Received: 24 October 2016 | Accepted: 13 June 2017

DOI: 10.1111/2041-210X.12846

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

Detecting hybridization by likelihood calculation of gene tree extra lineages given explicit models

Melisa Olave[1,2] | Luciano J. Avila[1] | Jack W. Sites Jr.[3] | Mariana Morando[1]

[1]Patagonian Institute for the Study of Continental Ecosystems – The National Scientific and Technical Research Council (IPEEC-CONICET), Puerto Madryn, Chubut, Argentina

[2]Department of Biology, University of Konstanz, Konstanz, Germany

[3]Department of Biology and M. L. Bean Life Science Museum, Brigham Young University (BYU), Provo, UT, USA

Correspondence
Melisa Olave
Email: olave@cenpat-conicet.gob.ar

Funding information
Office of the Director, Grant/Award Number: 0530267; ANPCYT-FONCYT, Grant/Award Number: PICT 2006-506 and 33789; Consejo Nacional de Investigaciones Científicas y Técnicas; Fulbright-Bunge y Born fellowship; the Brigham Young University Kennedy Center for International Studies; Department of Biology; Bean Life Science Museum; NSF-PIRE award

Handling Editor: Robert Freckleton

Abstract

1. Explanations for gene tree discordance with respect to a species tree are commonly attributed to deep coalescence (also known as incomplete lineage sorting [ILS]), as well as different evolutionary processes such as hybridization, horizontal gene transfer and gene duplication. Among these, deep coalescence is usually quantified as the number of extra linages and has been studied as the principal source of discordance among gene trees, while the other processes that could contribute to gene tree discordance have not been fully explored. This is an important issue for hybridization because interspecific gene flow is well documented and widespread across many plant and animal groups.
2. Here, we propose a new way to detect gene flow when ILS is present that evaluates the likelihood of different models with various levels of gene flow, by comparing the expected gene tree discordance, using the number of extra lineages. This approach consists of proposing a model, simulating a set of gene trees to infer a distribution of expected extra lineages given the model, and calculating a likelihood function by comparing the fit of the real gene trees to the simulated distribution. To count extra lineages, the gene tree is first reconciled within the species tree, and for a given species tree branch the number of gene lineages minus one is counted. We develop a set of R functions to parallelize software to allow simulations, and to compare hypotheses via a likelihood ratio test to evaluate the presence of gene flow when ILS is present, in a fast and simple way.
3. Our results show high accuracy under very challenging scenarios of high impact of ILS and low gene flow levels, even using a modest dataset of 5–10 loci and 5–10 individuals per species.
4. We present a powerful and fast method to detect hybridization in the presence of ILS. We discuss its advantage with large dataset (such as genomic scale), and also identifies possible issues that should be explored with more complex models in future studies.

KEYWORDS

deep coalescence, gene flow, hybridization, likelihood, model-based analysis

1 | INTRODUCTION

There has been a growing interest among biologists in understanding how frequent the hybridization process occurs in nature (Abbott, Barton, & Good, 2016; Abbott et al., 2013; Mallet, 2005, 2007), as it is a clue for understating the maintenance of breeding barriers between diverging linages and how new species are born (Payseur & Rieseberg, 2016). Hybridization leaves detectable footprints in genomes, and thus DNA information can be used to evaluate the interspecific gene flow hypothesis. Gene trees are often discordant with each other and also with the species tree, and common explanations for this pattern include deep coalescences (also known as incomplete lineage sorting [ILS]) and hybridization (Funk & Omland, 2003; Maddison, 1997). The deep coalescences occur due to stochastic segregation and persistence of gene lineages during the speciation process (Figure 1), and this is more likely with short speciation times and large effective population sizes (N_e; Leaché & Rannala, 2011). This stochastic segregation of multiple independent loci has been incorporated into mathematical models based on coalescent theory, in which the evolutionary history of a set of samples is studied by moving backward in time (Kingman, 1982; Pamilo & Nei, 1988; Tajima, 1983; Takahata & Nei, 1985; Wakeley, 2008). Many studies have focused on deep coalescence properties (e.g. Degnan & Rosenberg, 2006, 2009; Degnan & Salter, 2005; Rosenberg, 2003, 2013; Rosenberg & Tao, 2008; Than & Rosenberg, 2013), and others have used this source of variation as information to infer phylogenies (e.g. ASTRAL: Mirarab & Warnow, 2015; *BEAST: Heled & Drummond, 2010; BEST: Liu & Pearl, 2007; BUCKy: Ané, Larget, Baum, Smith, & Rokas, 2007; MDC: Maddison & Knowles, 2006; Than & Nakhleh, 2009; MP-EST: Liu, Yu, & Edwards, 2010; STEM: Kubatko, Carstens, & Knowles, 2009). Researchers have been able to test and distinguish among alternative hypotheses to infer evolutionary patterns and processes using explicit coalescent models.

While deep coalescences have been studied as the principal source of discordance among gene trees, other sources that could contribute to this discordance are not yet well explored. This is especially true for hybridization, even though gene flow between species is well documented and more common than previously thought (Mallet, 2005, 2007). Some studies have estimated the effect of hybridization as a source of variation among genomic sequences. For example, the ABBA/BABA algorithm tests for an excess of shared derived variants (Green et al., 2010), as well as gene trees, by observing a deviation of the null hypothesis from a strict coalescent model (e.g. Blanco-Pastor, Vargas, & Pfeil, 2012; Buckley, Cordeiro, Marshall, & Simon, 2006; Joly, McLenachan, & Lockhart, 2009; Maureira-Butler, Pfeil, Muangprom, Osborn, & Doyle, 2008). Others (Gerard, Gibbs, & Kubatko, 2011) have explicitly tested the hybridization hypothesis by extending Meng and Kubatko's (2009) model to estimate speciation and hybridization events in the presence of ILS.

One approach that uses gene tree discordance to evaluate hypotheses is to simulate data under a coalescent model, estimate a distribution of expected variation, and then compare the fit of real data to the alternative hypotheses (e.g. Richards, Carstens, & Knowles, 2007). The coalescence times for alleles from different species are expected to be greater than the species divergence times. Thus, one way to quantify gene tree discordance is by calculating the deep coalescence cost of a given gene tree within a species tree by counting the number of "extra lineages," as proposed by Maddison (1997) (see also Than & Nakhleh, 2009, 2010; Than & Rosenberg, 2011, 2013). To count extra lineages, the gene tree should first be reconciled within the species tree (Figure 1), and for each branch of the species tree the number of gene lineages minus one is counted (to count "extra" lineages; Maddison, 1997). The number of extra lineages is interpreted as the deep coalescence cost.

Here, we show that it is possible to evaluate the likelihood of different models including various levels of gene flow by comparing the number of extra lineages. We propose a new approach that can be applied to test the hybridization hypothesis in any organism from which independent nuclear loci can be sequenced. The method consists of proposing a model, simulating a set of gene trees to estimate a distribution of expected extra lineages given the model, and calculating a likelihood function by comparing the fit of the real gene trees to the simulated distribution. The rationale for this approach is that by including post-divergence gene flow, there is a displacement of the distribution of extra lineages expected under a coalescent-only model, which results in a greater mean. Thus, counting extra lineages could be a useful way to quantify the gene tree discordance in an empirical dataset with respect to a model, even when the source of variation is due to both ILS and gene flow. Although similar approaches based on comparing variation among gene trees have been studied before (Buckley et al., 2006; Joly et al., 2009), here we extend the method by explicitly testing hybridization hypotheses, incorporating gene flow into the models, and calculating the likelihood of the data. We also developed a set of R functions to parallelize software to allow simulations, and evaluate alternative hypotheses with a likelihood ratio test to assess the significance of gene flow in the presence of ILS, in a fast and simple way. Our results show high accuracy under challenging scenarios of extensive ILS and low gene

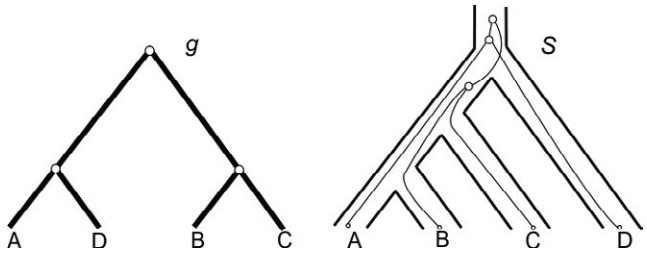

FIGURE 1 Counting extra lineages; this example is taken from Than and Rosenberg (2013). The gene tree (*g*, left) is mapped within the species tree (*S*) on the right. The number of extra lineages is counted as the summation of $n_e - c_e - 1$; where n_e is the total of elements in an edge (*e*) and c_e is the number of internal nodes in *e*. In this example, e_3 contains two lineages and no nodes, thus 2–0–1 = 1. In contrast, e_2 has three lineages and one node, thus 3–1–1 = 1, whereas there are no extra lineages in any of the pendant edges or in the edge above the root of *S* (e_1), given three lineages and two nodes results in 3–2–1 = 0. Hence, the total number of extra lineages is 2

flow levels, given modest datasets of 5–10 loci and 5–10 individuals per species.

