.1111/mec.14274

NEWS AND VIEWS

Opinion

WILEY MOLECULAR ECOLOGY

On the independent gene trees assumption in phylogenomic studies

W. Bryan Jennings

Departamento de Vertebrados, Museu Nacional, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Correspondence
W. Bryan Jennings, Departamento de Vertebrados, Museu Nacional, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.
Email: wbjenn@gmail.com

Abstract

Multilocus coalescent methods for inferring species trees or historical demographic parameters typically require the assumption that gene trees for sampled SNPs or DNA sequence loci are conditionally independent given their species tree. In practice, researchers have used different criteria to delimit "independent loci." One criterion identifies sampled loci as being independent of each other if they undergo Mendelian independent assortment (IA criterion). O'Neill et al. (2013, *Molecular Ecology*, 22, 111–129) used this approach in their phylogeographic study of North American tiger salamander species complex. In two other studies, researchers developed a pair of related methods that employ an independent genealogies criterion (IG criterion), which considers the effects of population-level recombination on correlations between the gene trees of intrachromosomal loci. Here, I explain these three methods, illustrate their use with example data, and evaluate their efficacies. I show that the IA approach is more conservative, is simpler to use and requires fewer assumptions than the IG approaches. However, IG approaches can identify much larger numbers of independent loci than the IA method, which, in turn, allows researchers to obtain more precise and accurate estimates of species trees and historical demographic parameters. A disadvantage of the IG methods is that they require an estimate of the population recombination rate. Despite their drawbacks, IA and IG approaches provide molecular ecologists with promising a priori methods for selecting SNPs or DNA sequence loci that likely meet the independence assumption in coalescent-based phylogenomic studies.

KEYWORDS
distance threshold, IGUs, independent assortment, independent genealogical units, multilocus coalescent analyses

1 | INTRODUCTION

A key assumption of multilocus coalescent methods for inferring species trees and historical demographic parameters such as population divergence times, effective population sizes and gene flow holds that gene trees of sampled SNPs or DNA sequence loci are conditionally independent of each other given the species tree (Bryant, Bouckaert, Felsenstein, Rosenberg, & RoyChoudhury, 2012; Chifman & Kubatko, 2014; Edwards, Liu, & Pearl, 2007; Hey & Nielsen, 2004; Hudson & Coyne, 2002; Liu & Pearl, 2007; Rosenberg & Nordborg, 2002; RoyChoudhury, 2011; Wakeley, 2008). Although the coalescent theory framework underlying many multilocus methods assumes that the gene trees of sampled loci are completely independent of each other, in reality, gene trees of all loci in diploid genomes are intercorrelated with each other to varying degrees due to their shared population pedigree (Wakeley, King, Low, & Ramachandran, 2012). Despite this caveat, eukaryotic genomes harbour a number of loci that are independent enough—in a statistical sense—to meet

coalescent theory expectations (King, Wakeley, & Carmi, 2017; Wakeley et al., 2012). Hudson and Coyne (2002) described such statistically independent loci as "independent genealogical units" or "IGUs," which they defined as "the number of genomic segments whose passage to monophyly is nearly independent of that for all other segments." A given pair of markers, which can be individual sites (e.g., single nucleotide polymorphisms [SNPs]) or DNA sequence loci, are expected to represent IGUs when they are found on different (i.e., nonhomologous) chromosomes or are spaced far enough apart on the same chromosome (King et al., 2017; Wakeley, 2008; Wakeley et al., 2012). This assumption is not to be confused with the "no recombination within loci" assumption that is also common to many inferences derived from multilocus data sets (see Edwards et al., 2016; Jennings, 2016; Lanier & Knowles, 2012; Springer & Gatesy, 2016). Here, we are only concerned with SNPs or nonrecombined DNA sequence loci (i.e., all sites within each locus share the same gene tree; Pluzhnikov & Donnelly, 1996).

The independent gene trees assumption is important because SNPs or loci that meet this assumption will have gene trees that effectively represent replicate samples of the evolutionary process (Arbogast et al., 2002; Chifman & Kubatko, 2014; Edwards et al., 2007; Hey & Nielsen, 2004; Liu & Pearl, 2007; Rosenberg & Nordborg, 2002; Wakeley, 2008). The property of genealogical independence of SNPs or loci confers benefits to coalescent-based analyses because larger numbers of IGUs are expected to enhance the precision and accuracy of coalescent-based parameter estimates (Arbogast et al., 2002; Edwards & Beerli, 2000; Pluzhnikov, Di Rienzo, & Hudson, 2002; Pluzhnikov & Donnelly, 1996; Rosenberg & Nordborg, 2002), an assertion that has received substantial empirical corroboration (Costa, Prosdocimi, & Jennings, 2016; Felsenstein, 2006; Jennings & Edwards, 2005; Lee & Edwards, 2008; Smith, Harvey, Faircloth, Glenn, & Brumfield, 2013). Owing to advances in next-generation sequencing and methods for obtaining genomewide data sets, researchers can now obtain unprecedented phylogenomic data sets consisting of hundreds to thousands of targeted loci (e.g., Faircloth et al., 2012; Lemmon, Emme, & Lemmon, 2012; McCormack et al., 2012; Meiklejohn et al., 2016). Moreover, as full-genome data for nonmodel organisms become increasingly available in coming years, researchers will be able to exhaustively sample all available SNPs or DNA sequence loci of interest in genomes using software (Costa et al., 2016; Jennings, 2016). Accordingly, a need exists for methods that can identify IGUs otherwise large multilocus data sets may inadvertently include nonindependent or "pseudoreplicated" samples (Costa et al., 2016). While violation of the independence assumption by a subset of SNPs or loci found in a data set may not necessarily lead to spurious species tree inferences or biased parameter estimates, the presence of nonindependent samples will likely impact the estimation of variances by making confidence intervals that are too narrow (Bryant et al., 2012; Gutenkunst, Hernandez, Williamson, & Bustamante, 2009; RoyChoudhury, 2011).

