Title: The genomic view of diversification

Julie Marin[1], Guillaume Achaz[1,2], Anton Crombach[1,3,4], Amaury Lambert[1,5]

[1] *Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS UMR 7241, INSERM UMR 1050, PSL Research University, Paris, France;*

[2] *Institut de Systématique, Évolution, Biodiversité (ISYEB), MNHN, CNRS, Sorbonne Université, EPHE, Paris, France;*

[3] *Inria, Lyon Antenne La Doua, Villeurbanne, France*

[4] *Université de Lyon, INSA-Lyon, LIRIS, UMR 5205, Villeurbanne, France*

[5] *Laboratoire de Probabilités, Statistique et Modélisation (LPSM), Sorbonne Université, CNRS UMR 8001, Paris, France.*

Corresponding author: julie.marin@college-de-france.fr, +33 1 44 27 14 09

Short running title: The genomic view of diversification

ABSTRACT: Evolutionary relationships between species are traditionally represented in the form of a tree, the species tree. Its reconstruction from molecular data is hindered by frequent conflicts between gene genealogies. Usually, these disagreements are explained by incomplete lineage sorting (ILS) due to random coalescences of gene lineages inside the edges of the species tree. This paradigm, the multi-species coalescent (MSC), is constantly violated by the ubiquitous presence of gene flow, leading to incongruences between gene trees that cannot be explained by ILS alone. Here we argue instead in favor of a vision acknowledging the importance of gene flow and where gene histories shape the species tree rather than the opposite. We propose a new framework for modeling the joint evolution of gene and species lineages relaxing the hierarchy between the species tree and gene trees. We implement this framework in two mathematical models called the gene-based diversification models (GBD): 1) GBD-forward following all evolving genomes and 2) GBD-backward based on coalescent theory. They feature four parameters tuning colonization, gene flow, genetic drift and genetic differentiation. We propose a quick inference method based on differences between gene trees. Applied to two empirical data-sets prone to gene flow, we find a better support for the GBD model than for the MSC model. Along with the increasing awareness of the extent of gene flow, this work shows the importance of considering the richer signal contained in genomic histories, rather than in the mere species tree, to better apprehend the complex evolutionary history of species.

Keywords: coalescent theory, gene flow, gene tree, gene-based diversification model, multi-species coalescent, phylogeny, population genetics, speciation, species tree, introgression.

INTRODUCTION

The most widely used way of representing evolutionary relationships between contemporary species is the so-called species tree, or phylogeny. The high efficiency of statistical methods using sequence data to reconstruct species trees, hence called 'molecular phylogenies', led to precise dating of the nodes of these phylogenies (Heled & Drummond, 2010; Kishino, Thorne, & Bruno, 2001; Tamura et al., 2012). Notwithstanding the debatable accuracy of these datings, the use of time-calibrated phylogenies, sometimes called 'timetrees' (Hedges & Kumar, 2009), has progressively overtaken a view where phylogenies merely represent tree-like relationships between species in favor of a view where the timetree is the exact reflection of the diversification process (Morlon, 2014; Pyron & Burbrink, 2013; Stadler, 2013a). In this view, the nodes of the phylogeny are consequently seen as punctual speciation events where one daughter species is instantaneously 'born' from a mother species. In this paper, we explore an alternative view of diversification, acknowledging that speciation is a long-term process (Etienne, Morlon, & Lambert, 2014; Lambert, Morlon, & Etienne, 2015; Rosindell, Cornell, Hubbell, & Etienne, 2010) and not invoking any notion of mother-daughter relationship between species as done in the timetree view. This alternative view is gene-based rather than species-based, comparable with Wu's genic view of speciation (2001). We use here the term 'gene' in the sense of "non-recombining locus", *i.e.*, a region of the genome with a unique evolutionary history. Our view is meant in particular to accommodate the well-recognized existence of gene flow between incipient species, which persists during the speciation process and long after (Mallet, Besansky, & Hahn, 2016).

The timetree view of phylogenies does acknowledge that gene trees are not independent and may disagree with the species tree (Maddison, 1997). However, current methods jointly inferring gene trees and species tree rely on two assumptions that we question in the next section: (1) there is a unique species tree, (2) the species tree shapes the gene trees and (3) the species tree is the only factor mediating all dependencies between gene trees (they are independent conditional on the species tree).

This view is materialized in a model called the 'multispecies coalescent' (MSC) (Knowles & Kubatko, 2011) where conditional on the species tree, the evolutionary histories of genes follow independent coalescents constrained to take place within the hollow edges of the species tree. Many methods have been developed to estimate the species tree under the MSC, such as full likelihood methods (e.g. BEAST, Heled

and Drummond 2010, BPP, Yang 2015) which average over gene trees and parameters (Xu & Yang, 2016), and the approximate or summary coalescent methods (e.g. ASTRAL, Mirarab et al. 2014, MP-EST, L. Liu, Yu, and Edwards 2010, and STELLS, Y. Wu 2012) which use a two-step approach: gene trees are first inferred and then combined to estimate the species tree that minimize conflicts among gene trees. Discordance between gene topologies is then explained, as a first approximation at least, by the intrinsic randomness of coalescences resulting in incomplete lineage sorting (ILS) (Fig. 1).

However, the presence of gene flow (introgression, hybridization, horizontal transfer) is now widely recognized between closely related species, and even between distantly related species (Mallet et al., 2016). Porous species boundaries, allowing for gene exchange because of incomplete reproductive isolation, are indeed regularly observed in diverse taxa such as amphibians (Fontenot, Makowsky, & Chippindale, 2011; Pereira, Monahan, & Wake, 2011), arthropods (De Busschere et al., 2010), cichlids (Willis, Macrander, Farias, & Ortí, 2012), cyprinids (Buonerba et al., 2015; Gante, Collares-Pereira, & Coelho, 2004; Gante, Doadrio, Alves, & Dowling, 2015; Gante, Santos, & Alves, 2010; Sousa-Santos et al., 2014), insects (Nadeau et al., 2013; Peccoud, Ollivier, Plantegenest, & Simon, 2009; Wahlberg, Weingartner, Warren, & Nylin, 2009), and even more frequently among bacteria (Mallet et al., 2016; Soucy, Huang, & Gogarten, 2015). Long neglected, gene flow has recently been recognized as an important evolutionary driving force, through adaptive introgression or the formation of new hybrid taxa (Abbott et al., 2013). The ubiquity of genetic exchange across the Tree of Life between contemporary species suggests that gene flow has occurred many times in the evolutionary past, and might actually be the most important cause of discrepancies between gene histories (e.g. Clark and Messer 2015; Cui et al. 2013; Gallus, Janke, Kumar, and Nilsson 2015; Jónsson et al. 2014) (Fig. 1). Accordingly, several extensions to the MSC model have been considered allowing for gene flow between species (Kubatko, 2009; Yu, Dong, Liu, & Nakhleh, 2014). These models acknowledge that species boundaries can be permeable at a few specific timepoints (Harrison & Larson, 2014). Unfortunately, because of the heavy computational cost of modeling the coalescent with gene flow, these methods are limited to small data-sets (Yu et al., 2014). More importantly, they might not be appropriate to realistically model gene flow, given the frequency of gene flow across time and clades described in empirical studies (Solís-Lemus, Yang, & Ané, 2016). Additionally, some of these methods, for instance

ASTRAL and MP-EST, might infer erroneous gene trees when gene flow is present (Long & Kubatko, 2018). These observations urge for novel approaches where gene flow is the rule rather than the exception.

To fill this void, we propose here an alternative framework and two accompanying models (one in forward time and one in backward time), the gene-based diversification (GBD) models, framed with minimal assumptions arising from recent empirical evidence. Those models rely on the property of populations to spontaneously differentiate genetically while simultaneously undergoing gene flow. This genetic differentiation is accompanied by a decrease in gene flow until reproductive isolation is complete. Moreover, unlike previous models, we place ourselves in the case of pervasive gene flow among species that may have occurred countless times in the past, as suggested by recent studies. The GBD models are anchored in a new conceptual framework, that we call the genomic view of diversification. Unlike the timetree view, the present framework does not put the emphasis on the species tree (which in our model becomes a network rather than a tree) and assumes that gene trees shape the species tree rather than the opposite.

THE GENOMIC VIEW OF DIVERSIFICATION

Gene flow and the questionable existence of a species genealogy

The biological species concept (BSC, Mayr 1942) defines species as groups of interbreeding populations that are reproductively isolated from other groups. This definition postulates the non-permeability of species boundaries, which is contradicted by the growing body of evidence describing permeable or semi-permeable genomes, even between distantly related taxa. To integrate the possibility of gene flow into the definition of species, Wu (2001) shifted the emphasis from isolation at the level of the whole genome to differential isolation at the gene level. Species are thus defined as differentially adapted groups for which inter-specific gene flow is allowed except for genes involved in differential adaptation (a well-defined form of divergence in which the alternative alleles have opposite fitness effects in the two groups). Because a fraction of the genome may still be exchanged after speciation is complete, a mosaic of gene genealogies is expected between divergent genomes (Wu, 2001). Much evidence supports this prediction with the observation of highly conflicting gene trees, e.g. Darwin's finches (B. R. Grant & Grant, 1998; P. R. Grant & Grant, 1996), sympatric sticklebacks (Rundle, Nagel, Boughman, & Schluter, 2000; Schluter, 1998), Iberian barbels (Gante et

al., 2015), and *Rhagoletis* species (Berlocher, 2000).