2 | MATERIALS AND METHODS

2.1 | Methodological outline

This approach involves inferring a distribution of expected extra lineages given a model and the fit with real data (i.e. gene trees) to evaluate the likelihood of the different proposed models. The sequence of steps from 1 to 4 is shown in Figure 2 and corresponds to:

1. *Model construction*: the proposed model includes the species tree topology and branch lengths (*S*) in coalescent units (CU = t/N_e: where t is generations and N_e is the effective population size) of any given number of species, in a coalescent framework, and could potentially include any magnitude of exchange of the migrant parameter ($M = N_e m$; where m is the proportion of individuals that migrate per generation) between two or more species, occurring at any given time during their speciation histories. Thus, *M* represents the number of migrants coming into a population per generation.
2. *Inference of expected extra lineages given a model*: a set of *H* gene trees is simulated following the model described in (1). Extra lineages are counted between each simulated gene tree and the species tree *S* proposed in (1), providing a distribution of the expected number of extra lineages under the model in (1). To simulate this multi-species coalescent, we used the software MS (Hudson, 2002) to simulate 10,000 gene trees. Although here we needed high precision in distribution inference to explore the quality of the approach, we also obtained very similar results by only simulating 500 gene trees. We used the package PHYLONET (Than & Nakhleh, 2009) to count extra lineages. For more simulation details see Section 2.2.
3. *Likelihood of the empirical data*: Consider a sample of gene trees $g = (g_1, g_2,..., g_N)$ for N loci, and suppose that the vector g will be summarized by classifying each gene tree by the number of extra lineages it requires to be reconciled with the species tree *S* and the migration parameter *M*. Then the data are $n = (n_1, n_2,..., n_j)$, where n_i denotes the number of gene trees in the sample of *N* loci that requires i extra lineages to be reconciled with species tree *S* having migration parameter *M*. Let the vector $p = (p_1, p_2,..., p_j)$ denote the probabilities of various numbers of extra lineages, e.g., p_i; this gives the probability of a gene tree generated from species tree *S* with migration parameter *M* requiring i extra lineages. Then the vector $n = (n_1, n_2,..., n_j)$ is a sample from a multinomial distribution with probability vector p and *N* trials, for which the likelihood function is calculated as:

$$L\left(n_1,n_2,\ldots,n_j|S,M\right) = \frac{N!}{n_1!,n_2!,\ldots,n_j!}\, p_1^{n_1} p_2^{n_2} \ldots p_j^{n_j}$$

Note that here $j < N$ and it must be chosen so that any possible observation of the number of extra lineages, whether observed in the sample or not, be classified into a multinomial category (otherwise, the total of all the cell probabilities will not be (1). Now, let the vector $h = (h_1, h_2,..., h_j)$ be the counted number of gene trees with each number of extra lineages, then we can estimate the vector p with $\hat{p} = (\frac{h_1}{H}, \frac{h_2}{H}, \ldots \frac{h_j}{H})$.

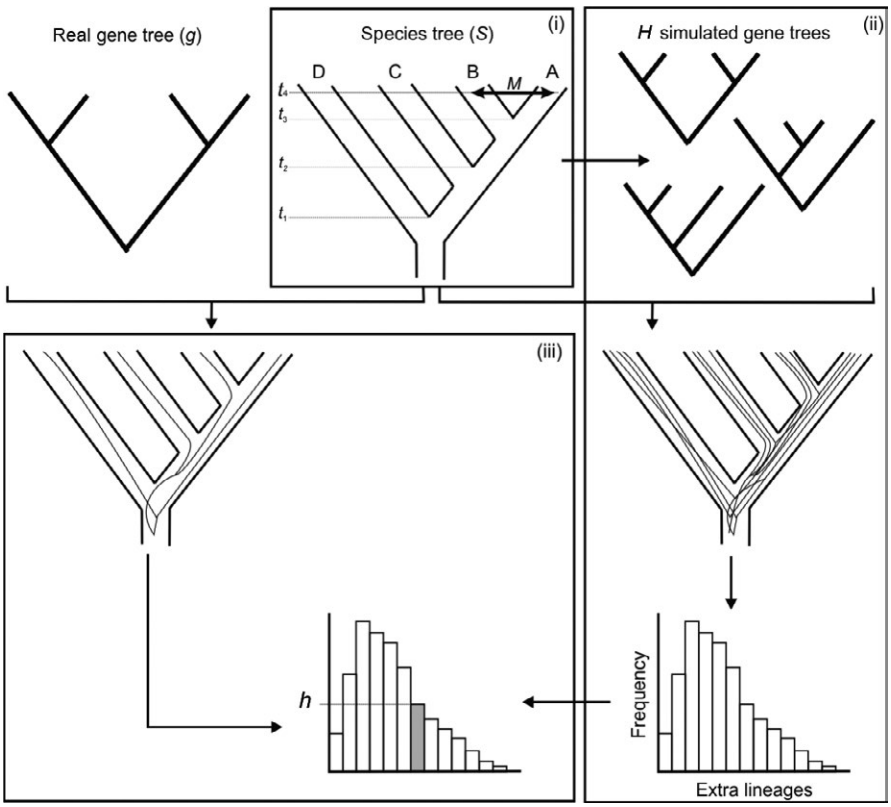

FIGURE 2 Methodological outline. The first step (i) corresponds to the proposed model, including coalescence, a species tree topology including branch lengths (*S*), a magnitude of *M* parameter and time of occurrence. Then (step ii), a set of *H* gene trees is simulated, the number of extra lineages between each of these and the species tree *S* are counted and a distribution of expected extra lineages is estimated. Finally (step iii), the number of extra lineages in the real gene tree within the species tree *S* is counted and compared to the distribution obtained in (ii). The number of simulated gene trees matching exactly the same number of extra lineages as the empirical gene tree *g*, is equal to *h*. Thus, *h*/*H* approximates the probability of observing the number of extra lineages in the empirical gene tree *g* given the model (*S*, *M*)

4. *Model selection*: A likelihood ratio test is then used to examine whether there is evidence of hybridization in the presence of ILS (χ^2). The likelihood ratio test statistic *G* is calculated as:

$$G = 2(\ln L_1 - \ln L_0)$$

where both L_0 and L_1 are calculated following description in iii, with $M = 0$ for the case of L_0 and L_1 corresponds to the maximum value of the likelihood of the data, calculated for a model with $M > 0$. The likelihood ratio test computes χ^2 with *k* degrees of freedom, where *k* is the difference of the number of parameters between L_0 and L_1. Here, we calculated the likelihood of different models including different magnitudes of the *M* parameter (i.e. $M = 0$; $M > 0$), and all other parameters remain equal (including speciation times and θ, see Section 2.3 for recommendation for estimation of this parameters in empirical analyses).

We developed a set of R functions to parallelize software and automate these steps. Function attributes are summarized in Table S1, and are available in https://github.com/melisaolave (see Olave, Avila, Sites, & Morando, 2017).

2.2 | Simulations

To evaluate the accuracy in detecting hybridization in the presence of ILS, a total of 25 replicated analyses were performed under different scenarios and combinations of loci (a total of 12,000 analyses), assuming a known species tree topology and branch lengths. Our simulations included two focal recently divergent sister species (A and B) and two extra species (C and D); note that detecting gene flow between sister taxa represents a more challenging scenario than for more distantly divergent species. The four-taxon symmetric and asymmetric species trees (Figure 2ii) were generated in Mesquite v2.74 (Maddison & Maddison, 2010).

Because lineages are likely to have coalesced within each population after $5N_e$ generations along species tree branches (here N_e is the effective number of chromosomes), and monophyly of lineages is probable (and, therefore, congruence is expected between gene trees and the species tree (Hudson & Coyne, 2002; Hudson & Turelli, 2003)), we simulated symmetric and asymmetric tree topologies fixing the divergence of the focal species (A and B) to $0.66N_e$, $1.33N_e$ and $2.66N_e$, for a total tree depth (t.d.) of $2N_e$, $4N_e$ and $8N_e$ scenarios, respectively. For the symmetric tree scenario, the divergence of C and D species was also fixed to the same divergence time as the A and B focal species, and for the asymmetric tree topology, C species was fixed to $1.33N_e$ ($2N_e$ t.d.), $2.33N_e$ ($4N_e$ t.d.) and $5.33N_e$ ($8N_e$ t.d.), and D species was fixed to $2N_e$ ($2N_e$ t.d.), $4N_e$ ($4N_e$ t.d.) and $8N_e$ ($8N_e$ t.d.). These scenarios of different tree depths represent more and less difficult conditions respectively, due to higher ILS impact respectively (Degnan & Rosenberg, 2009).