In practice, researchers have identified putative IGUs in genomes by selecting these markers from different chromosomes and using a priori distance thresholds to choose multiple intrachromosomal sites (e.g., SNPs) or DNA sequence loci. A "distance threshold" represents the minimum genetic or physical distance separating IGUs found on the same chromosome (Costa et al., 2016). While it is straightforward to choose loci from different chromosomes, knowing what particular distance threshold to use for intrachromosomal loci is more challenging. Although some early genomewide studies (e.g., Sachidanandam et al., 2001) used distance thresholds to select presumable IGUs, to my knowledge none of them presented explicit methods for obtaining such thresholds. In a previous issue of *Molecular Ecology*, O'Neill et al. (2013) addressed this need by presenting an objective methodology for choosing IGUs and illustrating its use in an empirical study. More recently, Costa et al. (2016) developed an alternative approach aimed at accomplishing the same task. Before all of these studies, Pluzhnikov and Donnelly (1996) also provided a theoretical framework for delimiting IGUs but I am unaware of any attempts to further develop and apply their idea in any empirical phylogenomic studies. Below, I review the O'Neill et al., Costa et al. and Pluzhnikov & Donnelly approaches (hereafter OEA, CEA and P&D, respectively), illustrate their use with example data and evaluate their advantages and disadvantages.

2 | O'NEILL ET AL. APPROACH TO DELIMITING IGUS

OEA inferred the population structure of the North American tiger salamander species complex (*Ambystoma tigrinum*) using 95 presumably independent DNA sequence loci. Moreover, they selected their loci from different chromosomes and, for intrachromosomal loci, they chose loci separated from each other by at least 50 centimorgans (cM). The authors did not state the significance of the "50 cM" distance, but presumably they chose this criterion for delimiting independent intrachromosomal loci because such loci would be expected to undergo independent assortment (Hartl & Jones, 2006). It should be emphasized that the IGU assumption in phylogenomics is not equivalent to the independent loci assumption of Mendelian genetics. Indeed, for intrachromosomal loci, the IGU assumption concerns the *conditional independence of gene trees* of those loci given their demographic history, which is achieved via multigenerational effects of population-level recombination. In contrast, for intrachromosomal loci under the Mendelian assumption, the *independence of loci* is due to single-generation effects of recombination in individuals (i.e., no demographic component). Nonetheless, it may still be acceptable practice to use an independent assortment criterion (hereafter "IA" criterion) to identify IGUs, as Felsenstein (2004) noted that loci separated by genetic distances less than 30 cM are expected to have gene trees that are completely different.

OEA used their method with a genetic linkage map from a representative species of tiger salamander to obtain a sample of 95 loci. They did not estimate the maximum number of IGUs in the tiger salamander genome using the IA criterion; rather, they filtered their candidate loci to find only enough loci to populate a 96-sample PCR

plate. However, given the map length of 5,251 cM for the tiger salamander (Smith, Kump, Walker, Parichy, & Voss, 2005), we can use this information to estimate the maximum number of IGUs that could be harvested from this genome under the IA criterion. The resulting number is 5,251 cM/50 cM = 105 loci, which is similar to the number obtained by OEA. With the total IGU number, we can estimate physical distance thresholds between loci provided genome size is known. The tiger salamander genome is 3.2×10^{10} base pairs (bp) or 32 gigabases (Gb) in size (Keinath, Timoshevskly, Tsonis, Voss, & Smith, 2015), which suggests that IGUs are separated from each other by 3.2×10^{10} bp/105 = 304,761,905 bp or 305 megabases (Mb), on average.

The 32-Gb tiger salamander genome is roughly an order of magnitude larger than the genomes of many other vertebrates. Thus, how many IGUs can be found in a more typical vertebrate genome? We can begin developing an understanding of IGU numbers in other vertebrates by examining Australian estrildid finches, a group which has been the subject of multilocus phylogeographic and whole-genome studies. One of these phylogeographic studies (Jennings & Edwards, 2005) focused on three closely related species of grass finches (*Poephila acuticauda*, *P. hecki and P. cincta*), while another (Balakrishnan & Edwards, 2009) concerned the closely related zebra finches (*Taeniopygia guttata*). Because both the complete genome sequence and linkage map are available for the zebra finch, we can use this information to estimate the number of IGUs and physical distance thresholds for this species and the grass finches. Zebra finches have a genome size of 1.2×10^9 bp or 1.2 Gb (Warren et al., 2010) and a map length of ~2,400 cM (Backström et al., 2010). Using this map with the OEA approach, the number of IGUs in the genomes of grass finches and zebra finches is estimated to be only 48 loci while the physical distance separating IGUs is around 25 Mb, on average.

3 | COSTA ET AL. APPROACH TO DELIMITING IGUS

In a full-genome phylogenomic study of the hominoids, CEA employed a 200 kilobase (kb) threshold to select 292 anonymous loci for coalescent-based analyses. Their method for estimating physical distance thresholds is based on the original work of Hudson and Coyne (2002). In contrast to the IA-based approach, this method employs an independent genealogies criterion (hereafter "IG" criterion). This approach consists of first estimating the total number of IGUs per genome followed by conversion of this estimate into a physical distance threshold. A complete explanation of this methodology, which has not been provided before, is given below.

Hudson and Coyne (2002), hereafter H&C, derived a formula for estimating the total IGUs per genome. First, these authors noted that under a neutral evolution model the statistical dependence between a pair of intrachromosomal loci (or sites) depends on $4N_ec$, where N_e is effective population size and c represents the recombination rate between those loci or sites per generation. Before we continue with the IGU formula derivation, we must clarify a couple potentially confusing issues. First, many previous studies have denoted the per generation recombination rate as r instead of c (e.g., Frisse et al., 2001; Hartl & Clark, 1997). Here, we will use c to represent this parameter following H&C. Second, it is important to notice how this parameter is defined in studies. For example, c (and r) have been defined as *per site* (i.e., between adjacent nucleotides; e.g., Kaplan, Hudson, & Langley, 1989; Wall, 2003) and *per locus* (e.g., Hudson, 1987; Wall, 2000) per generation recombination rates. To estimate the number of IGUs in the *Drosophila melanogaster* genome, H&C used a genetic linkage map for this species to estimate the number of recombinations (or crossovers) *per genome* per generation. Therefore, they treated "c" as a per genome per generation recombination rate in their formula. Thus, to avoid confusion, we will define c as the per site per generation recombination parameter and c_G as the per genome per generation recombination parameter. Furthermore, we assume that c scales linearly with physical distance in the genome (i.e., $c = c_G/G$, where G is genome size in bp) and that there is no recombination rate heterogeneity in the genome (i.e., single recombination rate). The quantity $4N_ec_G$, which represents the population recombination rate for the entire genome, can be viewed as the expected number of recombinations in the ancestors of two random gene copies going back to their most recent common ancestor. In other words, it is equal to the average expected coalescence time for each copy ($2N_e$ generations $\times$ 2) times c_G recombination events per generation. In the next step of their derivation, H&C pointed out that r^2, a measure of linkage disequilibrium, is ~$1/4N_ec$ for large populations, a finding they attributed to Ohta and Kimura (1971). They further stated that if r^2 is .001, then a pair of intrachromosomal loci could be considered IGUs because of the low correlation. Taking this into account, H&C obtained the result $4N_ec_G = 1/0.001 = 1{,}000$. Because this particular value of $4N_ec_G$ represents the assumed minimum amount of population-level recombination required to statistically decouple the genealogies of a pair of genomic loci, dividing 1,000 into $4N_ec_G$ gives the desired result—the approximate number of IGUs per genome. Thus, H&C's IGU formula is:

$$I = \frac{4N_ec_G}{1{,}000}, \tag{1}$$

where I represents the total IGUs per genome (Hudson & Coyne, 2002; Costa et al., 2016). Note that the c parameter used in CEA's version of this formula is also a per genome per generation recombination rate and therefore it is equivalent to c_G here. As can be seen in (1), N_e plays a key role in determining the number of loci with independent genealogies: for a given recombination rate, large N_e values translate to more IGUs per genome than smaller N_e values and vice versa. To estimate a global (within a genome) physical distance threshold, CEA first estimated I using (1), then used the formula:

$$D_T = G/I, \tag{2}$$

where D_T is the average physical distance in bp between IGUs. We can now use formulae (1) and (2) to obtain IG-based estimates of IGUs and physical distance thresholds for tiger salamanders, grass finches and zebra finches.

As H&C demonstrated, a linkage map can be used to obtain an estimate for the recombination parameter c_G in (1). However, an estimate of N_e must also be supplied, which is problematic in this case because North American tiger salamanders have varying N_e depending on species. For example, Wang, Johnson, Johnson, and Shaffer (2011) found that California tiger salamanders (*A. californiense*) had low N_e of 11–64, which may be explained by population bottlenecks or pond sizes. In contrast, Church, Kraus, Mitchell, Church, and Taylor (2003), who used mitochondrial DNA, estimated the effective number of females (N_f) in Eastern tiger salamanders (*A. tigrinum*) to be 134,000–144,000. Because autosomal loci have fourfold higher N_e than mitochondrial loci (Wilson et al., 1985), N_e for autosomal loci in these salamanders is likely higher. Given the pronounced effect of N_e on IGU and physical distance threshold estimates, this means that substantially different estimates of I (and hence D_T) can be obtained from formulae (1) and (2) depending on which N_e value is inserted into (1). Faced with this dilemma, which N_e should we use? Because OEA conducted a phylogeographic study of this entire species complex, we would thus be justified in using a value of N_e in the range of 10^5 to 5×10^5. However, to illustrate the CEA method, we will choose $N_e = 10^5$, which represents a conservative value from the low end of this range. Applying the IG criterion via (1) to estimate the number of IGUs in the tiger salamander genome, we first observe that $c_G = 5{,}251\ \text{cM} \times 0.01\ \text{rec cM}^{-1}\ \text{gen}^{-1} = 53\ \text{rec gen}^{-1}$, where rec and gen represent units of recombinations and generations, respectively. Next, assuming $N_e = 10^5$, we obtain:

$$I = [(4 \times 10^5\ \text{gen}) \times 53\ \text{rec gen}^{-1}]/1{,}000\ \text{rec}$$
$$I = 21{,}200$$

Once an IGU estimate is in hand, formula (2) is ready to be used to obtain a corresponding physical distance threshold if genome size for the study organism is known. In this case, the tiger salamander genome is 3.2×10^{10} bp (Keinath et al., 2015), which yields a physical threshold distance of 3.2×10^{10} bp/21,200 = 1,509,434 bp (~1.5 Mb).

Jennings (2016) provided estimates of I and D_T for grass and zebra finches using the CEA approach, but those results were based on initial genetic linkage maps that did not cover the entire genome. Backström et al. (2010) suggested that the complete map length of the zebra finch genome is closer to 2,400 cM rather than to previous estimates of 1,068 and 1,341 cM. We will therefore re-estimate I and D_T for these birds using an adjusted map length estimate of 2,400 cM. Grass finches and zebra finches appear to have substantially different effective population sizes. Although species-specific N_e estimates are not available for the three grass finch species, Jennings and Edwards (2005) did estimate the sizes of their most recent common ancestral populations (N_a), which are 384,000 for the ancestor of *P. acuticauda* and *P. hecki* (these were formerly considered to be a single species) and 521,000 for the ancestor of *P. acuticauda*, *P. hecki* and *P. cincta*. In contrast, N_e for zebra finches found on the Australian mainland (they also occur in Indonesia and East Timor) has been estimated to range between 1.3 and 7 million (Balakrishnan & Edwards, 2009). This disparity in effective population sizes between grass finches and zebra finches is not surprising when considering that the former group is restricted to the monsoon tropics of northern Australia, while the latter are widely distributed across Australia. Thus, to obtain conservative estimates of I and D_T for these finches, we will use the lower-bound values of effective population size for each (i.e., 384,000 for grass finches and 1.3×10^6 for zebra finches). Using this information together with the adjusted map length in formula (1), we obtain estimates of 36,864 and 124,800 IGUs for grass and zebra finches, respectively. To obtain estimates of D_T, we only need to use our estimates of I with the known genome size for zebra finches (i.e., 1.2×10^9 bp, Warren et al., 2010) in formula (2), which gives us 32,552 bp (~33 kb) and 9,615 bp (~10 kb) for grass and zebra finches, respectively.