Accordingly, the notion of a species genealogy as the binary division of species into new independently evolving lineages in bifurcating phylogenetic trees, appears inappropriate. To avoid this misleading vision of speciation, we here wish to relax the species tree constraint by considering only gene genealogies as real genealogies, thereby laying aside, at least temporarily, the notion of species genealogy. To do so, we do not specify mother-daughter relationships between species, yet we postulate the existence of species at any time, and assume that we can unambiguously follow the genealogies of genes (defined as non-recombining loci, as mentioned above).

The notion of a species genealogy as a binary bifurcating tree is hardly compatible with gene flow, and a direct consequence is to challenge the notion of a unique ancestral species. If all genes ancestral to species S have travelled through the same species in the past, then species S has only one single ancestor species at any time. But because of gene flow, these genes may lie in different species living at a given time in the past, such that species S can have several ancestral species at this time. In other words, several species have contributed to the present-day genome of the species S.

Genomic coadaptation under continuous gene flow

While some genes (e.g., genes involved in divergent adaptation) are hardly exchanged between populations, other genes (e.g., neutral genes unlinked to genes under divergent selection) can be subject to gene flow between different species (Pinho & Hey, 2010; Wu, 2001). Gene flow can persist for long periods of time, with evidence suggesting introgression events occurring over periods lasting up to 20 Myr (Buonerba et al., 2015; Gante et al., 2015; Willis et al., 2012). Over time, genetic differences will accumulate in regions of low recombination and expand via selective sweeps, leading eventually to complete reproductive isolation (Wu, 2001). Because populations differentially accumulate new alleles, their compatibility (hybrid fitness) will be affected. This process has been conceptualized by Bateson, Dobzhanzky and Muller in the so-called Bateson-Dobzhanzky-Muller (BDM) model (Coyne & Orr, 2004; Dobzhansky, 1936; Muller, 1942). This model proposes that genetic incompatibilities, hence called BDM incompatibilities, are characterised by negative epistatic interactions between alleles at two or more genes that have fixed differentially, in each

of the parental populations, by local adaptation or genetic drift. The selective value of hybrids is reduced because the new alleles, divergently selected in each population, are incompatible when carried by the same genome. On the other hand, in the parental populations the co-adapted combinations of alleles have neutral or even beneficial effects (Seehausen et al., 2014; Turelli & Orr, 2000). These incompatibilities have been hypothesized to increase at a rate proportional to the square of time (Orr, 1995). Accordingly, pairs of species will likely exhibit greater genetic incompatibility as a function of time since divergence, *i.e.* be less permeable to gene flow, as has been observed for Iberian barbels (Gante et al., 2015), pea aphids (Peccoud et al., 2009), or salamanders (Pereira et al., 2011). In other words, gene lineages remaining too long isolated within different species decrease their ability to introgress the genome of the other, a property that we name *genomic coadaptation* and which is the consequence of spontaneous mutation.

The gene-based diversification (GBD) models

We propose here a new plastic framework, derived from the genomic view of diversification described above, that acknowledges the importance of gene flow and relaxes the hierarchy between the species tree and gene trees. We built two models, one in forward time that follows the standard view of the main biological processes responsible for diversification under gene flow, and one in backward time using coalescent theory, less computationally intensive, with matching backward parameters (Fig. 2). These models that we named the gene-based diversification (GBD-forward and GBD-backward) models, describe the joint evolution of gene and species lineages, reconciling phylogenomics with our current knowledge of species diversification. The biological mechanisms first, then the corresponding parameters, are detailed thereafter for each model.

The GBD-forward model

The GBD-forward model describes the joint action of four processes affecting the diversification of genomes (see Fig. 2): colonization, mutation, drift and gene flow.

We consider a stochastically varying number of *populations*, all populated with individual genomes. We neglect extinctions and focus on *colonization* events, at which one population seeds a daughter population

founded by one or several of its individuals. Genes independently accumulate *mutations* with time, under the infinite-allele model assumption. Mutations can be fixed or lost due to selection and genetic drift, that we summarize here under the term *drift*.

As a result of mutations and drift, populations differentiate genetically through time, which results in the decrease of gene flow. To model this, we follow what we term the *co-adaptation* between non-homologous genes and assume that *introgression* is governed by the numbers of co-adapted alleles in the receiver and donor populations. Right after colonization, all the genes of the daughter and mother populations carry the same alleles and so are co-adapted. Now an allele having arisen at time t by mutation on some gene is co-adapted only with the alleles carried by its genome at time t. This assumption underlies the well-known model of BDM incompatibilities described previously. Each time a mutation occurs the number of co-adapted genes among populations will decrease, reducing in turn the possibility of genetic exchange between populations.

Two populations that are completely differentiated, in the sense that all pairs of non-homologous alleles sampled from each of them are not co-adapted, can no longer exchange genes and can thus be seen as different species. Because populations are constantly differentiating from each other, we name populations in the prospective point of view (GBD-forward) what will become species only from a retrospective point of view (GBD-backward).

Demographic events are assumed to be much faster than other processes. In the time scale considered here, (1) the fixation of alleles within populations is instantaneous so that all genomes in a population are identical (we thus do not model the co-existence of several different homologous alleles within a population) and (2) a colonization event can be seen as the instantaneous replication of one population into two, actually because of (1), of one genome into two.

Parametrization

At $t = 0$, we consider a single monomorphic population, summarized into a single genome harboring n genes. During the diversification process, the genome of this population (n genes) will be replicated, mutations will be differentially fixed in each population, and the genomes of these populations can be replicated

again. We follow the lineages of these n genes in forward time, assuming a time-continuous Markov chain with 4 events occurring at the following rates.

- **Genetic differentiation** (rate α). At any time t, each gene lineage in each population can acquire a new allele (infinite-allele model) at rate α. By definition, a new allele occurring at gene L on genome G is co-adapted with the allele present at a gene L', for any L' (different of L) of genome G. On the contrary, a mutation arising at gene L of genome G and a mutation arising at gene L' of genome G' are not co-adapted.

- **Colonization** (rate β). At any time t, each population can be replicated at rate β into a new population which will evolve independently in the future. The newborn population is assumed to carry the same genome as carried by the mother population.

- **Genetic drift** (rate γ). Each population undergoes Moran-type births and deaths at rate γ. In this work, we assume γ to be much larger than all other parameters, so that each population is actually monomorphic at all times.

- **Gene flow** (rate δ). At any time t, each gene lineage at locus L on genome G can be replicated and introgress genome G' at rate $\delta(n-1)$, proportional to the number of non-homologous loci in genome G'. If accepted by the target genome G', the replicated lineage replaces its homologous gene lineage (at locus L in G'). The introgression is accepted with a probability equal to the fraction of the $n-1$ non-homologous genes on G' carrying an allele co-adapted with the allele carried by L.

Diversification occurs until a number K of different populations is reached and the whole process is stopped when the K populations are genetically isolated, that is, when no pair of alleles carried by different genomes is co-adapted (i.e., when all probabilities of introgression are equal to 0).

This framework can be made more complex by letting the parameters depend on time, on the gene, or on any prescribed category of genes.

The GBD-backward model

The GBD-backward model is not the exact backward picture of the GBD-forward model but relies on the same idea that genomes in different populations tend to diverge with time until they cannot exchange alleles. The consequence of this fact is that genes sampled in the same genome today will tend to be found in the same population in the past more often than by chance. We model this phenomenon by saying that the ancestral lineages of genes sampled in the same present-day genome are *co-adapted*, and that co-adapted genes are *attracted* towards each other. The GBD-backward model describes the joint action of four processes (see Fig. 2): non-homologous attraction, homologous attraction, coalescence and erosion.

As explained above, in the retrospective point of view (GBD-backward), we name species the populations in which the ancestral gene lineages travel.

Each gene lineage can move from its species to another species. This happens as a result of homologous attraction, non-homologous attraction and erosion. As explained previously, (non-homologous) *co-adapted* genes move into the same species as a result of *non-homologous attraction*, which can be viewed as the backward consequence of *reproductive isolation*. Homologous gene lineages move into the same species as a result of *homologous attraction*, which can be viewed as the backward picture of a *colonization* event, when populations and their genomes have been replicated. Last, any gene lineage can move from its species by *erosion* to an empty species, i.e., a species containing no other gene lineage ancestral to the sample (the term erosion refers to the fact that the block of ancestral lineages lying in the same species loses one element).

When two homologous gene lineages are in the same species they can *coalesce* when finding their common ancestor, that is merge into a single lineage (hence within the same genome).

Note that after coalescence of two homologous lineages, the resulting lineage is now ancestral to at least two genomes and thus co-adapted with all gene lineages ancestral to these genomes. As a consequence of the mere *non-homologous attraction*, going further back in time, all other genes will then move to the same species and further coalesce, until all homologous gene lineages have coalesced.