Coalescent genealogies were generated for five and ten individuals per species for each species tree using the program MS (Hudson, 2002), under a model of constant population size and no recombination within loci. Following Maddison and Knowles (2006), DNA sequences were simulated with the program SEQ-GEN (Rambaut & Grassly, 1997) under parameters that may constitute a good representation of reality (Huang, He, Kubatko, & Knowles, 2010; Olave, Solà, & Knowles, 2014; Tonini, Moore, Stern, Shcheglovitova, & Ortí, 2015). Specifically, the HKY model is a commonly used model in phylogenetics literature, and is characterized by a moderate level of complexity and flexibility in terms of the number of estimated parameters. Thus, the HKY model was selected for simulation of the nucleotide substitution process, with a transition-transversion ratio of 3.0, a gamma distribution with shape parameter of 0.8, and nucleotide frequencies of $A = 0.3$, $C = 0.2$, $T = 0.3$, and $G = 0.2$.

Specifically, 500 base pairs were generated, with two θ values = 0.01 and 0.001, describing higher and lower mutations rates respectively ($\theta = 4N_e\mu$; where μ is the mutation rate). Gene trees were reconstructed using the neighbour joining method implemented in PAUP* v4 software (Swofford, 2002; note that we detected high power despite the potential for errors in gene tree inferences with the NJ approach, we used this model because it provides a fast inference that makes our simulation method computationally feasible). We simulated strict coalescence models ($M = 0$), as well as the presence of ILS and different magnitudes of migration parameters occurring at single events between two focal species (A and B), corresponding to $M = 0.5$, 1, 2 and 5, from A to B, describing lower and higher rates of gene flow between species. We selected simulated *M* values to have a good representation of gene flow levels by considering that $M = 1$ is sufficient to overcome the effects of genetic drift and that $M > 4$ indicates that there has been general mixing of the populations (Wright, 1931). We simulated relatively recent migration, corresponding to times of 0.1, 0.25, 0.5 coalescent units, from the tips to the past, for total depths of 2, 4 and $8N_e$ respectively. Datasets were simulated for analyses of combinations of five, 10, 20 and 30 loci.

We analysed simulated datasets following the steps listed in Section 2.1. For each of these, we evaluated the likelihood of a total of five different models, setting $M = 0$, 0.5, 1, 2, 5. The species tree topologies (i.e. symmetric or asymmetric) and branch lengths (i.e. 2, 4 or $8N_e$) were set concordant with the real species tree used to simulate the data. We are interested in detecting hybridization, thus any magnitude estimated for $M > 0$ when gene flow is present is considered a successful result, as well as a value of $M = 0$ for the case of a strict coalescent model (no gene flow).

2.3 | Real data analyses

We also analysed the Olave, Avila, Sites, and Morando (2011) empirical dataset for the Argentinean lizards *Liolaemus gracilis* and *L. bibronii*. Hybridization between these species was hypothesized based on mitochondrial introgression (Morando, Avila, Turner, & Sites, 2007), and was further tested by Olave et al. (2011) that included more samples, three nuclear genes and morphological data in a phylogeographic framework. This study demonstrated hybridization between *L. gracilis* and *L. bibronii* species, and here we incorporated this dataset to show that our method is effective not just in simulated sequences but also in empirical datasets (see Olave et al., 2011 for more details).

We estimated a species tree with *BEAST v1.6.2 (Heled & Drummond, 2010) using two mitochondrial and three nuclear genes and 16 individuals of *L. bibronii*, 18 of *L. gracilis* and three outgroup species with only one sample each (missing data = 5%); a total of four gene trees (the mitochondrial loci were used to infer a single gene tree). All analyses were run for 100 million generations, sampled every 10,000 generations, and 10% of the data discarded as burn-in. Convergence was diagnosed by observing effective sample size (ESS) values equal to or greater than 200.

Because hybrid samples are used to estimate the species tree, the divergence time between *L. gracilis* and *L. bibronii* is expected to be underestimated. This is because *BEAST assumes all gene tree discordance is due only to deep coalescence, and post-divergence gene flow forces the inferred speciation times to delay the gene flow event (see Leaché, Harris, Rannala, & Yang, 2014). Thus, we evaluated the impact of using the full matrix (i.e. including hybrid species), and the effect of removing those hybrid samples (i.e. individuals with introgressed mitochondria) from the species tree estimations.

Species tree branch lengths were converted to coalescent units to be compatible with the MS program (MS branch length units = $4N_e$ generations). This was done by multiplying the branch length by the population parameter $\frac{1}{\theta}$ ($\theta = 4N_e\mu$; N_e is the effective population size and μ the mutation rate) for simulating nuclear gene trees, and for mitochondrial gene tree by $\frac{4}{\theta}$. We used LAMARC v2.1.8 software (Kuhner, 2006) to estimate θ values, also including gene flow in the model, in a Bayesian inference of 31,000 MCMC steps, with 10% burnin. Diagnosis of convergence was made by observing ESS values equal or greater than 200.

We compared the likelihood of seven and nine different models for the full and no-hybrid matrices, respectively. Olave et al. (2011) estimated two possible values for *M* (1.7 and 2.64, inferred based on haplotype information (Nei, 1973), and a F_{st} summary statistic (Hudson, 1992 respectively). Therefore, we included combination of parameters: *M* = 0, 1.7 and 2.64, and migration time at *t* = 0.01, 0.02, and $0.04N_e$ for the case of the full matrix (ingroup divergence estimated $0.08N_e$; $\theta = 0.020639$), and *t* = 0.01, 0.25, 0.5, and 0.75 for the case of no-hybrid matrix (ingroup divergence estimated $1N_e$; $\theta = 0.0212$). To infer the distribution of expected extra lineages for each model, we simulated 10,000 gene trees. Simulating an independent distribution for each gene tree prevents incorrect estimates due to missing data (this was not needed for the simulation study, as it did not have missing data).

2.4 | Robustness test to violation of species tree prior

The method proposed here assumes a known species tree topology and branch lengths. This prior information could be problematic, since the distribution of expected extra lineages depends on the proposed species tree. Note that this method is designed to be applied only to a small number of species, and we strongly recommend restricting the number of outgroups (see Discussion comments below). Following this recommendation will prevent prior errors in the topology of the species tree. Thus, here we will focus on testing the violation to branch length estimations. Because species tree methods tend to underestimate divergence of lineages with past or ongoing gene flow in order to accommodate the amount of gene tree discordance (Leaché et al., 2014), the estimation of the expected number of extra linages will be higher. This could lead to an increase in detection of false negatives in hybridization, because the amount of gene tree discordance could potentially be explained by only ILS. Thus, we have included a robustness test for branch length errors in the species tree prior. We followed the same type of simulations described above, including 25 replicates for the most extreme scenarios of the lowest and highest migration parameters (*M* = 0.5 and 5) and $0.66N_e$ and $2.66N_e$ divergence of sister taxa. In order to approximate a better estimation of this deviation in the results, we only included focal species A and B, under $\theta = 0.01$ and combinations of five and 10 individuals per species and five, 10, 20 and 30 loci. Three different scenarios of branch length underestimation were considered, including 10%, 25% and 50% of underestimation (with respect to the species divergence = $0.66N_e$ and $2.66N_e$).

2.5 | Comparisons with available methods

We compared the accuracy and time required by our method with LAMARC v2.1.8 (Kuhner, 2006) and IMA2 v8.27 (Hey, 2010), and performed ten replicate analyses using the matrices simulated for the analyses described above, but we focused only on the most challenging scenarios. We analysed simulated matrices based on the same four-taxon (symmetric and asymmetric) species trees with $0.66N_e$ and $1.33N_e$ ingroup divergence, with *M* = 0.5 and five individuals per species, for the cases of five and 30 loci. We focused on the level of successful results (i.e. detecting *M* > 0) and the time required for each analysis, and compared them with the method proposed here. The time needed by our method is highly dependent on the number of models evaluated and the number of simulated gene trees. We tested a total of five different models (*M* = 0, 0.5, 1, 2, 5) and estimated the expected distribution of extra linages based on *H* = 10,000 gene trees, as in the simulation study. In LAMARC, Bayesian searches were run for 31,000 MCMC steps, with 10% burnin, including the migration parameter in the model, and also estimating θ values. Diagnosis of convergence was made by observing ESS values equal or greater than 200. The migration parameter estimated in LAMARC is the per-generation migration rate, divided by μ, the per-site mutation rate. Thus, to convert this migration value into *M* (as treated here), the results were multiplied by the θ value of the recipient population. Although IMa2 allows multiple lineages, including multiple species also increase significantly the time consumption, thus we only tested for hybridization between focal species A and B (i.e. outgroups exluded). We ran four MCMC chains saving 20,000 geanologies, with geometric heating ($h1 = 0.96$, $h2 = 0.90$). Upper bounds were set for θ (100), τ (25), and *M* (10), and discarding 25,000 initial states. We provided the true species tree to be used as the guide tree and specified a full model with migration between sampled populations only (option –j3). We calculated a likelihood ratio test for two models, including (1) No migration

and (2) Equal migration from A to B and from B to A. Constant populations size was assumed in both cases. All analyses were run on a desktop IMAC computer (2.4 GHz Intel Processador 2 GB RAM 667 MHz).