H&C's IGU formula requires a linkage map for estimating the recombination parameter c_G, a requirement that makes the CEA approach challenging to use with nonmodel organisms owing to the difficulties of constructing linkage maps. However, other methods for estimating recombination rates exist, which do not involve linkage maps. This other class of methods relies on population genetics software to directly estimate recombination rates from genomic sequences. These programs can estimate either per site (or per locus) per generation recombination rates (e.g., LAMARC 2.0, Kuhner, 2006) or per site (or per locus) population recombination rates (e.g., SITES, Hey & Wakeley, 1997; INFERRHO, Wang & Rannala, 2008, 2009). However, to integrate these recombination estimation methods into the CEA approach, we must first modify formulae (1) and (2) to accommodate additional recombination parameters.

The quantity $4N_ec_G$ represents the population recombination rate scaled to the genome. However, when using the bioinformatics approach to estimate population-level recombination from DNA sequences, researchers instead are usually more interested in estimating the population recombination rate scaled to bp (i.e., population recombination rate between adjacent nucleotides). This *per site* population recombination rate, which is denoted by the parameter ρ, is described by the formula $\rho = 4N_ec$ where c is the recombination rate per site per generation (e.g., Becquet & Przeworski, 2007). Given that $c = c_G/G$, the quantities ρ, $4N_ec_G$, and $4N_ec$ can be related to each other in the following manner:

$$G\rho = 4N_ec_G = 4N_ecG. \quad (3)$$

We can now use formulae (2) and (3) to construct an IGU formula that contains all parameters of interest. By rearranging (2) and substituting the terms in (3) for $4N_ec_G$ in (1), we can obtain:

$$I = G/D_T = G\rho/1{,}000 = 4N_ec_G/1{,}000 = 4N_ecG/1{,}000. \quad (4)$$

Although the G/D_T term in (4) may not be useful for estimating I, its presence in this formula enables us to construct an equation for directly estimating D_T. Thus, if we solve for D_T in (4), we get:

$$D_T = G/I = 1{,}000/\rho = 1{,}000G/4N_ec_G = 1{,}000/4N_ec. \quad (5)$$

Formulae (4) and (5) are more versatile than (1) and (2) because they offer researchers different options for estimating I and D_T depending on which recombination parameter is to be estimated. For example, a linkage map can provide an estimate for c_G while population genetics software can be used to estimate c or ρ. Notice

that if an estimate of ρ is used in (4) and (5), then an estimate of N_e is not needed. Also, as (4) and (5) assume that ρ is a per site estimate, be aware of the fact that some software programs instead generate an estimate of the *per locus* ρ. Thus, a per locus ρ estimate must first be converted into a per site ρ (i.e., $\rho = \rho_{locus}/L$, where L is the locus length in bp) before it can be used in these IGU and physical distance threshold formulae. This also holds for the per generation recombination parameter *c* (i.e., $c = c_{locus}/L$).

4 | PLUZHNIKOV & DONNELLY APPROACH TO DELIMITING IGUS

In a study based on computer simulations, P&D investigated the relationships between physical distances and per site population recombination rates (ρ) on correlations between gene genealogies. In their study, these authors made the assumption that population recombination rates scale linearly in the genome with physical distance. Therefore, they defined the term Dρ, which represents the scaled population recombination rate, where D is the physical distance in bp between sites or loci (Pluzhnikov & Donnelly, 1996). Their results suggested that the gene trees of two intrachromosomal loci could be considered uncorrelated with each other when $D\rho \geq 10$. Notice that this formula can be used to estimate physical distance thresholds given an estimate of ρ (i.e., $D_T = 10/\rho$). Also, if we re-examine formula (4), we see that $G\rho/1{,}000 = G/D_T$ can be rearranged and simplified to yield $D_T\rho = 1{,}000$, which resembles P&D's $D\rho \geq 10$ expression except that in the former the r^2 cut-off for IGUs is ~.001, whereas it is ~.1 in the latter. Although the P&D IG criterion is far less stringent than the H&C criterion (by two orders of magnitude), the fact that both studies used a common mathematical framework to identify IGUs means we can construct new, albeit less conservative, IGU and physical distance threshold formulae similar to (4) and (5), but which are based on the P&D criterion for IGUs:

$$I = G/D_T = G\rho/10 = 4N_e c_G/10 = 4N_e cG/10 \quad (6)$$

and,

$$D_T = G/I = 10/\rho = 10G/4N_e c_G = 10/4N_e c. \quad (7)$$

With these new formulae, we can estimate *I* and D_T in tiger salamanders and finches using the IG criterion of P&D. Using (6) and the available information for tiger salamanders, *I* is estimated to be an astonishing 2.1 million while formula (7) yields a D_T of ~15 kb. The P&D approach yields IGU estimates of 3,686,400 and 12.5 million for grass and zebra finches, respectively, and D_T of ~0.3 kb and ~0.1 kb for each, respectively.

5 | ADVANTAGES AND DISADVANTAGES OF EACH APPROACH

An evaluation of the three methods for delimiting IGUs using the tiger salamander and finch examples shows that the IA-based method is, as expected, quite conservative compared to the two IG-based methods (Table 1). Indeed, for each example species or species group, the IA-based approach only yielded ~100 or fewer IGUs per genome, whereas IG-based methods produced IGU estimates in the tens of thousands to millions (Table 1). However, an advantage of the IA approach is that it is simple to use provided a genetic linkage map is available. The IG methods are more complicated to use because they require an estimate of the population recombination rate. Accordingly, to use an IG-based method a researcher must obtain estimates of N_e and *c* or an estimate of ρ. Another problem evident with the IG-based methods is that they yield dramatically different results with the CEA approach being far more conservative than the P&D approach (Table 1). These and other issues particular to the IG methods are further considered below.

As we observed earlier, a number of different N_e (or N_a) estimates may exist for each study species or species group, a circumstance that complicates the use of IG-based methods. Given this situation, how should a researcher choose an N_e or N_a value for use in IG formulae? We addressed this issue here by selecting the smallest reasonable estimate of N_e (or N_a), a strategy that appeared to perform well as demonstrated by the extremely large IGU estimates obtained for tiger salamanders, grass finches and zebra finches (Table 1). Thus, for studies with multiple available N_e or N_a values, perhaps the researcher should adopt this conservative strategy for estimating *I* and D_T until future studies provide alternative guidelines.