Equivalently to the *drift* process in forward time, we will assume that *coalescences* are fast, so that in backward time homologous attraction events are immediately followed by coalescence of the two gene lineages.

Parameterization

At $t = 0$, n homologous genes are sampled in each of N distinct species. Retrospectively, the genomes of these N species (each harbouring n genes) will merge progressively into one genome of n genes at some time t in the past. Homologous genes, one by one, will merge (*homologous attraction* and *coalescence*). Merged genes will then attract all the genes of their original genomes (*non-homologous attraction*), until the *coalescence* of all homologous genes. We follow the lineages of these n genes in backward time, assuming a time-continuous Markov chain with 4 events occurring at the following rates.

- **Non-homologous attraction** (rate a). At any time t in the past, as a consequence of **genomic coadaptation**, each gene lineage L escapes from its species S at rate $a(n-1)$ per target species S', proportional to the number of non-homologous loci in the genome G' hosted by S'. It is accepted in S' based on its co-adaptation with G'. If G_0 denotes the genome harboring the descendant lineage of L at time $t = 0$, then all gene lineages harbored by G' that are ancestral to G_0 are said co-adapted with L. Then L is accepted in S' with a probability proportional to the fraction of the $n-1$ non-homologous loci of G' that are co-adapted with it. The parameter a corresponds to the parameter α of the GBD-forward model (genetic differentiation).

- **Homologous attraction** (rate b). At any time t in the past, each gene lineage at rate b per homologous gene lineage, moves to the species harboring this homologous lineage (or in an alternative, more specific version of the model, each gene lineage belonging to some previously prescribed category, like genes contributing to reproductive isolation). This parameter corresponds to the parameter β of the GBD-forward model (colonization).

- **Coalescence** (rate c). At any time t in the past, each pair of homologous genes lying within the same species coalesces at rate c. This parameter corresponds to the parameter γ of the GBD-forward model (genetic drift).

- **Erosion** (rate d). At any time t in the past, each gene lineage escapes from its genome at rate d and enters an empty species (also called ghost species, *i.e.*, harboring no other gene lineage ancestral to

the samples, see Fig. 2). This parameter corresponds to the parameter δ of the GBD-forward model (gene flow). To model the flow of bigger chunks of DNA, we could alternatively assume that instead of one lineage, a given fraction of the lineages of a genome can simultaneously move to an otherwise empty species. We will not consider this possibility in the present work.

We define the number of *ancestral species* of a given genome at time t, as the number of species at time t containing gene lineages ancestral to this genome. Let us briefly expose the expected effects of these parameters on gene trees.

- Large non-homologous attraction (a) values will result in most gene lineages concentrated in one or two ancestral species.
- Large homologous attraction (b) values will result in short waits between speciation events.
- Large coalescence (c) values hinder incomplete lineage sorting.
- Large erosion (d) values will result in a large number of ancestral species per genome.

In this manuscript we wish to explore the impact of gene flow rather than ILS to explain gene tree conflicts, and thus consider a large c value (coalescence rate) so that coalescence events are instantaneous, which is consistent with the large γ value of the forward model. Therefore, only the parameters a, b, and d have an influence on the gene genealogies in the GBD-backward model.

The GBD models were implemented in R (https://www.r-project.org) and evaluated under different sets of parameters. Because the GBD-forward model is computationally prohibitive, while giving comparable qualitative results with the GBD-backward model (see Results), we conducted most of the analyses and the inferences with the GBD-backward model. We provide a first ABC-like inference method by minimizing the difference (Kullback-Leibler divergence) between the distributions of Kendall and Colijn (KC) distances (pairwise distances between gene trees) (Kendall & Colijn, 2016) in empirical vs simulated data. We applied this inference method to two empirical multi-locus data-sets showing complex evolutionary patterns due to gene flow, comprising six morphologically and ecologically distinct species, the Ursinae (a bear subfamily) species (Kutschera et al., 2014) and the *Geospiza* clade (a genus of Darwin's finches) (Farrington, Lawson,

Clark, & Petren, 2014). We estimated in particular 1) the relative amount of gene flow that has shaped each data-set, and 2) the corresponding average number of ancestral species.

MATERIAL AND METHODS

Inference method for the GBD-models

When considering several sampled genomes all containing n genes, a set of n gene trees is obtained for each particular parameter setting and each realization of the model. To characterize a set of gene trees, we employed a multidimensional summary statistic defined as the distribution of pairwise distances between gene trees. Because the GBD-models are time oriented, a tree metric for rooted trees was necessary. Among this class of metrics, we evaluated three metrics: one accounting only for topology, the Robinson–Foulds (RF) metric (Robinson & Foulds, 1981), and two metrics accounting for both branch lengths and topological differences, the Billera-Holmes-Vogtmann (BHV) metric (Billera, Holmes, & Vogtmann, 2001) and the Kendall and Colijn (KC) metric (Kendall & Colijn, 2016).

The RF metric between two trees, *i.e.* the distance between two trees, is computed as the number of bipartitions of the taxon set congruent with one tree but not with the other tree. The BHV metric is based on a view of tree space as a quadrant complex with quadrants sharing faces. Two trees with the same topology lie in the same quadrant, otherwise they lie in two distinct quadrants. At a common edge between two quadrants, the incongruent internal branches between trees have lengths equal to zero. Then a distance can be calculated between two rooted trees as the shortest path across these interconnected quadrants. The KC metric corresponds to the Euclidean distance between two vectors (two by tree). One vector records the number of edges between the most recent common ancestor and each pair of tips and the second records the path length on the considered edges in the first vector (lengths of each tip are also recorded in the second vector). BHV and KC distances do not rely only on the topology but also on branch lengths. The difference in topology is weighted by the branch lengths supporting these topologies, therefore uncertainties causing polytomies (or a branching pattern close to a polytomy) in gene trees will only marginally affect our results.

To compare trees that did not evolve on the same time scale, RF, BHV and KC distances were computed

on re-scaled trees. For each set of gene trees issued from a single simulation or data-set, we rescaled all the trees so that the median of the most recent node depth is 1. After this step, the relative difference in branch lengths remains the same among each set of gene trees.

To estimate model parameters on empirical or test (simulated) trees, we employed the Kullback-Leibler (KL) divergence (package 'FNN' in R) as a distance metric by minimizing this distance between the distributions of pairwise distances (BHV, KC or RF) of empirical, and test trees, with simulated trees. The lower the KL divergence the better is the fit.

Inference method accuracy

To define which distance metric was the more suitable, we tested the accuracy of our inference method, detailed above, using either BHV, KC (with the parameter defining the balance between branch lengths and topology set to 0.5) and RF distances.

Using the GBD-backward model, we built gene trees for 204 parameter combinations (with $N = 6$ and $n = 10$) by varying two parameters, a and b, and fixing $d = 1$ and $c = 200$. The number of time units t was set to 5,000. We performed 85 replicates (75 replicates to build the reference distributions and 10 replicates as a test data-set) under each parameter combination in a grid of ($\frac{1}{a}$, b) with $\frac{1}{a} \in [0.3, 3.5]$, every 0.2, and $b \in [0.01, 0.12]$, every 0.01. For each of the 204 test data-sets, the parameters ($\frac{1}{a}$, b) were inferred by minimizing the KL distance between the pairwise distance distribution of the test trees and the pairwise distance distribution (95% confidence interval) of the reference trees from the grid.

Because KC distances showed a better performance in inferring the model parameters (see supplementary results), we used only this metric for the other analyses of the study.

Comparison of the GBD-models

Next, we aimed to evaluate i) the dynamic of coalescence events between two genomes, called here the coalescence profiles, ii) the number of ancestral species through time for one genome, and iii) the maximal number of gene lineages of one genome located in one (ancestral) species. We performed simulations for $n = 20$ genes, with $\alpha = 0.1$, $\beta = 0.2$, $\delta = 0.06$ and $K = 30$ for the GBD-forward, and $a = 2$, $b =$

0.2, $d = 6$ and $N = 30$ for the GBD-backward model. Additionally, to visually compare the reconstructed genealogies obtained with the GBD-forward and the GBD-backward model we performed simulations for genomes containing $n = 5$ genes, with $\alpha = 0.5$, $\beta = 1$, $\delta = 0.2$ and $K = 30$ for the GBD-forward, and $a = 1$, $b = 0.1$, $d = 2$ and $N = 10$ for GBD-backward model.

Both models gave qualitatively similar results (see Results section). However because the GBD-forward model is computationally prohibitive, all the following analyses were performed with the GBD-backward model. A simulation, with $N = 6$ ($K = 30$ for the GBD-forward to be able to reconstruct genealogies of $N = 6$ genomes), $n = 10$, $a = \alpha = 1$, $b = \beta = 1$, $c = 200$ and $d = \delta = 1$, took about 10 hours for the GBD-forward model and 10 minutes for GBD-backward model (Intel(R) Core(TM) i7-6700 CPU).