3 | RESULTS

3.1 | Inference of expected extra lineages

We followed steps (1) and (2) described in Section 2.1 (Figure 2) and generated density plots based on 10,000 simulated gene trees. Density plots showing the expected number of extra lineages given each model simulated here, for asymmetric and symmetric trees, are quite similar (Figure 3 and Figure S1 respectively), with little deviation increasing the number of extra lineages in the symmetric tree.

As expected, density plots show that increasing the total tree depth (from $2N_e$ to $8N_e$) results in a lower variance. This pattern is also observed by decreasing the number of individuals from 10 to 5 per species. Although decreasing the total depth of the tree (from $8N_e$ to $2N_e$) shifts the curve to a more normally distributed shape, increasing the *M* parameter results in a deviation of the distribution, and its displacement on the *x*-axis that increases the mean and variance. This is because the program is reconciling each gene tree within the species tree, treating a priori all gene tree discordance as deep coalescence, and thus forcing all gene trees to coalesce before the speciation event. Adding migration to the model leads to higher gene tree-species tree conflict, and a greater number of extra linages is counted. Increasing the number of individuals from five to ten has little impact when ILS is low (≥$4N_e$ t.d. scenarios) and the *M* parameter is absent. However, density plots of higher impact of ILS (<$4N_e$ t.d.) and *M* (≥5) parameters show a clear separation between the distribution curves using 5 versus 10 individuals.

3.2 | Power in detecting hybridization

Simulation results for both symmetric and asymmetric species trees are quite similar (Figure 4 and Figure S2 under $\theta = 0.01$; Figure 5 and Figure S4 under $\theta = 0.001$). However, the method shows higher power under the case of the asymmetric tree scenario, in the particular case of the most challenging scenario treated here, corresponding

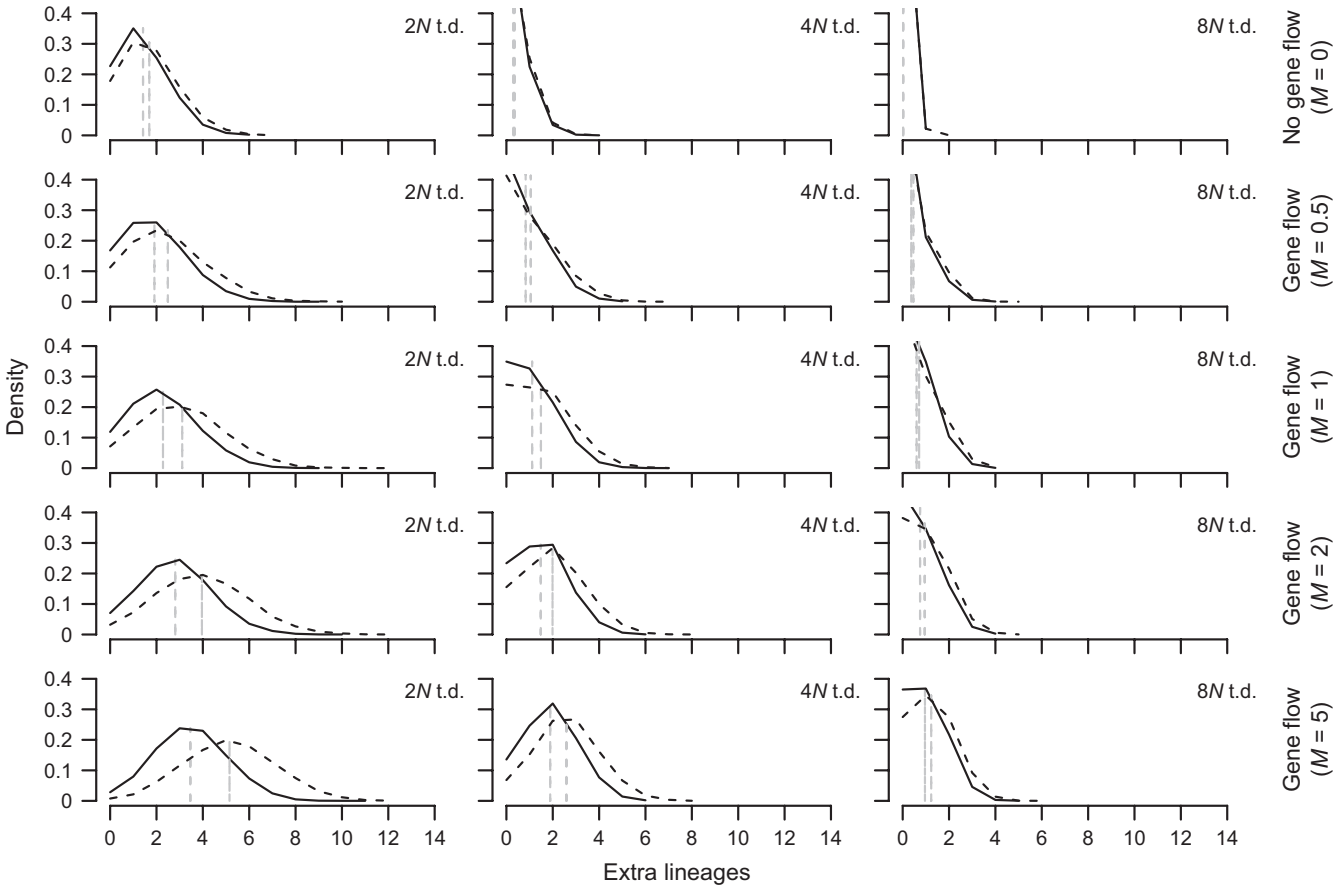

FIGURE 3 Density plots for the asymmetric species tree showing the distribution of extra lineages counted for a total of 10,000 simulated trees (step [ii] in Figure 2). The *x*-axis corresponds to the number of extra lineages counted, and *y*-axis corresponds to the frequency of observed extra lineages. Different total depths (t.d.) are shown in columns, and different values for the *M* parameter are shown in rows. Matrices of five individuals per species are shown in bold lines, and ten individuals per species in dotted lines. Vertical lines correspond to the mean calculated per each distribution

to 0.66N_e for the ingroup divergence and M = 1. Scenarios simulated under a lower θ = 0.001 show a decreasing power of the method (Figure 5) when very recent divergence times are proposed (=0.66N_e), but power remains high when divergence times become larger (≥1.33N_e).

Under the scenario of θ = 0.01 (Figure 4), high impact of ILS (=0.66N_e) and small parameter values (M = 1), detecting gene flow when it is present becomes powerful in larger matrices with ten individuals per species and/or with more loci. In the symmetric tree with larger values of M (≥2), five individuals per species and five loci are enough for >90% accuracy. However, adding more individuals or loci was needed to achieve the same power for the situation on the asymmetric tree (specifically ten individuals per species were needed).

Under a more favourable scenario (≥1.33N_e), we found that gene flow was detected in >90% of the replicated analyses of the scenarios studied here, including very low values of M (=0.5) and low mutation rates (θ = 0.001); this result holds for the smaller dataset of five loci and five individuals per species, with M = 0.5.

3.3 | Real data analysis

Results for the empirical data from the *L. gracilis* - *L. bibronii* study recovered M > 0 as the most likely result for both cases of the full and no-hybrid matrices (Table 1). The model selected is for the case of estimating the species tree using the full matrix with M = 2.64 and t = 0.01, and M = 1.7 and t = 0.01, using the no-hybrid matrix.