Given that the P&D approach yields a far larger number of presumable IGUs, should this method be preferred over the more conservative CEA method? We begin addressing this question by taking a closer look at the distance thresholds obtained from each IG method, specifically with a focus on the required minimum correlation level between genealogies to delimit IGUs. In their simulation results, P&D observed that the correlation in estimates of coalescent-based parameters was "low" when $D\rho \geq 10$ (i.e., $r^2 \sim .1$). While

TABLE 1 Estimates of the number of independent genealogical units (*I*) and physical distance thresholds (D_T) in the genomes of tiger salamanders, grass finches and zebra finches using three different approaches. OEA = O'Neill et al. (2013), CEA = Costa et al. (2016), P&D = Pluzhnikov and Donnelly (1996), IA = independent assortment criterion, IG = independent genealogies criterion, kb = kilobase and Mb = megabase

Approach (criterion)	Tiger salamander *I*	Grass finch *I*	Zebra finch *I*	Tiger salamander D_T	Grass finch D_T	Zebra finch D_T
OEA (IA)	105	48	48	305 Mb	25 Mb	25 Mb
CEA (IG)	21,200	36,864	124,800	1.5 Mb	33 kb	10 kb
P&D (IG)	2,120,000	3,686,400	12,480,000	15 kb	0.3 kb	0.1 kb

the correlations obtained by P&D varied somewhat depending on simulation scenario, we can estimate the predicted correlation in coalescence times between pairs of intrachromosomal loci under both the P&D and CEA criteria using formula 7.28 in Wakeley (2008). Doing this for the tiger salamander and finches reveals that the predicted correlation for loci chosen via the P&D approach is approximately 0.1, while it is only 0.001 for loci obtained using the CEA approach (Figure 1). In the case of tiger salamanders, the curve shown in Figure 1a suggests that the correlation in coalescence times essentially becomes zero at ~200 kb. Thus, the ~1.5 Mb distance threshold estimated by the CEA approach appears to be more than adequate—if not too conservative. For grass and zebra finches, the correlation closely approaches zero at ~1.5 and ~1 kb, respectively—both well below their own CEA-based thresholds of ~30 and ~10 kb (Figure 1b; Table 1).

A potentially serious weakness of both IG-based methods is that they assume a single genomewide recombination rate. This is a concern because the existence of recombination coldspots and hotspots in eukaryotic genomes may negatively affect the estimation of IGUs and physical distance thresholds based on a single-rate model. Indeed, this does seem to be the case, as predicted correlations in coalescence times obtained under a single recombination rate model have been found to greatly underestimate correlations based on a mixed-rate recombination model and real data (Wakeley, 2008). It is therefore essential for us to determine how well these current IG-based methods perform in organisms with genomes having spatially variable recombination landscapes.

We can begin to examine this issue by making use of the results illustrated in figure 7.9 of Wakeley (2008). This figure, which is not shown here, shows two downward trending curves depicting the predicted correlations in coalescence times between intrachromosomal loci in the human genome vs. physical distance based on single-rate (i.e., similar to Figure 1 in this study) and mixed-rate recombination models. The single- and mixed-rate curves closely approach the zero correlation threshold by ~200 kb and ~1 Mb, respectively. As Wakeley pointed out, correlations based on real data, which are also shown in his figure 7.9, agreed with the mixed-model quite well, thereby suggesting that the "legitimate" threshold distance for delimiting IGUs in humans is ~1 Mb. Given $\rho = 5.2 \times 10^{-4}$ for the human genome (Reich et al., 2002; Wakeley, 2008), we can use formulae (5) and (7) to estimate D_T in the human genome using each IG-based method and then compare these values to the legitimate threshold value. The CEA method yielded a D_T ~2 Mb, which is twice the legitimate threshold distance. In contrast, D_T for the P&D method was estimated to be only ~19 kb, which, according to Wakeley's figure 7.9, suggests that the effective correlation in coalescence times for these loci under this criterion is only around 0.5 rather than the 0.1 correlation expected under the single-rate model.

H&C evidently chose such a conservative r^2 of .001 as the statistical cut-off value for their IGU formula to preclude the possibility that recombination coldspots in a typical genome could adversely impact the estimation of IGUs. Therefore, it is not surprising that the distance threshold estimate generated via the CEA method was shown to be more than sufficient for delimiting IGUs in the human genome despite the existence of spatially variable recombination rates. In contrast, the P&D method, which was based on a far less conservative cut-off value, appeared to not be effective at establishing a valid threshold for IGUs in the human genome.