A single sampled genome (GBD-backward model)

To evaluate the variation in the number of ancestral species with the intensity of gene flow, we performed simulations for a single sampled genome containing n genes (with $n = 20, 50, 100, 200$), and varied the relative amount of gene flow (erosion rate d) compared to genetic differentiation (*non-homologous attraction* rate a), ratio $\frac{d}{a}$ (with $a = 1$ and $d \in [0.2, 2]$, every 0.2). The simulation was run for 10,000 steps. We sampled the number of ancestral species every 500 steps starting at time $t = 5,000$, and averaged these 11 values for each simulation. For each set of parameters, 5 replicates were performed and averaged.

A model is said to be *sampling consistent* if the same outcome is expected for any k sampled genes independently of the total number n of genes in the genome. To evaluate the validity of this property, we randomly sampled $k = 20$ genes from each genome of $n \geq 20$ genes and computed their average number of ancestral species.

A sample of several genomes (GBD-backward model)

We evaluated the influence of the number n of genes (with $n = 10, 15, 20$), of the number of species N (with $N = 6, 10$), and of the relative amount of gene flow $\frac{d}{a}$ (with $d = 1$ and $\frac{1}{a} = 0.3, 0.5, 0.9, 1.3, 1.7, 2.1, 2.5, 2.9, 3.3$) on gene tree diversity (KC distances) (Fig. 6A). The other parameters were fixed, with $b = 0.05$ and $c = 200$.

For the same values of d and c but with $\frac{1}{a} \in [0.3, 3.5]$, every 0.2, and for $n = 10$, $N = 6$, we also evaluated the influence of the *homologous attraction* rate b (with $b = 0.01$, 0.02, 0.05, 0.12) on gene tree diversity (KC distances) (Fig. 6B).

Inference from empirical data-sets

Empirical data-sets

To evaluate if the GBD-backward model correctly reproduces the signal left by gene flow in gene trees we compared empirical gene trees (shaped under gene flow) with simulated gene trees under the GBD-backward model and under the MSC model. The adequacy between the simulated trees and empirical gene trees was estimated by comparing the distributions of pairwise gene tree distances of simulated vs empirical data-sets. The empirical clades have been chosen for their moderate phylogenetic depth, good sampling coverage and known conflicting gene trees. The bear data-set comprised 14 autosomal introns for 6 bear species (*Helarctos malayanus*, *Melursus ursinus*, *Ursus americanus*, *U. arctos*, *U. maritimus*, and *U. thibetanus*) and 2 outgroups (*Ailuropoda melanoleuca* and *Tremarctos ornatus*) (Kutschera et al., 2014). The sequences were downloaded from GenBank (supplementary Table S1). As done by Kutschera et al. (2014), all variation within and among individuals was collapsed into one single 50% majority-rule-consensus sequence for each of the 8 species. The phylogenetic trees were built with the program BEAST v. 1.8.3. (Drummond, Suchard, Xie, & Rambaut, 2012), with the parameters used by Kutschera et al. (2014): Yule prior to model the branching process, strict clock, a normal prior on substitution rates (0.001 ± 0.001) (mean $\pm$ SD), minimum age of 11.6 My for the divergence of *A. melanoleuca* from other bears (exponential prior: mean $= 0.5$; offset $= 11.6$), and 10 million generations with sampling every 1000 generations. The models of DNA evolution were estimated by modeltest (function 'modelTest', package 'phangorn' in R (Schliep, 2011)) (supplementary Table S2). The monophyly of the ingroup and the topology among the outgroups were constrained according to the topology depicted by Kutschera et al. (2014).

The second data-set comprised 7 nuclear markers for 6 finch species (*Geospiza conirostris*, *G. fortis*, *G. fulginosa*, *G. magnirostris*, *G.scandens*, and *G. septentrionalis*) and 2 outgroups (*Camarhynchus psittacula* and *Platyspiza crassirostris*) (Farrington et al., 2014). The sequences were downloaded from GenBank

(supplementary Table S3). The phylogenetic trees were built with the program BEAST v. 1.8.3. (Drummond et al., 2012) with the parameters used by (Farrington et al., 2014): coalescent constant size prior to model the branching process, strict clock, substitution rate equal to 1, specific models of DNA evolution defined by the authors (supplementary Table S2), and 10 million generations with sampling every 1000 generations. The monophyly of the ingroup and the topology among the outgroups were constrained according to the topology depicted by Farrington et al. (2014).

Estimation of parameters under the multi-species coalescent (MSC) model

We optimized the MSC model for $N = 6$ species by varying two parameters, the speciation rate λ and the extinction rate μ, and fixing the coalescence rate to 1. Birth-death trees of 6 tips (function 'sim.bdtree', package 'geiger' in R) were simulated in a grid of (λ, $\mu = m\lambda$) with $\lambda \in [0.02, 0.34]$, every 0.02, and $m \in [0.1, 0.65]$, every 0.05. Because we simulated small trees (6 tips), the degree of variation between trees simulated with the same parameters was high. Therefore for each value of (λ, μ) we randomly selected 75 species trees for which the crown age did not differ by more than 2.5% from the expected crown age. Next, we simulated 10 gene genealogies for each species tree (coalescence rate fixed to 1).

If the diversification rate (speciation rate minus extinction rate) is low, all the homologous genes will coalesce before the next node in the species tree, so that all the gene trees will have the same topology. On the contrary, if the diversification rate is too fast, some homologous genes will not have time to coalesce before the next node of the species tree, resulting in incongruent gene trees due to the randomness of coalescences (ILS).

Estimation of parameters under the gene-based diversification (GBD-backward) model

Equivalently, we optimized the GBD-backward model for $N = 6$ by varying two parameters, here a and b, and fixing $d = 1$ and $c = 200$ (recall c is given a sufficiently large value that coalescences are instantaneous). Since increasing n has no effect on KC distances (see results and Fig. 6), we simulated genomes with $n = 10$ genes. Each simulation was conducted until the coalescence of all homologous genes. We performed 75 replicates under each parameter combination in a grid of ($\frac{1}{a}$, b) with $\frac{1}{a} \in [0.3, 3.5]$,

every 0.2, and $b \in [0.01, 0.12]$, every 0.01.

For both models (MSC and GBD-backward) we employed the Kullback-Leibler (KL) divergence (package 'FNN' in R) as a distance metric to find the best set of parameters by minimizing this distance between the distributions of KC pairwise distances of empirical and simulated trees. The lower the KL divergence is the better is the fit.

RESULTS

Inference method accuracy

Using simulated data-sets we showed that our inference method, comparison of the distribution of KC pairwise distances among a set of trees, was able to give reliable estimates of simulated parameters despite its simplicity (see supplementary results figure S1). This simple inference method is sufficient to estimate the parameters of the model having supposedly shaped the gene trees of the data set. More subtle methods will be developed in the future to account for more complex features, such as differential gene flow depending on putative gene categories, and to infer the very history of the embedding of gene lineages into species.

Comparison of the GBD-models

Even if the two models, GBD-forward and GBD-backward, are not rigorously the inverted image of one another, they showed a qualitatively similar pattern in gene genealogies and in the distribution of gene lineages among species through time (Figs. 3 and 4).

The coalescence profiles (*i.e.* the cumulative number of coalescence events through time) were similar in shape for both models (Figs. 3 and S2) with a rapid increase followed by a plateau until the last coalescence event. However, contrary to the GBD-backward model, the GBD-forward model can generate simultaneous coalescences (as in our simulation, Fig. 3A). If gene flow is too weak relatively to the other parameters, some of the simultaneous coalescences modelled at a colonization event will remain together (*i.e.* happening at the same time t) and appear as a multifurcation in reconstructed trees.

The number of ancestral species through time was also comparable between the two models (Fig. 3B and E). Going backward in time, after a first increase in the number of ancestral species, gene lineages tend

to be brought back together in fewer ancestral species whether it is due to non-homologous attraction in the GBD-backward model or because genes were co-adapted at this time in the GBD-forward model. In the GBD-backward model, once at least a pair of homologous genes have coalesced, the two genomes will be attracted toward this coalesced gene lineage increasing the probability for two other homologous genes to be in the same species and thus to coalesce. The attraction will be even stronger when many genes have coalesced, explaining the latter decrease in the number of ancestral species (Fig. 3E).

In the GBD-forward model, going forward in time, the number of ancestral species increases and suddenly decreases. Just after a *colonization* event, the new species can easily exchange genes with related species. Gene lineages will then acquire new mutations, isolating each genome from the others, explaining the rapid decrease in the number of ancestral species toward the present (Fig. 3B). Note that in both cases, going backward in time, once all the homologous genes have coalesced they follow the exact same path through species (the blue and pink lines are confounded in the Fig. 3B-F).

Moreover, because they are co-adapted, genes sampled in the same species at present time should have spent time together in the same species more often than by chance in the past. This property was indeed observed in both models, with genes sampled at present time frequently found together in the same species in the past (see Figs. 3C, F and 4), where at least five gene lineages (over 20) lie in the same ancestral species at all times).