3.4 | Violation of known species tree assumption

Results for violation of the known species tree prior show a decrease in power under the most complicated scenario of 50% underestimation of branch length priors, and specifically for the case of most recent species divergence (=0.66N) and with a very low migration parameter (M = 0.5; Figure 6). In this scenario our simulations show that most cases of hybridization are rejected. However, under a scenario of stronger gene flow (M = 5) and/or given enough data (e.g. 10 individuals per species or 30 loci), hybridization can be detected in most of the cases (>0.75). When species divergence was increased (=2.66N), the

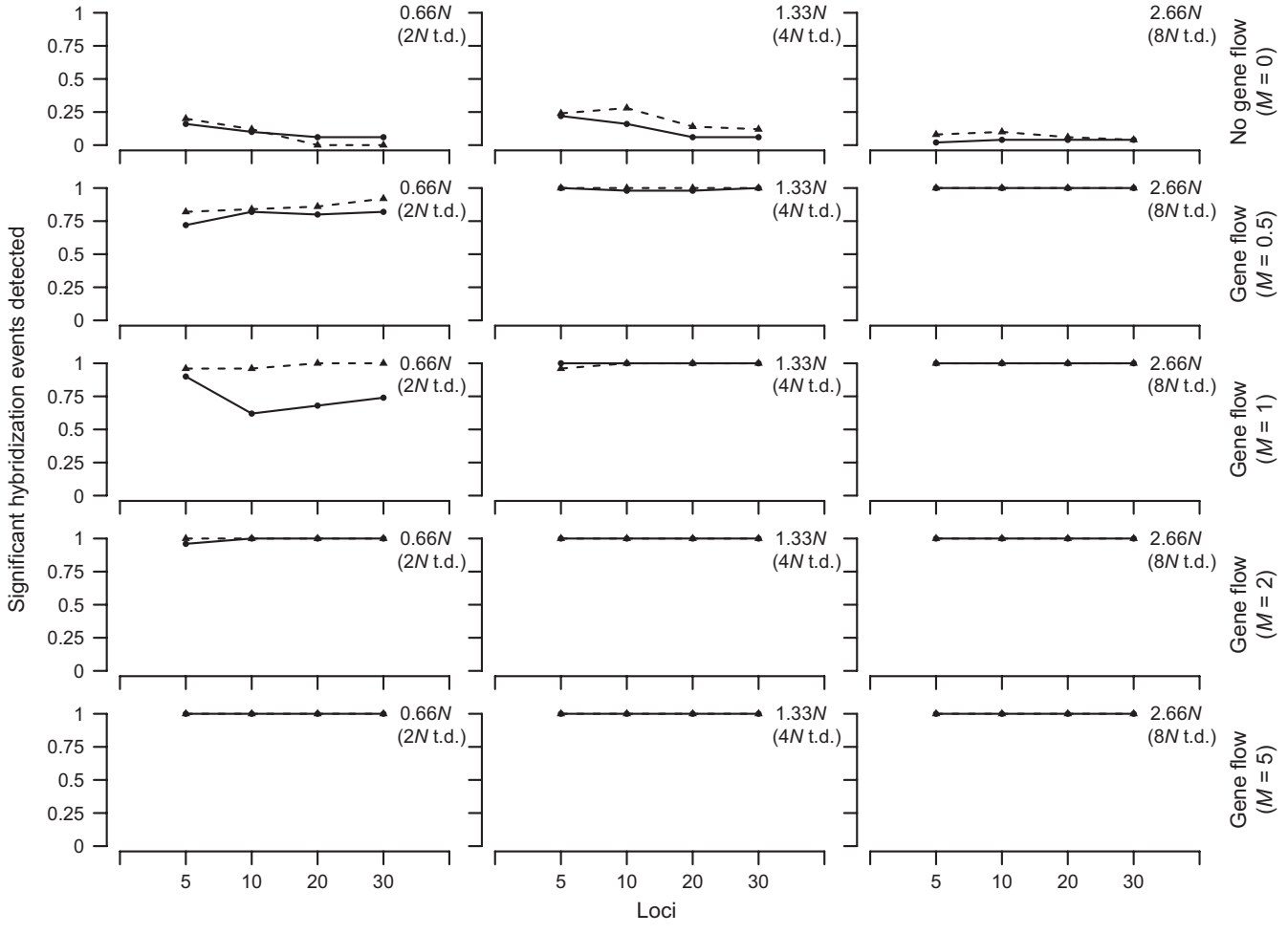

FIGURE 4 Power of the method applied to the asymmetric species tree among 25 replicated analyses per scenario studied here, using simulated data under θ = 0.01. The *x*-axis corresponds to the number of loci used, and the *y*-axis corresponds to the proportion of significant hybridization events detected among analyses. Different depths for ingroup divergence are shown in columns, and different values for M parameter are shown in rows. Matrices of five individuals per species are shown in bold lines, and ten individuals per species in dotted lines

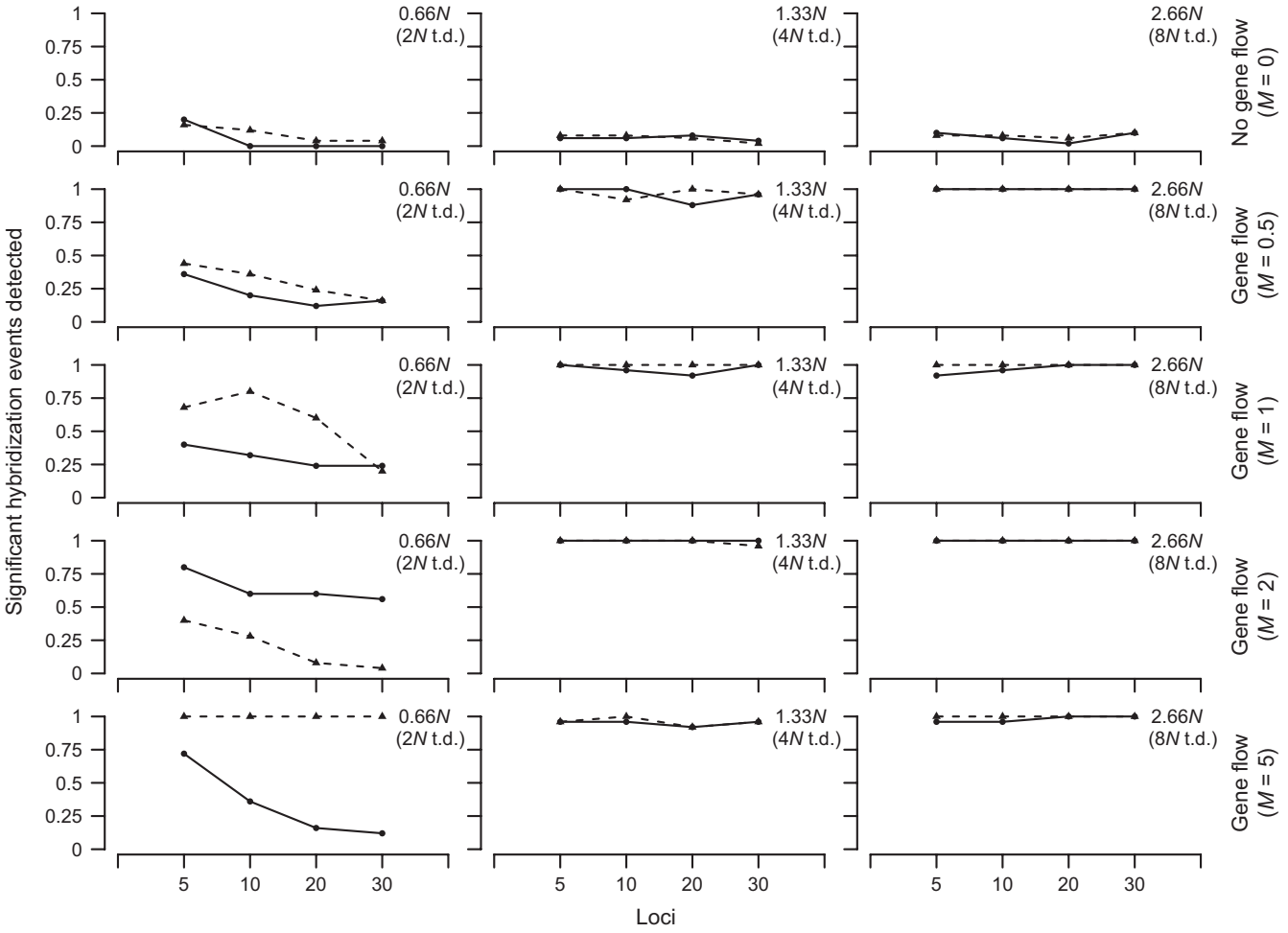

FIGURE 5 Power of the method applied to the asymmetric species tree among 25 replicated analyses per scenario studied here, using simulated data under $\theta = 0.001$. The *x*-axis corresponds to the number of loci used, and the *y*-axis corresponds to the proportion of successful results. Different depths for ingroup divergence are shown in columns, and different values for *M* parameter are shown in rows. Matrices of five individuals per species are shown in bold lines, and ten individuals per species in dotted lines

power of the method remains high even under branch length underestimation = 50%, and the largest proportion of results were able to detect hybridization.

3.5 | Comparisons with available methods

LAMARC software results were powerful, recovering $M > 0$ in each replicate when it was present (Table S2). Also IMa2 detected hybridization sign in all replicates, by selecting equal gene flow from A to B and B to A (model 2 described in Section 2.5).

There are not important differences in time consumption in any method when analysing either asymmetric or symmetric trees, as well as between different scenarios of ingroup divergence $0.66N_e$ and $1.33N_e$ (Figure 7 [asymmetric tree] and Figure S4 [symmetric tree]). With small matrices of five loci, LAMARC required in average ~72 min to estimate the migration parameter. For the same case, IMa2 required an average of ~20 hr and 12 min, whereas our method needed ~6 hr 50 min. In contrast, with the large 30-locus matrix, LAMARC needed ~19 hr 48 min and IMa2 required ~141 hr and 30 min, compared to the same average time of ~6 hr and 50 min for our method.