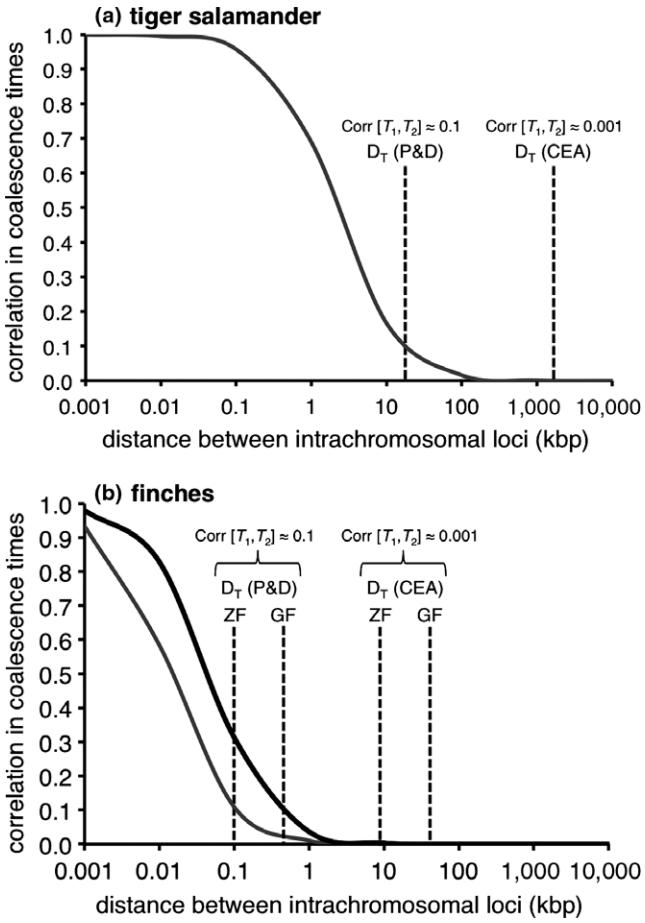

FIGURE 1 Predicted correlations in coalescence times (T_1, T_2) for pairs of intrachromosomal loci in relation to physical distances between loci. (a) distance vs. correlation in coalescence times for the tiger salamander. (b) distance vs. correlation in coalescence times for grass finches (black curve) and zebra finches (grey curve). Each curve was generated using the formula Corr $[T_1, T_2] \approx \rho + 18/\rho 2 + 13\rho + 18$ (from formula 7.28 in Wakeley, 2008) and assuming a per site population recombination rate of $\rho = 6.63 \times 10^{-4}$ for tiger salamanders, $\rho = 3.07 \times 10^{-2}$ for grass finches, and $\rho = 1.04 \times 10^{-1}$ for zebra finches (ρ estimates were obtained using data in the main text). A single recombination rate is assumed to scale linearly with distance; that is, in the formula above, $D\rho$ replaces ρ, where D represents the physical distances in base pairs (bp) between loci. For example, a distance of 100 bp in the tiger salamander genome yields a scaled recombination rate between the two loci of 6.63×10^{-2} and a correlation $\approx$ .96. Vertical dashed lines indicate physical distance thresholds (D_T) along the horizontal axis obtained using the Pluzhnikov & Donnelly, 1996 (P&D) and Costa et al., 2016 (CEA) approaches. GF = grass finch, ZF = zebra finch, and kbp = kilobase pair. Drawn after figure 7.9 in Wakeley (2008)

Recombination coldspots and hotspots in eukaryotic genomes could adversely affect the results obtained from IG-based methods in another way. As pointed out earlier, a researcher can estimate recombination parameters (i.e., c or ρ) from genomic sequences for use in IG formulae. However, because these IG methods assume a single recombination rate across the genome, it does not seem advisable to obtain a single estimate (per site or per locus) of one of the recombination parameters that is based only on one small genomic region. If only a single point estimate were to be used in IG formulae, then it is possible that resulting IGU and physical distance threshold estimates could be at the extreme low or high end of the range of possible values depending on whether the estimate was obtained from a coldspot or hotspot region. Instead, a better strategy may be to obtain a number of estimates from random genomic locations and then use the corresponding average in IG formulae.

6 | NUMBERS OF IGUS VS. THE PRECISION AND ACCURACY OF COALESCENT-BASED INFERENCES

The IA method is less complicated and has fewer assumptions than IG methods, but the fact that it only estimated a maximum of ~100 IGUs in a genome as large as the tiger salamander also shows that this method is extremely conservative. In contrast, IG-based methods evidently have the potential to delimit thousands to millions of independent SNPs or DNA sequence loci per genome (Table 1). But would such enormous data sets likely generate estimates of species trees and historical demographic parameters that are better quality than inferences based on the largest current data sets? From a statistical standpoint, this is expected to be the case because increasing the number of replicate samples should lead to concomitant decreases in the variances of estimated parameters. Indeed, this theoretical expectation is being corroborated by empirical studies because regardless of whether a comparison involved ~5–15 vs. ~16–30 loci (Jennings & Edwards, 2005; Lee & Edwards, 2008), 53 vs. 292 loci (Costa et al., 2016), or 166 vs. 776–1,516 loci (Smith et al., 2013), the anticipated decreases in parameter variances for the larger data sets were observed. Moreover, the variances associated with the larger data sets were frequently twofold to threefold narrower compared to estimates based on the smaller data sets. It is therefore reasonable to expect this trend to continue further as loci numbers per data set reach into the tens of thousands and beyond until either the supply of loci in a given genome is exhausted or the variance cannot be further reduced (King et al., 2017). In addition to gains in statistical power, larger numbers of loci can also improve the accuracy of parameter estimates. For example, Costa et al. (2016) found that the posterior means of historical demographic parameters based on a 53-locus data set suffered problems with being strongly influenced by the priors, whereas their larger 292-locus data set was immune to this problem. Thus, larger numbers of IGUs may also help overcome the problem of poorly specified priors in Bayesian-based software programs, thereby providing more accurate inferences.

7 | IGU DELIMITATION IN SHALLOW-SCALE VS. DEEP-SCALE PHYLOGENOMIC STUDIES

The IGU delimitation methods reviewed here are expected to be especially useful for phylogeographic and shallow-scale phylogenomic studies for at least a couple of reasons. First, researchers are increasingly using multilocus coalescent methods to estimate species trees and historical demographic parameters in these types of studies. Second, because it is reasonable to expect little variation in chromosome numbers, recombination rates and genome sizes among individuals sampled from recently diverged populations (or species) compared to samples obtained from more highly diverged species, it should be less complicated to apply IG-based methods to the former type of study than to the latter. However, as researchers are also increasingly using multilocus coalescent methods to infer deep-scale species trees (e.g., Edwards et al., 2016; McCormack et al., 2012), how should a researcher use an IG-based method when faced with significant variation in chromosome numbers, recombination rates or genome sizes among sampled individuals? One strategy is to take a conservative approach by choosing the genome from among the sampled genomes that yields the lowest IGU counts and highest physical distance thresholds. Provided that N_e is not too small ($<10^4$), it seems likely that the researcher can still potentially delimit hundreds or thousands of IGUs for use in a study. Another strategy would be to first obtain a range of IGU and physical distance threshold estimates from various genomes in a study, which are then used to construct different data sets with varying numbers of loci—and possibly, with varying numbers of nonindependent loci. Each data set can then be separately analysed using multilocus coalescent analyses to assess variation in estimates of species trees and historical demographic parameters. This sensitivity analysis is worthwhile to do because it is possible that such an exercise will not produce results that vary in a biologically meaningful way. At worst, if results do vary in an important way depending on which data set is analysed, then at least the phylogenomics community will realize that this is a more serious problem that must be further studied.