A single sampled genome (GBD-backward)

With $N = 1$ sampled genome containing n genes, we let $A(t) = (A_1(t), \ldots, A_n(t))$ denote the sorting of genes into ancestral species t units of time before the present. More precisely, $A_k(t)$ denotes the number of ancestral species containing k gene lineages, so that $n = \sum_{k=1}^{n} kA_k(t)$ and $S(t) = \sum_{k=1}^{n} A_k(t)$ is the total number of species at t ancestral to the sampled genome. For each $\varepsilon \in (0, 1]$, we will also be interested in the number $S_\varepsilon(t) = \sum_{k=[\varepsilon n]}^{n} A_k(t)$ of ancestral species containing at least a fraction ε of the genome (with $[x]$ denoting the smallest integer larger than x). All stationary quantities will be denoted by the same symbols, replacing t with ∞.

We will call a *block* at (backward) time t a (maximal) set of gene lineages that lie in the same species

at time t. The transition rates can be specified as follows in terms of the configuration of gene lineages into blocks (*i.e.*, ancestral species). For each pair of blocks containing (j, k) lineages, non-homologous attraction occurs at rate ajk and results in the configuration $(j-1, k+1)$. For each block containing j lineages, erosion occurs at rate dj and results in the block losing one lineage; simultaneously a new block containing 1 single lineage is created. These are exactly the same rates as in the well-known Moran model with mutation under the infinite-allele model (Moran, 1958), replacing 'block' with 'allele', 'attraction' by 'resampling' (simultaneous birth from one of the j carriers of a given allele and death of one of the k carriers of another given allele) and 'erosion' with 'mutation' (mutation appearing in one of the j carriers of a given allele into a new allele never existing before). For this Moran model,

- the total population size is n;
- at rate a for each oriented pair of individuals independently, the first individual of the pair gives birth to a copy of herself and the second individual of the pair is simultaneously killed;
- mutation occurs at rate d independently in each individual lineage.

As a consequence, $A(t)$ has the same distribution as the allele frequency spectrum in the Moran model with total population size n, resampling rate a and mutation rate d, starting at time $t = 0$ from a population of clonal individuals (one single block). In particular, the distribution of $A(\infty)$ is the stationary distribution of the allele frequency spectrum, which is known to be given by Ewens' sampling formula with scaled mutation rate d/a (Durrett, 2008; Ewens, 1972; Ewens & Tavaré, 2006). Expectations of this distribution are:

$$\mathbb{E}(A_k(\infty)) = \frac{d}{d + a(k-1)},$$

so that

$$\mathbb{E}(S(\infty)) = \sum_{k=1}^{n} \frac{d}{d + a(k-1)} \tag{1}$$

and

$$\mathbb{E}(S_\varepsilon(\infty)) = \sum_{k=[\varepsilon n]}^{n} \frac{d}{d + a(k-1)}. \tag{2}$$

In particular, as $n \to \infty$,

$$\mathbb{E}(S(\infty)) \sim \frac{d}{a} \ln(n) \quad \text{and} \quad \mathbb{E}(S_\varepsilon(\infty)) \sim \frac{d}{a} \ln(1/\varepsilon).$$

At stationarity, and particularly for large values of $\frac{d}{a}$, the mean number of ancestral species $S(\infty)$ obtained from simulations was equal to the mathematical prediction (figure 5A). In particular, the mean number of ancestral species at stationarity increases with $\frac{d}{a}$.

An additional key feature of this model is *sampling consistency*. In words, the history of a sample of k genes taken from a genome of n genes does not depend on n. This property can again be deduced from the representation of our model in terms of the better known Moran model. Indeed, the dynamics of a sample of k individuals in the Moran model does not depend on the population size, as can be seen from the so-called lookdown construction (due to P. Donnelly and T. Kurtz and clearly exposed by Etheridge 2011). The simulations performed with k genes randomly sampled from each genome of n genes, are in agreement with this claim of sampling consistency: the number of ancestral species at stationarity $\mathbb{E}(S_\varepsilon(\infty))$ is independent of the number of genes n (Fig. 5B).

A sample of several genomes (GBD-backward)

Using simulations, we evaluated the GBD-backward model for several sampled genomes ($N > 1$) under several combinations of parameters. As expected, gene tree variation, measured by KC distances, increased with $\frac{d}{a}$, *i.e.* the relative amount of gene flow, and with the number of species N. Conversely our results showed that the number of genes n had no effect on distances (Fig. 6A). This last result, the lack of influence of n on gene tree variation, is of particular interest, because one usually has only access to a fraction of a genome. It shows that regardless of the number of genes sampled, the resulting gene tree variation will remain the same as long as gene trees have been shaped by processes with similar parameter values.

Our results also showed that as the *homologous attraction* rate b decreases, and for the same value of $\frac{d}{a}$, gene trees were more similar (lower KC distances) (Fig. 6B). When a long period of time elapses between two homologous attraction events (low b), all the genes belonging to the two genomes that have started to coalesce, have enough time to be attracted toward the same ancestral species, and thus coalesce before the next homologous attraction event, in spite of gene flow.

The GBD-backward model correctly captures the signal left by gene flow in empirical data-sets

To estimate model parameters, we minimized the Kullback-Leibler (KL) divergence between the distributions of KC pairwise distances of empirical and simulated trees (Fig. 7). Under the multi-species coalescent (MSC) model, the most likely parameter values were $\mu = 0.2 \times \lambda$ and $\lambda = 0.08$ (KL divergence $= 0.39$) for the bears and $\mu = 0.60 \times \lambda$ and $\lambda = 0.20$ (KL divergence $= 0.19$) for the finches. Under the gene-based diversification (GBD-backward) model, the most likely parameter values were $b = 0.03$ and $\frac{d}{a} = 2.1$ (KL divergence $= 0.24$) for the bears and $b = 0.08$ and $\frac{d}{a} = 1.7$ (KL divergence $= 0.01$) for the finches (Fig. 7). We noted longer tailed distributions for the distances between trees modeled under the MSC model than for the empirical data-sets (Fig. 8). This skewed distribution obtained with the MSC model explains why we did not detect a sharp peak in the optimization landscape for the MSC model (Fig. 7).

However, for both data-sets, the selected simulated distribution (GBD model) and the empirical distribution do not match perfectly. This difference, an excess of intermediate distances and a lack of large distances among empirical trees could reflect the exclusion of too incongruent genes when the data-sets were built, and/or the presence of functionally or physically linked genes (preferentially found together in ancestral species).

Comparing the parameters λ and μ to b and $\frac{d}{a}$ is not straightforward as the two models, MSC and GBD-backward, are built under different assumptions. However in both cases, the parameters influence the diversity among trees (shape of the distribution of KC pairwise distances). A greater variation among trees is expected with increasing λ and decreasing μ, and with increasing $\frac{d}{a}$ and b, allowing us to explore the parameter landscape to find the setting that minimizes the distance between simulations and empirical data-sets for each model.

Given our results and the mathematical predictions, the time-averaged number $S_\varepsilon(\infty)$ of ancestral species to the sampled genome containing at least 10% of the genome ($\varepsilon = 0.1$) when $n \to \infty$ is 4.8 for the bear data-set and 3.9 for the finch data-set.

DISCUSSION

Within species, gene flow allows the maintenance of species cohesion in the face of genetic differentiation (Morjan & Rieseberg, 2004; Slatkin, 1987), preventing genetic isolation of populations and the subsequent emergence of reproductive barriers leading to speciation (Coyne & Orr, 2004). Among species, the existence of gene flow challenges the notion of a species genealogy as well as the current concepts of species. Indeed, if gene flow is as pervasive as recent empirical studies suggest (Clark & Messer, 2015; Cui et al., 2013; Gallus et al., 2015; Jónsson et al., 2014), the genealogical history of species should be represented as a phylogenetic network encompassing the mosaic of gene genealogies. Similarly, it seems very conservative to delineate species based on the widely used biological species concept (reproductive isolation) (Mayr, 1942), or phylogenetic species concept (reciprocal monophyly) (Papadopoulou et al., 2008). Because of the ubiquity of gene flow, which can persist for several millions of years after the lineages have started to diverge (*i.e.*, onset of speciation) (Bolnick, Near, & Noor, 2005; Mallet, 2005), species should be rather defined by their capacity to coexist without fusion in spite of gene flow (Mallet, 2008; Samadi & Barberousse, 2006).

The simplified view of diversification, consisting in representing lineages splitting instantaneously into divergent lineages with no interaction (gene exchange) after the split, has been preventing evolutionary biologists from fully apprehending diversification at the genomic level and from correctly interpreting discrepancies between gene histories. Indeed, conflicting gene trees make the interpretation of their evolutionary history difficult. However, we argue that phylogenetic incongruence among gene trees should not be considered as a nuisance, but rather as a meaningful biological signal revealing some features of the dynamics of genetic differentiation and of gene flow through time and across clades. Current phylogenetic methods rely on the assumption that gene trees are constrained within the species tree, and that gene flow occurs infrequently between species. For many data-sets such as sequence alignments of genomes sampled from young clades, such methods could lead to an evolutionary misinterpretation of gene trees, and in the worst case to species trees with high node support while the gene trees had very different evolutionary histories (Long & Kubatko, 2018). These observations urge for a change of paradigm, where gene flow is fully part of the diversification model. To consider the ubiquity of gene flow across the Tree of Life and its broad effect on genomes described by many recent studies, we have developed a new framework focusing on gene

genealogies and relaxing the constraints inherent to the MSC paradigm. This framework is implemented in a mathematical model that we named the gene-based diversification (GBD-forward) model. We have also developed a complementary version of this model, the GBD-backward model, speeding up the simulations thanks to a coalescent approach.