4 | DISCUSSION

4.1 | Power of the method

The method proposed here has the power to detect gene flow for several hybridization scenarios. In general, we show that matrices of a small number of individuals per species and a modest number of loci are enough to detect hybridization (Figure 4, 5). Our empirical analyses successfully detected hybridization between *L. gracilis* and *L. bibronii*, and while estimates for the *M* parameter revealed different values for the two matrices used here (full = 2.64, and no-hybrids = 1.7), all results detected $M > 0$.

Detecting smaller values of the *M* parameter becomes more challenging when the impact of ILS is higher (i.e. lower tree depth) and the *M* parameter becomes smaller (Figures 4 , 5), but it is still possible to detect hybridization in most cases when using larger datasets in the

TABLE 1 Empirical results; selected models are shown in grey

Matrix	*M*	*t*	ln Likelihood
Full	0	—	−29,5300
	1.7	0.01	−13,3850
		0.02	−14,0253
		0.04	−15,6835
	2.64	0.01	−12,2532
		0.02	−12,4610
		0.04	−13,8941
Likelihood ratio test: $4{,}1465 \times 10^{-09}$			
No-hybrids	0	—	−46,8811
	1.7	0.01	−10,8451
		0.25	−26,1016
		0.50	−30,5512
		0.75	−46,1574
	2.64	0.01	−12,1573
		0.25	−25,8643
		0.50	−28,5602
		0.75	−45,9541
Likelihood ratio test: $2{,}0749 \times 10^{-17}$			

scenarios explored here. In the particular case of lower mutation rate ($\theta = 0.001$) and very recent divergence times (=$0.66N_e$), the power of the method decreases (Figure 5). Lack of mutations led to increase the number of polytomies in the gene trees, which impacts the number of expected extra lineages. However, in more favorable scenarios of $1.33N_e$, five individuals and five loci are sufficient to detect $M > 0$, when present in >90% of our replicated analyses, even under lower mutation rates of $\theta = 0.001$ (Figure 5).

The power of the method was lower for the case of symmetric species trees; the extra lineages counted for these trees and variances are slightly greater (Figure S1). Although the divergence of our ingroup (A and B) remains the same in both symmetric and asymmetric trees, the outgroup divergence is different. The asymmetric tree topology is: (D,(C,(A,B))), while the symmetric tree topology is ((C,D),(A,B)). Divergence of C and D occurred more recently in the symmetric than in the asymmetric tree, and the probability of deep coalescence is higher in the latter. The result is a slightly more challenging scenario due to a higher impact of ILS, and this is reflected in the results depicted in Figure S2. Based on these results, we predict that increasing the number of species is likely to decrease the power of our test, and we strongly recommend using this method with few species and a restricted number of outgroups.

We have also shown good power of the method under a range of errors in the estimation of species tree branch length priors (Figure 6). Specifically, under more complicated scenarios and larger datasets (loci and individuals), it will still be important to detect hybridization. Further, branch length underestimation seems not to be a problem when hybridization signal is high ($M = 5$). If possible, we recommend that users incorporate the strategy described above for real data analyses (Section 2.3), where individuals with a strong hybridization signature are removed for species tree estimation, and later incorporated for preforming this test. This will prevent high underestimation of species divergence, and then results are more likely to be accurate.

4.2 | Comparison with available methods and advantages with large datasets

Massive datasets are challenging to analyse (Than & Nakhleh, 2009), and our algorithm provides a relatively fast and powerful method to detect hybridization (Figure 4, 5), making it a good option for analysing larger matrices and genomic data. Our method uses most of its time estimating the probability vector, as described in Section 2.1(3), and time is also very dependent on the number of *H* gene trees simulated (=10,000 here), as well as the number of models to be tested (=5 here). Although the programs selected for comparisons (i.e. LAMARC and IMA2) also returned excellent results in detecting gene flow when it was present (Table S2), and particularly LAMARC ran faster than the method presented here with smaller matrices (five loci), our method provides a clear advantage when larger matrices are analysed (Figure 7 and Figure S4). This highlights the utility of our method relative to its efficiency in computational time requirements, by having a significantly faster algorithm than LAMARC and IMA2. Given the growing interest in generating genomic scale datasets (Lemmon & Lemmon, 2013), developing methodologies that can handle large datasets represents an important contribution.

Another method that handles large matrices is the ABBA/BABA algorithm, which tests for an excess of shared derived variants (Green et al., 2010). This test considers ancestral "A" and derived "B" alleles and is based on the prediction that two particular SNP patterns, termed "ABBA" and "BABA," should be equally frequent under a scenario of ILS without gene flow, and an excess of ABBA patterns (detected using the Patterson's D statistic) is interpreted as a sign of gene flow. The method is computationally efficient and constitutes a useful way to address genomic scale datasets, but it has been criticized by Martin, Davey, and Jiggins (2015) on the grounds that an excess of shared derived variants can arise from processes other than recent introgression, in particular non-random mating in a structured ancestral population.

Other methods are available to indirectly test for hybridization. For example, Buckley et al. (2006) have proposed a method to accommodate expected gene tree discordance given a strict coalescent model and comparing the fit of real data, but this approach assumes that all violations to the null hypothesis are due to gene flow. Similar to this approach, a parametric method to infer hybridization proposed by Joly et al. (2009) estimates and compares genetic distances between sequences from two species with the prediction given by a strict coalescent model. Similarly to the Buckley et al. (2006) method, any deviation from the null hypothesis is attributed to gene flow. Although the method we describe here is similar to both of the above methods, it constitutes a significant advance because we have incorporated an explicit test for hybridization in the presence of ILS.