8 | PHYSICAL MAPPING OF IGUS AND THEIR REPORTING IN PHYLOGENOMIC STUDIES

With the availability of complete genome sequences that match or are similar to the study group of interest, a researcher can design phylogenomic loci from any part of the genome and, potentially, very large numbers of them. For example, McCormack et al. (2012) used complete genome sequences of the chicken, *Anolis* lizard, and zebra finch to design hundreds of UCE loci for use in their phylogenomic

study of placental mammals. These authors took advantage of their complete genome data by calculating the physical distances between pairs of intrachromosomal loci. Because their selected loci were generally at least 2 Mb apart, they suggested their loci likely met the independent gene trees assumption. Although justification for this assertion was not given, their use of complete genome sequences to design loci with known interlocus physical distances nonetheless represented an advance in phylogenomics.

The use of in silico-based methods to identify IGUs is expected to increase as complete genome sequences become increasingly easier to acquire for nonmodel organisms (Costa et al., 2016; Jennings, 2016). Regardless of which IGU delimitation method is chosen, researchers should report (perhaps in supplementary tables) all interlocus physical distances or chromosomal coordinates for each locus as some recent studies have done (e.g., Costa et al., 2016; Lemmon et al., 2012). If an IG-based method is employed, then researchers should also show all parameter values used to estimate IGUs and physical distance thresholds as well as provide a table containing $D\rho$ estimates for all pairs of intrachromosomal loci. By providing this information, others can gauge the degree of genealogical independence for each pair of sampled loci in the light of different IGU criteria (e.g., $D\rho = 10$ for P&D, $D\rho = 1{,}000$ for CEA, etc.). Thus, $D\rho$ may prove useful as a test statistic for determining whether observed recombinational distances are "statistically significant" or not relative to each criterion (Pluzhnikov & Donnelly, 1996).

9 | DISTANCE THRESHOLDS AND THE DELIMITATION OF NEUTRAL IGUS

The identification of IGUs that may satisfy the neutrality assumption of multilocus coalescent analyses represents another application of IG-based distance thresholds (Costa et al., 2016; Jennings, 2016). These *neutral* IGUs are nonfunctional sites or loci (i.e., not targets of selection) that are genealogically independent of other sampled IGUs and functional genomic elements. Although the genomes of many eukaryotes apparently contain vast tracts of nonfunctional DNA, much of this DNA may still not meet the neutrality assumption owing to the influences of indirect selection such as hitchhiking, background selection and balancing selection (Costa et al., 2016; Jennings, 2016). To delimit presumably neutral IGUs, one must find candidate sites or loci in nonfunctional parts of the genome and ensure that they are located far enough from sites under selection for their respective gene trees to be independent of each other. Thus, physical distance thresholds can be used to select neutral IGUs.

A number of studies have used physical distance thresholds to help delineate presumably neutral loci (e.g., Burgess & Yang, 2008; Chen & Li, 2001; Peng, Elango, Wildman, & Soojin, 2009). However, to my knowledge none of these studies employed an objective IG-based criterion for estimating these distances. If one has access to an appropriate annotated genome sequence, which includes a General Features Format (GFF) file showing the locations of all annotated elements (e.g., genes and regulatory elements), then the same IG-based physical distance threshold used to delimit IGUs (D_T) can also be used to find IGUs that are likely neutral. In their phylogenomic study of hominoids, CEA used this approach to delimit hundreds of anonymous loci that are expected to have gene trees largely unperturbed by the effects of selection acting elsewhere in the genome.

10 | CONCLUSIONS

Of the three methods reviewed here for delimiting IGUs, the IA approach was by far the most conservative. Although the IA method only identified 105 and 48 IGUs in the tiger salamander and finch genomes, respectively, vs. the thousands to millions found using IG-based methods, it has the advantage of its simplicity assuming a linkage map is available. In contrast, the two IG-based methods are more complicated to use because they require an estimate of the population recombination rate. This disadvantage of the IG-based methods may be compensated, however, by the benefit of bringing much larger numbers of SNPs or DNA sequence loci to coalescent-based analyses, which is expected to allow researchers to obtain more precise and accurate estimates of species trees and historical demographic parameters. Of the two IG-based methods, the CEA method is likely to perform best owing to its robustness to violation of the single-rate recombination assumption. IG-based methods can also be used to estimate appropriate physical distance thresholds for identifying IGUs that may meet the neutrality assumption of coalescent-based analyses. As researchers increasingly gain access to complete genome data in coming years, the importance of methods for delimiting IGUs will grow as well. Methods such as these will help researchers acquire the maximum available IGUs and thus allow coalescent-based phylogenomic studies to reach their full potential.

ACKNOWLEDGEMENTS

I am grateful to Senior Editor Nolan Kane, Jeet Sukumaran, Joseph Felsenstein, Alex Pyron and five anonymous reviewers for providing many helpful comments, which greatly improved this study.

AUTHOR CONTRIBUTIONS

W.B.J. conceived and wrote the article.

REFERENCES

Arbogast, B. S., Edwards, S. V., Wakeley, J., Beerli, P., & Slowinski, J. B. (2002). Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics*, *1*, 707–740.

Backström, N., Forstmeier, W., Schielzeth, H., Mellenius, H., Nam, K., Bolund, E., … Ellegren, H. (2010). The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research*, *20*, 485–495.

Balakrishnan, C. N., & Edwards, S. V. (2009). Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (*Taeniopygia guttata*). *Genetics*, *181*, 645–660.

Becquet, C., & Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Research*, *17*, 1505–1519.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., & RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: Bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, *29*, 1917–1932.

Burgess, R., & Yang, Z. (2008). Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Molecular Biology and Evolution*, *25*, 1979–1994.

Chen, F. C., & Li, W.-H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *American Journal of Human Genetics*, *68*, 444–456.

Chifman, J., & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, *30*, 3317–3324.

Church, S. A., Kraus, J. M., Mitchell, J. C., Church, D. R., & Taylor, D. R. (2003). Evidence for multiple Pleistocene refugia in the postglacial expansion of the eastern tiger salamander, *Ambystoma tigrinum tigrinum*. *Evolution*, *57*, 372–383.

Costa, I. R., Prosdocimi, F., & Jennings, W. B. (2016). In silico phylogenomics using complete genomes: A case study on the evolution of hominoids. *Genome Research*, *26*, 1257–1267.

Edwards, S. V., & Beerli, P. (2000). Perspective: Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution*, *54*, 1839–1854.

Edwards, S. V., Liu, L., & Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 5936–5941.

Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., … Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, *94*, 447–462.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*, 717–726.

Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland: Sinauer Associates.

Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, *23*, 691–700.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., & Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *American Journal of Human Genetics*, *69*, 831–843.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*, e1000695.

Hartl, D. L., & Clark, A. G. (1997). *Principles of population genetics* (Vol. 116). Sunderland: Sinauer Associates.

Hartl, D. L., & Jones, E. W. (2006). *Essential genetics: A genomics perspective*. Sudbury: Jones & Bartlett Publishers.

Hey, J., & Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, *167*, 747–760.

Hey, J., & Wakeley, J. (1997). A coalescent estimator of the population recombination rate. *Genetics*, *145*, 833–846.

Hudson, R. R. (1987). Estimating the recombination parameter of a finite population model without selection. *Genetics Research*, *50*, 245–250.

Hudson, R. R., & Coyne, J. A. (2002). Mathematical consequences of the genealogical species concept. *Evolution*, *56*, 1557–1565.

Jennings, W. B. (2016). *Phylogenomic data acquisition: Principles and practice*. Boca Raton: CRC Press/Taylor & Francis.

Jennings, W. B., & Edwards, S. V. (2005). Speciational history of Australian Grass Finches (*Poephila*) inferred from thirty gene trees. *Evolution*, *59*, 2033–2047.

Kaplan, N. L., Hudson, R. R., & Langley, C. H. (1989). The "hitchhiking effect" revisited. *Genetics*, *123*, 887–899.

Keinath, M. C., Timoshevskly, N. Y., Tsonis, P. A., Voss, P. A., & Smith, J. J. (2015). Initial characterization of the large genome of the salamander *Ambystoma mexicanum* using shotgun and laser capture chromosome sequencing. *Scientific Reports*, *5*. https://doi.org/10.1038/srep16413

King, L., Wakeley, J., & Carmi, S. (2017). A non-zero variance of Tajima's estimator for two sequences even for infinitely many unlinked loci. *Theoretical Population Biology*, [Epub ahead of print] https://doi.org/10.1016/j.tpb.2017.03.002

Kuhner, M. K. (2006). LAMARC 2.0: Maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics*, *22*, 768–770.

Lanier, H. C., & Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? *Systematic Biology*, *61*, 691–701.

Lee, J. Y., & Edwards, S. V. (2008). Divergence across Australia's Carpentarian barrier: Statistical phylogeography of the Red-backed Fairy Wren (*Malurus melanocephalus*). *Evolution*, *62*, 3117–3134.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, *61*, 727–744.

Liu, L., & Pearl, D. K. (2007). Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Systematic Biology*, *56*, 504–514.

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, *22*, 746–754.

Meiklejohn, K. A., Faircloth, B. C., Glenn, T. C., Kimball, R. T., & Braun, E. L. (2016). Analysis of a rapid evolutionary radiation using ultraconserved elements: Evidence for a bias in some multispecies coalescent methods. *Systematic Biology*, *65*, 612–627.

Ohta, T., & Kimura, M. (1971). Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics*, *68*, 571–580.

O'Neill, E. M., Schwartz, R., Bullock, C. T., Williams, J. S., Shaffer, H. B., Aguilar-Miguel, X., … Weisrock, D. W. (2013). Parallel tagged amplicon sequencing reveals major lineages and phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*) species complex. *Molecular Ecology*, *22*, 111–129.

Peng, Z., Elango, N., Wildman, D. E., & Soojin, V. Y. (2009). Primate phylogenomics: Developing numerous nuclear non-coding non-repetitive markers for ecological and phylogenetic applications and analysis of evolutionary rate variation. *BMC Genomics*, *10*, 247.

Pluzhnikov, A., Di Rienzo, A., & Hudson, R. R. (2002). Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics*, *161*, 1209–1218.

Pluzhnikov, A., & Donnelly, P. (1996). Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, *144*, 1247–1262.

Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., … Altshuler, D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature Genetics*, *32*, 135–142.

Rosenberg, N. A., & Nordborg, M. (2002). Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, *3*, 380–390.

RoyChoudhury, A. (2011). Composite likelihood-based inferences on genetic data from dependent loci. *Journal of Mathematical Biology*, *62*, 65–80.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., … Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, *409*, 928–933.

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2013). Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology*, *63*, 83–95.

Smith, J. J., Kump, D. K., Walker, J. A., Parichy, D. M., & Voss, S. R. (2005). A comprehensive expressed sequence tag linkage map for tiger salamander and Mexican axolotl: Enabling gene mapping and comparative genomics in *Ambystoma*. *Genetics*, *171*, 1161–1171.

Springer, M. S., & Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, *94*, 1–33.

Wakeley, J. (2008). *Coalescent theory: An introduction* (Vol. 1). Greenwood Village: Roberts & Company Publishers.

Wakeley, J., King, L., Low, B. S., & Ramachandran, S. (2012). Gene genealogies within a fixed pedigree, and the robustness of Kingman's Coalescent. *Genetics*, *190*, 1433–1445.

Wall, J. D. (2000). A comparison of estimators of the population recombination rate. *Molecular Biology and Evolution*, *17*, 156–163.

Wall, J. D. (2003). Estimating ancestral population sizes and divergence times. *Genetics*, *163*, 395–404.

Wang, I. J., Johnson, J. R., Johnson, B. B., & Shaffer, H. B. (2011). Effective population size is strongly correlated with breeding pond size in the endangered California tiger salamander, *Ambystoma californiense*. *Conservation Genetics*, *12*, 911–920.

Wang, Y., & Rannala, B. (2008). Bayesian inference of fine-scale recombination rates using population genomic data. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*, 3921–3930.

Wang, Y., & Rannala, B. (2009). Population genomic inference of recombination rates and hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 6215–6219.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Kunstner, A., … Wilson, R. K. (2010). The genome of a songbird. *Nature*, *464*, 757–762.

Wilson, A. C., Cann, R. L., Carr, S. M., George, M., Gyllenstein, U. B., Helm-Bychowski, K. M., … Stoneking, M. (1985). Mitochondrial DNA and two perspectives on evolutionary genetics. *Biological Journal of the Linnean Society*, *26*, 375–400.

How to cite this article: Jennings WB. On the independent gene trees assumption in phylogenomic studies. *Mol Ecol*. 2017;00:1–10. https://doi.org/10.1111/mec.14274