The GBD-backward model

Under the GBD-backward model, gene genealogies are governed by four parameters: non-homologous attraction rate a , homologous attraction rate b, coalescence rate a, and erosion rate d (Fig. 2).

Non-homologous attraction models genetic differentiation resulting in reproductive isolation. The slower genes accumulate mutations and differentiate, the more time can be spent by gene lineages in different species. Hence when genomes differentiate slowly, the rate of non-homologous attraction is low. *Homologous attraction* corresponds to finding the most recent common ancestor of the two species at the genomic level. The time spent between homologous attraction events depends crucially on the (phylogenetic distance of the) species sampled at the present. *Coalescence* is in direct correspondence with genetic drift in the GBD-forward model and *erosion* with gene flow.

Each of these parameters influences differently the resulting tree variation, *i.e.* the distribution of the KC distances among trees, that we used here as a summary statistic. Instead of focusing on the main phylogenetic signal alone as done by the current phylogenetic methods, the GBD-backward model makes use of the whole signal generated by all gene trees.

Higher amount of gene flow (large d values) and reduced time to untangle gene genealogies before the connection of two other genomes (large b values) increase variation among trees. Conversely, when homologous genes coalesce faster (large c values) and genes are recalled faster toward the species harboring the other genes of their genome (large a values) gene trees are expected to be more similar.

After evaluating this model under various sets of parameters, we applied it to analyze two empirical multi-locus data-sets for which gene tree conflicts obscure the evolutionary history.

Gene flow among bears and among finches

In many cases, such as among bears and finches, gene flow is frequent and complicates the relationships between species, challenging the notion of a unique species tree. A strictly bifurcating lineage-based model will not adequately reflect those complex evolutionary patterns. On the contrary, models developed under the *genomic view of diversification* framework, *i.e.* relaxing species boundaries and accounting for gene flow, will better reproduce the complex history of gene genealogies under pervasive gene flow. Note that we considered a simple scenario with no ILS and statistically exchangeable genes resulting in a model with only three parameters, but given the simplicity and the flexibility of our model, many extensions may be considered to address scenarios that could not have been considered previously, opening up new perspectives in the study of speciation and macro-evolution.

Our results showed support for the hypothesis that gene flow has shaped the gene trees of bears and finches (Fig. 8). For the bear data-set, we found that each species had on average in the past about 4.8 ancestral species carrying at least 10% of its present genome (Equation (2)). This result is in line with previous studies reporting gene flow between pairs of bear species (Cahill et al., 2013; Hailer et al., 2012; Kutschera et al., 2014; S. Liu et al., 2014; Miller et al., 2012). Moreover, a recent phylogenomic study (869 Mb divided into 18,621 genome fragments) confirmed the existence of gene flow between sister species as well as between more phylogenetically distant species (Kumar et al., 2017). The authors used the D-statistic (gene flow between sister species) and D_{FOIL}- statistic (gene flow among ancestral lineages, Pease and Hahn 2015) to detect gene flow among the 6 bear species. Using their results, for each pair of species ij among the N species, we determined if the species j has contributed ($g_{ij} = 1$) or not ($g_{ij} = 0$) to the genome of the species i (with $g_{ii} = 1$), and calculated the average number of ancestral species $\overline{S}$ as follows:

$$\overline{S} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} g_{ij}. \tag{3}$$

Using this equation on the results of the phylogenomic study (Kumar et al., 2017), we found on average 5.3 ancestral species for each of the Ursinae bears, close to the estimate of 4.8 obtained with the GBD-backward model.

We detected lower gene flow among finches than among bears. Each finch species had on average in the past 3.9 ancestral species (for the subsample of gene trees analyzed here), which is also consistent with the extensive evidence that many species hybridize on several islands (Freeland & Boag, 1999; P. R. Grant

& Grant, 1997; P. R. Grant, Grant, & Petren, 2005; Sato et al., 1999). Because of gene flow very little genetic structure was detected by a Bayesian population structure analysis, (only 3 genetic populations among the 6 *Geospiza* species, see Farrington et al. 2014). Each of the 2 species *G. magnirostris* and *G.scandens* was mostly characterized by a single genetic population, therefore had about 1 ancestral species each. Conversely 4 *Geospiza* species shared the same genetic population, suggesting 4 ancestral species for each of these 4 species. Taking together these results roughly indicate that each of the 6 *Geospiza* species had in average 3 ancestral species, which is slightly lower than the GBD-backward estimate of 3.9.

Perspectives

Phylogenetic models and methods inferring macro-evolutionary history, such as speciation and extinction rates, trait evolution or ancestral characters, have become increasingly complex (Morlon, 2014; Pyron & Burbrink, 2013; Stadler, 2013b). Yet, the raw material used by these methods is often reduced to the species tree, which can be viewed as a summary statistic of the information contained in the genome. We argue here that a valuable amount of additional signal, not accessible in species trees, is contained in gene trees, and is directly informative about the diversification process. Indeed, because genetic differentiation and gene flow impact each gene differently, genes may have experienced very different evolutionary trajectories.

In order to make use of the entire information conveyed by gene trees, we have proposed here a new approach to study diversification, the genomic view of diversification, under which gene trees shape the species tree rather than the opposite. This approach aims at better depicting the intricate evolutionary history of species and genomes. We hope that this view of diversification will contribute to pave the way for future developments in the perspective of inferring diversification processes directly from genomes rather than from their summary into one single species tree. One of the challenges in this direction will be to propose finer inference methods than the simple, but reasonably satisfactory, method used here, based on a single multidimensional summary statistic, the distribution of pairwise KC distances between gene trees.

SUPPLEMENTARY MATERIAL

The code for the models is available as Supplementary Material.

ACKNOWLEDGMENTS

The authors thank the *Center for Interdisciplinary Research in Biology* (Collège de France, CNRS) for funding. JM is funded by LabEx MemoLife, project *Genomics of Diversification*. The authors also thank the INRA MIGALE bioinformatics platform (http://migale.jouy.inra.fr) for providing computational resources.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., ... Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology*, *26*(2), 229–246. doi: 10.1111/j.1420-9101.2012.02599.x

Berlocher, S. H. (2000). Radiation and divergence in the Rhagoletis pomonella species group: inferences from allozymes. *Evolution*, *54*(2), 543–557.

Billera, L. J., Holmes, S. P., & Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, *27*(4), 733–767.

Bolnick, D. I., Near, T. J., & Noor, M. (2005). Tempo of hybrid inviability in centrarchid fishes (teleostei: centrarchidae). *Evolution*, *59*(8), 1754–1767. doi: 10.1554/04-563.1

Buonerba, L., Zaccara, S., Delmastro, G. B., Lorenzoni, M., Salzburger, W., & Gante, H. F. (2015). Intrinsic and extrinsic factors act at different spatial and temporal scales to shape population structure, distribution and speciation in Italian Barbus (Osteichthyes: Cyprinidae). *Molecular Phylogenetics and Evolution*, *89*, 115–129. doi: 10.1016/j.ympev.2015.03.024

Cahill, J. A., Green, R. E., Fulton, T. L., Stiller, M., Jay, F., Ovsyanikov, N., ... Slatkin, M. (2013). Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. *PLoS genetics*, *9*(3), e1003345.

Clark, A. G., & Messer, P. W. (2015). Conundrum of jumbled mosquito genomes. *Science*, *347*(6217), 27–28.

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sinauer Associates, Sunderland, MA.

Cui, R., Schumer, M., Kruesi, K., Walter, R., Andolfatto, P., & Rosenthal, G. G. (2013). Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes. *Evolution*, *67*(8), 2166–2179.

De Busschere, C., Hendrickx, F., Van Belleghem, S. M., Backeljau, T., Lens, L., & Baert, L. (2010). Parallel habitat specialization within the wolf spider genus Hogna from the Galápagos. *Molecular ecology*, *19*(18), 4029–4045.

Dobzhansky, T. H. (1936). Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. *Genetics*, *21*(2), 113.

Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution*, *29*(8), 1969–1973.

Durrett, R. (2008). *Probability Models for DNA Sequence Evolution*. Springer. (Google-Books-ID: o4_bMHy7jFoC)

Etheridge, A. (2011). *Some Mathematical Models from Population Genetics: École D'Été de Probabilités de Saint-Flour XXXIX-2009*. Springer Science & Business Media. (Google-Books-ID: miI9tdPCFdUC)

Etienne, R. S., Morlon, H., & Lambert, A. (2014). Estimating the duration of speciation from phylogenies. *Evolution*, *68*(8), 2430–2440.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, *3*(1), 87–112. doi: 10.1016/0040-5809(72)90035-4

Ewens, W. J., & Tavaré, S. (2006). Ewens Sampling Formula. In *Encyclopedia of Statistical Sciences.* American Cancer Society. doi: 10.1002/0471667196.ess0638.pub2

Farrington, H. L., Lawson, L. P., Clark, C. M., & Petren, K. (2014). The evolutionary history of Darwin's finches: speciation, gene flow, and introgression in a fragmented landscape. *Evolution*, *68*(10), 2932–2944.