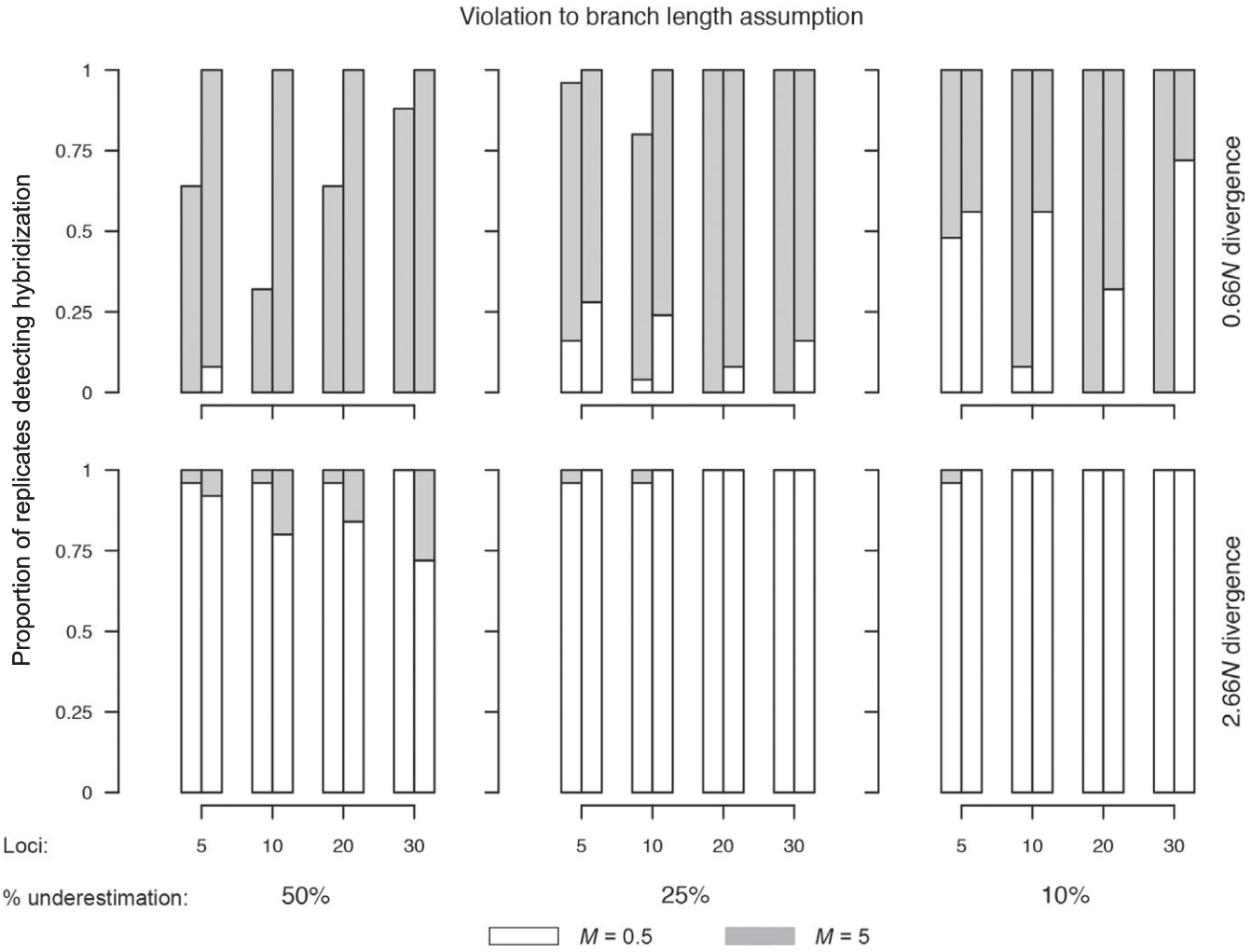

FIGURE 6 Robustness test to unknown species tree prior. Different scenarios where branch lengths in species tree prior were underestimated (50%, 25% and 10%, from the left to the right respectively), considering species divergence = 0.66N (top) and 2.66 (bottom). Different combinations of loci include five, 10, 20 and 30 loci for performing analyses are display in the x axes. All results using *M* = 5 conducted to larger proportions of replicates finding significant hybridization with respect to *M* = 0.5. Thus, white bars represent scenarios of *M* = 0.5 and grey bars for *M* = 5, and cases where only white bars are shown, means that little signal of *M* = 0.5 was enough to detect hybridization in all cases (same result for *M* = 5). Left bars are results including five individuals per species, and right bars represents results when including 10 individuals per species

Although the MESQUITE software is flexible and commonly used for several different types of phylogenetic analyses, including simulation of gene trees and counting the number of extra lineages, the combination of MS and PHYLONET parallelized using our R functions (Table S1) runs much faster than Mesquite. This advance provides options for testing more models and simulating more gene trees to improve accuracy of estimated distributions, thereby increasing the power of the method.

4.3 | Limitations and future directions

Here, we propose a new perspective among methods described above that includes: (1) using the concept of extra lineages as a measure for gene tree discordance given both ILS and gene flow, (2) calculating the likelihood of real gene trees given a model, and (3) evaluating the significance of gene flow in the presence of ILS using a likelihood ratio test. We also provide a set of functions written in R language to automate the process of simulation and model testing.

However, our methodology has some limitations. While we have shown that it is possible to study gene flow by counting the number of extra lineages, this is only a first step; other questions emerge and immediately suggest future studies. For instance, it is expected that as time-of-hybridization approaches the divergence time of the species involved, distinguishing between ILS and gene flow will be more complicated. Thus, one question is how close in time can both of these events be, yet still be detected and separated. Also, more complex models with additional variables need to be explored, such as temporal fluctuations in N_e with and without gene flow. This is important because temporal changes in N_e also affect the number of expected extra lineages. However, a more complex model including both N_e and *M* could obscure results, because similar patterns are expected due to the interaction of both parameters.

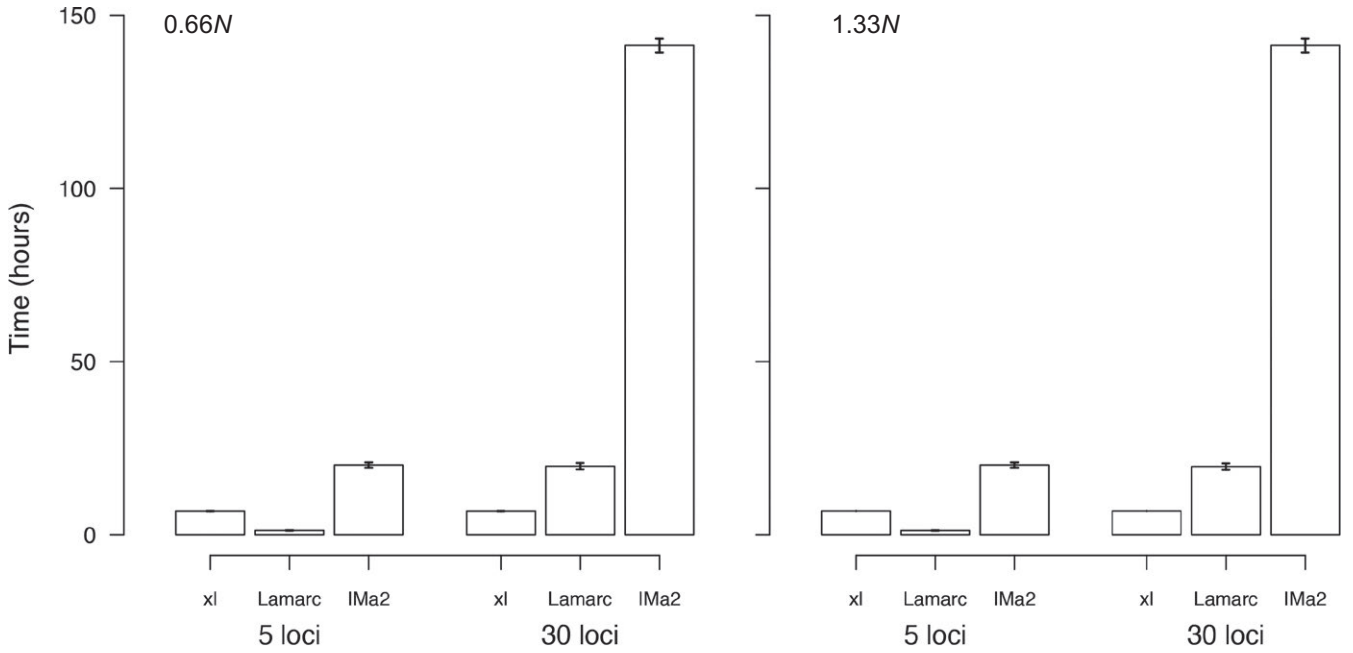

FIGURE 7 Comparisons of average time consumption by different methods: xI (extra lineage, the method presented here), Lamarc and IMa2 programs. Standard deviations are represented by lines on top of each bar. The plots show the results based on asymmetric tree scenario in Lamarc and our method (IMa2 was only ran with two species, see Section 2). Left and right plot represent the results of the scenario of ingroup divergence 0.66N and 1.33N, respectively. In each plot, the first three bars are results using a small matrix of five loci, and the next three bars are results of larger matrices of 30 loci

In addition, including more hybridizing species into the same data matrix could obscure results. For this case we predict that it will probably still be easy to detect at least one $M > 0$, but the power to detect multiple species hybridizations needs to be explored.

These cases of interaction among different processes during species divergence could be solved by incorporating different summary statistics to capture those signals under complex scenarios. It would be interesting to test also the utility of counting extra lineages in an approach such approximate Bayesian computation (ABC; see Beaumont, 2010; Beaumont, Zhang, & Balding, 2002; Csilléry, Blum, Gaggiotti, & François, 2010; Sunnaker et al., 2013), and some studies have recently focused on proposing new summary statistics that could be coupled with an ABC approach (e.g. Alvarado-Serrano & Hickerson, 2015; Peter & Slatkin, 2013). In this way, the R function that we provide here could be easily incorporated to the summary statistics calculation to compare expectations generated by simulation to real data under an ABC approach.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for useful comments and suggestions made on the first version of this article. We thank all members of the Grupo de Herpetología Patagónica (CENPAT-CONICET) for continuing support, and especially Arley Camargo for comments that improved this manuscript. We also thank Adam Leaché (University of Washington) for useful comments made on an earlier version of this manuscript. We especially thank Laura Kubatko (Ohio State University) for very useful suggestions and comments regarding the likelihood calculations. We thank all members of L. L. Knowles' lab (University of Michigan) for their time discussing preliminary results. We also thank Eduard Solà (Universitat de Barcelona) for his comments and discussions with MO, and all members of the J. Rozas' lab (Universitat de Barcelona). MO also thanks the Society for the Study of Evolution (SSE) and the European Society for Evolutionary Biology (ESEB) for travel awards (MO) to attend meetings in 2013 (Soc. for the Study of Evolution, Snowbird, UT; XIV Congress of the European Society for Evolutionary Biology, Lisbon, Portugal); both fostered multiple discussions that improved this work.

Financial support was provided by grants PICT 2006-506 ANPCYT-FONCYT (LJA), ANPCYT-FONCYT 33789 (MM), and a doctoral fellowship (MO) from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), the Fulbright-Bunge y Born fellowship (MO), the Brigham Young University Kennedy Center for International Studies, Department of Biology, and the Bean Life Science Museum, and NSF-PIRE award (OISE 0530267) for support of collaborative research on Patagonian Biodiversity granted to the following institutions (listed alphabetically): Brigham Young University, Centro Nacional Patagónico (AR), Dalhousie University, Instituto Botánico Darwinion (AR), Universidad Austral de Chile, Universidad de Concepción, Universidad Nacional del Comahue, Universidad Nacional de Córdoba, and University of Nebraska.