Fontenot, B. E., Makowsky, R., & Chippindale, P. T. (2011). Nuclear–mitochondrial discordance and gene flow in a recent radiation of toads. *Molecular Phylogenetics and Evolution*, *59*(1), 66–80. doi: 10.1016/j.ympev.2010.12.018

Freeland, J. R., & Boag, P. T. (1999). The mitochondrial and nuclear genetic homogeneity of the phenotypically diverse Darwin's ground finches. *Evolution; International Journal of Organic Evolution*, *53*(5),

1553–1563. doi: 10.1111/j.1558-5646.1999.tb05418.x

Gallus, S., Janke, A., Kumar, V., & Nilsson, M. A. (2015). Disentangling the relationship of the Australian marsupial orders using retrotransposon and evolutionary network analyses. *Genome biology and evolution*, *7*(4), 985–992.

Gante, H. F., Collares-Pereira, M. J., & Coelho, M. M. (2004). Introgressive hybridisation between two Iberian Chondrostoma species (Teleostei, Cyprinidae) revisited: new evidence from morphology, mitochondrial DNA, allozymes and NOR-phenotypes. *Folia Zoologica*, *53*(4), 423.

Gante, H. F., Doadrio, I., Alves, M. J., & Dowling, T. E. (2015). Semi-permeable species boundaries in Iberian barbels (Barbus and Luciobarbus, Cyprinidae). *BMC evolutionary biology*, *15*(1), 111.

Gante, H. F., Santos, C. D., & Alves, M. J. (2010). Phylogenetic relationships of the newly described species Chondrostoma olisiponensis (Teleostei: Cyprinidae). *Journal of Fish Biology*, *76*(4), 965–974. doi: 10.1111/j.1095-8649.2010.02536.x

Grant, B. R., & Grant, P. R. (1998). Hybridization and speciation in Darwin's finches: the role of sexual imprinting on a culturally transmitted trait. *Endless forms: species and speciation*, 404–422.

Grant, P. R., & Grant, B. R. (1996). Speciation and hybridization in island birds. *Phil. Trans. R. Soc. Lond. B*, *351*(1341), 765–772.

Grant, P. R., & Grant, B. R. (1997). Hybridization, Sexual Imprinting, and Mate Choice. *The American Naturalist*, *149*(1), 1–28.

Grant, P. R., Grant, B. R., & Petren, K. (2005). Hybridization in the recent past. *The American Naturalist*, *166*(1), 56–67. doi: 10.1086/430331

Hailer, F., Kutschera, V. E., Hallström, B. M., Klassert, D., Fain, S. R., Leonard, J. A., . . . Janke, A. (2012). Nuclear Genomic Sequences Reveal that Polar Bears Are an Old and Distinct Bear Lineage. *Science*, *336*(6079), 344–347. doi: 10.1126/science.1216424

Harrison, R. G., & Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. *Journal of Heredity*, *105*(S1), 795–809. doi: 10.1093/jhered/esu033

Hedges, S. B., & Kumar, S. (2009). *The Timetree of Life*. OUP Oxford. (Google-Books-ID: 9rt1c1hl49MC)

Heled, J., & Drummond, A. J. (2010). Bayesian inference of species trees from multilocus data. *Molecular*

Biology and Evolution, *27*(3), 570–580. doi: 10.1093/molbev/msp274

Jónsson, H., Schubert, M., Seguin-Orlando, A., Ginolhac, A., Petersen, L., Fumagalli, M., ... Orlando, L. (2014). Speciation with gene flow in equids despite extensive chromosomal plasticity. *Proceedings of the National Academy of Sciences*, *111*(52), 18655–18660. doi: 10.1073/pnas.1412627111

Kendall, M., & Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular biology and evolution*, *33*(10), 2735–2743.

Kishino, H., Thorne, J. L., & Bruno, W. J. (2001). Performance of a Divergence Time Estimation Method under a Probabilistic Model of Rate Evolution. *Molecular Biology and Evolution*, *18*(3), 352–361. doi: 10.1093/oxfordjournals.molbev.a003811

Knowles, L. L., & Kubatko, L. S. (2011). *Estimating species trees: practical and theoretical aspects*. John Wiley and Sons.

Kubatko, L. S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Systematic Biology*, *58*(5), 478–488.

Kumar, V., Lammers, F., Bidon, T., Pfenninger, M., Kolter, L., Nilsson, M. A., & Janke, A. (2017). The evolutionary history of bears is characterized by gene flow across species. *Scientific Reports*, *7*, 46487. doi: 10.1038/srep46487

Kutschera, V. E., Bidon, T., Hailer, F., Rodi, J. L., Fain, S. R., & Janke, A. (2014). Bears in a forest of gene trees: phylogenetic inference is complicated by incomplete lineage sorting and gene flow. *Molecular Biology and Evolution*, *31*(8), 2004–2017.

Lambert, A., Morlon, H., & Etienne, R. S. (2015). The reconstructed tree in the lineage-based model of protracted speciation. *Journal of mathematical biology*, *70*(1-2), 367–397.

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. doi: 10.1186/1471-2148-10-302

Liu, S., Lorenzen, E., Fumagalli, M., Li, B., Harris, K., Xiong, Z., ... Wang, J. (2014). Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears. *Cell*, *157*(4), 785–794. doi: 10.1016/j.cell.2014.03.054

Long, C., & Kubatko, L. (2018). The effect of gene flow on coalescent-based species-tree inference. *gener-*

ations, *1*, 2N.

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, *46*(3), 523–536. doi: 10.1093/sysbio/46.3.523

Mallet, J. (2005). Hybridization as an invasion of the genome. *Trends in ecology & evolution*, *20*(5), 229–237.

Mallet, J. (2008). Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1506), 2971–2986. doi: 10.1098/rstb.2008.0081

Mallet, J., Besansky, N., & Hahn, M. W. (2016). How reticulated are species? *BioEssays*, *38*(2), 140–149. doi: 10.1002/bies.201500149

Mayr, E. (1942). *Systematics and the origin of species, from the viewpoint of a zoologist*. Harvard University Press.

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., ... Wittekindt, N. E. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences*, *109*(36), E2382–E2390.

Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., & Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, *30*(17), i541–i548. doi: 10.1093/bioinformatics/btu462

Moran, P. A. P. (1958). Random processes in genetics. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 54, pp. 60–71). Cambridge University Press.

Morjan, C. L., & Rieseberg, L. H. (2004). How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular ecology*, *13*(6), 1341–1356.

Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecology Letters*, *17*(4), 508–525. doi: 10.1111/ele.12251

Muller, H. J. (1942). Recessive genes causing interspecific sterility and other disharmonies between Drosophila melanogaster and simulans. *Genetics*, *27*, 157.

Nadeau, N. J., Martin, S. H., Kozak, K. M., Salazar, C., Dasmahapatra, K. K., Davey, J. W., ... Jiggins, C. D.

(2013). Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Molecular Ecology*, *22*(3), 814–826. doi: 10.1111/j.1365-294X.2012.05730.x

Orr, H. A. (1995). The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics*, *139*(4), 1805–1813.

Papadopoulou, A., Bergsten, J., Fujisawa, T., Monaghan, M. T., Barraclough, T. G., & Vogler, A. P. (2008). Speciation and DNA barcodes: testing the effects of dispersal on the formation of discrete sequence clusters. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1506), 2987–2996.

Pease, J. B., & Hahn, M. W. (2015). Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology*, *64*(4), 651–662. doi: 10.1093/sysbio/syv023

Peccoud, J., Ollivier, A., Plantegenest, M., & Simon, J.-C. (2009). A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proceedings of the National Academy of Sciences*, *106*(18), 7495–7500. doi: 10.1073/pnas.0811117106

Pereira, R. J., Monahan, W. B., & Wake, D. B. (2011). Predictors for reproductive isolation in a ring species complex following genetic and ecological divergence. *BMC Evolutionary Biology*, *11*, 194. doi: 10.1186/1471-2148-11-194

Pinho, C., & Hey, J. (2010). Divergence with Gene Flow: Models and Data. *Annual Review of Ecology, Evolution, and Systematics*, *41*(1), 215–230. doi: 10.1146/annurev-ecolsys-102209-144644

Pyron, R. A., & Burbrink, F. T. (2013). Phylogenetic estimates of speciation and extinction rates for testing ecological and evolutionary hypotheses. *Trends in Ecology & Evolution*, *28*(12), 729–736. doi: 10.1016/j.tree.2013.09.007

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, *53*(1-2), 131–147.

Rosindell, J., Cornell, S. J., Hubbell, S. P., & Etienne, R. S. (2010). Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, *13*(6), 716–727.

Rundle, H. D., Nagel, L., Boughman, J. W., & Schluter, D. (2000). Natural selection and parallel speciation in sympatric sticklebacks. *Science*, *287*(5451), 306–308.