AUTHORS' CONTRIBUTIONS

M.O. and M.M. conceived the ideas and designed methodology. M.O. analysed the data and led the writing of the manuscript. M.M., L.J.A.

and J.W.S. provided the funding and facilities for completing optimally this research. All authors contributed critically to revise the drafts and gave final approval for publication.

DATA ACCESSIBILITY

R functions, explanations and a tutorial are available in https://github.com/melisaolave (see Olave et al., 2017). https://doi.org/10.5281/zenodo.810097.

Other related files available in Dryad Digital Repository https://doi.org/10.5061/dryad.q084s (Olave et al., 2017).

REFERENCES

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J., Bierne, N., … Butlin, R. K. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*, 229–246.

Abbott, R. J., Barton, N. H., & Good, J. M. (2016). Genomics of hybridization and its evolutionary consequences. *Molecular Ecology*, *25*, 2325–2332.

Alvarado-Serrano, D. F., & Hickerson, M. J. (2015). Spatially explicit summary statistics for historical population genetic inference. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.12489

Ané, C., Larget, B., Baum, D. A., Smith, S. D., & Rokas, A. (2007). Bayesian estimation of concordance among gene trees. *Molecular Biology and Evolution*, *24*, 412–426.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology Evolution and Systematics*, *41*, 379–406.

Beaumont, M. A., Zhang, W., & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*, 2025–2035.

Blanco-Pastor, J. L., Vargas, P., & Pfeil, B. (2012). Coalescent simulation reveal hybridization and incomplete lineage sorting in Mediterranean linaria. *PLoS ONE*, *7*, 1–16.

Buckley, T., Cordeiro, M., Marshall, D., & Simon, C. (2006). Differentiating between hypotheses of lineage sorting and introgression in New Zealand cicadas (*Maoricicada dugdale*). *Systematic Biology*, *55*, 411–425.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*, 410–418.

Degnan, J. H., & Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, *3*, 762–768.

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, *24*, 332–340.

Degnan, J. H., & Salter, L. (2005). Gene tree distributions under the coalescent process. *Evolution*, *59*, 24–37.

Funk, D. J., & Omland, K. E. (2003). Species-level paraphyly and polyphyly: Frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology Evolution and Systematics*, *34*, 397–423.

Gerard, D., Gibbs, H. L., & Kubatko, L. (2011). Estimating hybridization in the presence of coalescence using phylogenetic intraespecific sampling. *BMC Evolutionary Biology*, *11*, 291.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Pääbo, S. (2010). A draft of the neandertal genome. *Science*, *328*, 710–722.

Heled, J., & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data research article. *Molecular Biology and Evolution*, *27*, 570–580.

Hey, J. (2010). Isolation with migration models for more than two populations. *Molecular Biology and Evolution*, *27*, 905–920.

Huang, H., He, Q., Kubatko, L. S., & Knowles, L. L. (2010). Sources of error inherent in species-tree estimation: Impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Systematic Biology*, *59*, 573–583.

Hudson, R. R. (1992). Gene trees, species trees and the segregation of ancestral alleles. *Genetics*, *131*, 509–512.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, *2*, 337–338.

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

Hudson, R. R., & Turelli, M. (2003). Stochasticity overrules the "three-times rule": Genetic drift, genetic draft, and coalescence times for nuclear loci versus mitochondrial DNA. *Evolution*, *57*, 182–190.

Joly, S., McLenachan, P. A., & Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *American Naturalist*, *174*, 54–70.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Process and their Application*, *13*, 235–248.

Kubatko, L., Carstens, B. C., & Knowles, L. L. (2009). STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, *25*, 971–973.

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

Leaché, A. D., Harris, R. B., Rannala, B., & Yang, Z. (2014). The influence of gene flow on species tree estimation: A simulation study. *Systematic Biology*, *63*, 17–30.

Leaché, A. D., & Rannala, B. (2011). The accuracy of species tree estimation under simulation: A comparison of methods. *Systematic Biology*, *60*, 126–137.

Lemmon, E. M., & Lemmon, A. R. (2013). High-troughput genomic data in systematics and phylogenetics. *Annual Review of Ecology Evolution and Systematics*, *44*, 99–121.

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.

Liu, L., Yu, L., & Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, *10*, 302.

Maddison, W. (1997). Gene trees in species trees. *Systematic Biology*, *46*, 523–536.

Maddison, W. P., & Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology*, *55*, 21–30.

Maddison, W. P., & Maddison, D. R. (2010). Mesquite: A modular system for evolutionary analysis. Version 2.74. Retrieved from http://mesquiteproject.org.

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, *20*, 229–237.

Mallet, J. (2007). Hybrid speciation. *Nature*, *446*, 279–283.

Martin, S. H., Davey, J. W., & Jiggins, C. D. (2015). Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution*, *32*, 244–257.

Maureira-Butler, I. J., Pfeil, B. E., Muangprom, A., Osborn, T. C., & Doyle, J. J. (2008). The reticulate history of medicago (Fabaceae). *Systematic Biology*, *57*, 466–482.

Meng, C., & Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*, *75*, 35–45.

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*, 44–52.

Morando, M., Avila, L. J., Turner, C., & Sites, J. W. Jr. (2007). Molecular evidence for species complex in the Patagonian lizard *Liolaemus bibronii* and phylogeography of the closely related *Liolaemus gracilis* (Squamata: Liolaemini). *Molecular Phylogenetics and Evolution*, *43*, 952–973.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, *70*, 3321–3323.

Olave, M., Avila, L. J., Sites, J. W. Jr, & Morando, M. (2011). Evidence of hybridization in the Argentinean lizard *Liolaemus gracilis* and *Liolaemus bibronii* (Iguani: Liolaemini): An integrative approach based on genes and morphology. *Molecular Phylogenetics and Evolution*, *61*, 381–391.

Olave, M., Avila, L. J., Sites, J.W. .Jr, & Morando, M. (2017). Data from: Detecting hybridization by likelihood calculation of gene tree extra lineages given explicit models. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.q084s

Olave, M., Solà, E., & Knowles, L. L. (2014). Upstream analyses create problems with DNA-based species delimitation. *Systematic biology*, *63*, 263–271.

Pamilo, P., & Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, *5*, 568–583.

Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology*, *25*, 2337–2360.

Peter, B. M., & Slatkin, M. (2013). Detecting range expansion from genetic data. *Evolution*, *67*, 3274–3289.

Rambaut, A., & Grassly, N. (1997). SeqGen: An application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, *13*, 235–238.

Richards, C. L., Carstens, B. C., & Knowles, L. L. (2007). Distribution modeling and statistical phylogeography: An integrative framework for generating and testing alternative biogeographic hypotheses. *Journal of Biogeography*, *34*, 1833–1845.

Rosenberg, N. A. (2003). The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. *Evolution*, *57*, 1465–1477.

Rosenberg, N. A. (2013). Discordance of species trees with their most likely gene trees: A unifying principle. *Molecular Biology and Evolution*, *30*, 2709–2713.

Rosenberg, N. A., & Tao, R. (2008). Discordance of species trees with their most likely gene trees: The case of five taxa. *Systematic Biology*, *57*, 131–140.

Sunnaker, M., Busseto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian computation. *PLoS Computational Biology*, *9*, e1002803.

Swofford, D. L.. (2002). *PAUP*. Phylogenetic analysis using parsimony (*and other methods)*, Version 4. Sunderland, MA: Sinauer Associates.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, *105*, 437–460.

Takahata, N., & Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, *110*, 325–344.

Than, C. V., & Nakhleh, L. (2009). Species tree inference by minimizing deep coalescences. *PLoS Computational Biology*, *5*, e1000501. https://doi.org/10.1371/journal.pcbi.1000501

Than, C. V., & Nakhleh, L.. (2010). Inference of parsimonious species phylogenies from multi-locus data by minimizing deep coalescences. In L. L. Knowles & L. S. Kubatko (Eds.), *Estimating species trees: Practical and theoretical aspects* (pp. 79–98). Chichester, UK: Wiley-VCH.

Than, C. V., & Rosenberg, N. A. (2011). Consistency properties of species tree inference by minimizing deep coalescences. *Journal of Computational Biology*, *18*, 1–15.

Than, C. V., & Rosenberg, N. A. (2013). Mathematical properties of the deep coalescence cost. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*, 61–72.

Tonini, J., Moore, A., Stern, D., Shcheglovitova, M., & Ortí, G. (2015). Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. *PLOS Currents Tree of Life*, Edition 1, https://doi.org/10.1371/currents.tol.34260cc27551a527b124ec5f6334b6be

Wakeley, J. (2008). *Coalescent theory: An introduction*. Greenwood Village, CO: Roberts and Company.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Olave M, Avila LJ, Sites JW Jr, Morando M. Detecting hybridization by likelihood calculation of gene tree extra lineages given explicit models. *Methods Ecol Evol*. 2018;9:121–133. https://doi.org/10.1111/2041-210X.12846