Samadi, S., & Barberousse, A. (2006). The tree, the network, and the species. *Biological Journal of the Linnean Society*, *89*(3), 509–521.

Sato, A., O'hUigin, C., Figueroa, F., Grant, P. R., Grant, B. R., Tichy, H., & Klein, J. (1999). Phylogeny of Darwin's finches as revealed by mtDNA sequences. *Proceedings of the National Academy of Sciences*, *96*(9), 5101–5106. doi: 10.1073/pnas.96.9.5101

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592.

Schluter, D. (1998). Ecological causes of speciation. *Endless forms: species and speciation*, 114–129.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., . . . Saetre, G.-P. (2014). Genomics and the origin of species. *Nature Reviews Genetics*, *15*(3), 176–193.

Slatkin, M. (1987). Gene flow and the geographic structure of natural populations. *Science(Washington)*, *236*(4803), 787–792.

Solís-Lemus, C., Yang, M., & Ané, C. (2016). Inconsistency of species tree methods under gene flow. *Systematic Biology*, *65*(5), 843–851. doi: 10.1093/sysbio/syw030

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, *16*(8), 472–482. doi: 10.1038/nrg3962

Sousa-Santos, C., Gante, H. F., Robalo, J., Cunha, P. P., Martins, A., Arruda, M., . . . Almada, V. (2014). Evolutionary history and population genetics of a cyprinid fish (<Emphasis Type="Italic">Iberochondrostoma olisiponensis</Emphasis>) endangered by introgression from a more abundant relative. *Conservation Genetics*, *15*(3), 665–677. doi: 10.1007/s10592-014-0568-1

Stadler, T. (2013a). Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, *26*(6), 1203–1219. doi: 10.1111/jeb.12139

Stadler, T. (2013b). Recovering speciation and extinction dynamics based on phylogenies. *Journal of Evolutionary Biology*, *26*(6), 1203–1219. doi: 10.1111/jeb.12139

Tamura, K., Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A., & Kumar, S. (2012). Estimating divergence times in large molecular phylogenies. *Proceedings of the National Academy of Sciences*, *109*(47), 19333–19338. doi: 10.1073/pnas.1213199109

Turelli, M., & Orr, H. A. (2000). Dominance, epistasis and the genetics of postzygotic isolation. *Genetics*,

154(4), 1663–1679.

Wahlberg, N., Weingartner, E., Warren, A. D., & Nylin, S. (2009). Timing major conflict between mitochondrial and nuclear genes in species relationships of Polygoniabutterflies (Nymphalidae: Nymphalini). *BMC Evolutionary Biology*, *9*, 92. doi: 10.1186/1471-2148-9-92

Willis, S. C., Macrander, J., Farias, I. P., & Ortí, G. (2012). Simultaneous delimitation of species and quantification of interspecific hybridization in Amazonian peacock cichlids (genus cichla) using multilocus data. *BMC Evolutionary Biology*, *12*, 96. doi: 10.1186/1471-2148-12-96

Wu. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, *14*(6), 851–865. doi: 10.1046/j.1420-9101.2001.00335.x

Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution; international journal of organic evolution*, *66*(3), 763–775. doi: 10.1111/j.1558-5646.2011.01476.x

Xu, B., & Yang, Z. (2016). Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. *Genetics*, *204*(4), 1353–1368. doi: 10.1534/genetics.116.190173

Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology*, *61*(5), 854–865.

Yu, Y., Dong, J., Liu, K. J., & Nakhleh, L. (2014). Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, *111*(46), 16448–16453. doi: 10.1073/pnas.1407950111

Figure 1: Gene trees and species tree conflicts. The species tree of A, B, and C is depicted in black. In pink (gene 1) and green (gene 2) are two gene trees congruent with the species tree, *i.e.* with A and B being sister species. In light blue (gene 3), the tree of a gene subject to gene flow between species B and C. In dark blue (gene 4), the tree of a gene undergoing incomplete lineage sorting.

Figure 2: The gene-based diversification (GBD) models. Gene genealogies through species (or populations, depending on the point of view, retrospective *vs* prospective) are depicted for two hypothetical present-day genomes ($N = 2$ at $t = 0$) and five homologous genes ($n = 5$). Each grey ellipse represents a species (A-F). The grey lines represent the gene genealogies of non-sampled species at $t = 0$. The model assumes that species are quasi-static in the timescale of a few generations, and each species lineage is located in a separate column. The genealogies of genes depend on four processes: genetic differentiation/non-homologous attraction, colonization/homologous attraction, genetic drift/coalescence and gene flow/erosion.

Figure 3: Comparison of the GBD models. A) and D) Coalescence profiles of two genomes sampled at present time. B) and E) The corresponding number of ancestral species through time for each genome. C) and F) The maximal number of gene lineages belonging to each genome located in one species. The dashed line represents the last coalescence events between the two sampled genomes. Simulations were performed for $n = 20$ genes, with $\alpha = 0.1$, $\beta = 0.2$, $\delta = 0.06$ and $K = 30$ for the GBD-forward, and $a = 2$, $b = 0.2$, $d = 6$ and $N = 30$ for the GBD-backward model. Nb.: number.

Figure 4: Genealogies of a single genome generated with the GBD-forward (A) and GBD-backward models (B). The labels/locations of species (or populations, depending on the point of view, retrospective *vs* prospective) are neutral. A) Parameter settings: $\alpha = 0.5$, $\beta = 1$, $\delta = 0.2$, $n = 5$ and $N = 30$. B) Parameter settings: $a = 1$, $b = 0.1$, $d = 2$, $n = 5$ and $N = 10$.

Figure 5: Evaluation of the GBD-backward model for a single sampled genome with n genes. Parameter settings: $a = 1$, $d \in [0.2, 2]$, every 0.2, and $n = 20$, 50, 100, and 200. The number of time units t was set to

$10,000$. We sampled the number of ancestral species every 500 time units starting at time $t = 5,000$, and averaged them for each simulation. For each set of parameters, 5 replicates were performed and averaged. A) Number of ancestral species depending on the number of genes n and on the ratio $\frac{d}{a}$, for one sampled genome. B) To assess the sampling consistency of our models, k lineages were randomly sampled. The number of ancestral species reported is the number of ancestral species of these k genes only.

Figure 6: Kendall-Colijn (KC) distances among sets of gene trees simulated under the gene-based diversification (GBD-backward) model. For each set of parameters, with $t = 5,000$ (enough to reach the coalescence of all homologous genes), the median KC distances were calculated. A) Influence of the number of genes n (with $n = 10$, 15, and 20), of the number of species N (with $N = 6$ and 10), and of the ratio $\frac{d}{a}$ on the KC distances. Parameter settings: $b = 0.05$, $d = 1$, $c = 200$, and $\frac{1}{a} = 0.3$, 0.5, 0.9, 1.3, 1.7, 2.1, 2.5, 2.9, 3.3. For each set of parameters, 7 simulations were performed. B) Influence of the *homologous attraction* rate b and of the erosion-to-non-homologous attraction ratio $\frac{d}{a}$ on the KC distances. Parameter settings: $n = 10$, $N = 6$, $b = 0.01$, 0.02, 0.05, 0.12, $d = 1$, $c = 200$, and $\frac{1}{a} \in [0.3, 3.5]$, every 0.2. For each set of parameters, 75 simulations were performed.

Figure 7: Minimization of the Kullback-Leibler (KL) divergence between empirical and simulated trees, *i.e.* between their distributions of KC pairwise distances. Two parameters were optimized for each model. The *speciation* rate (λ) and the *extinction* rate (μ) for the multi-species coalescent (MSC) model (with coalescence rate set to 1). The *homologous attraction* b and the ratio of the *erosion* rate over the *non-homologous attraction* rate ($\frac{d}{a}$) for the gene-based diversification (GBD-backward) model (with d set to 1). For each set of variables, 75 simulations were performed and averaged. The same color scale was used for each empirical data-set. For each optimization analysis, the cell for which we found the best fit between empirical and simulated trees (smallest KL divergence) is framed.

Figure 8: Best fit between empirical and simulated trees, *i.e.* between their distributions of KC pairwise distances (selected cells of Fig. 7). For each set of variables, 75 simulations were performed and averaged.

a: *non-homologous attraction* rate, b: *homologous attraction* rate, d: *erosion* rate (set to 1), λ: *speciation* rate, μ: *extinction* rate, KL: Kullback-Leibler.

Figure 1

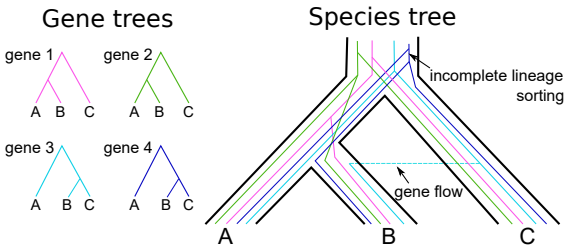

Gene trees
gene 1
A B C
gene 2
A B C
gene 3
A B C
gene 4
A B C
Species tree
incomplete lineage sorting
gene flow
A
B
C

Figure 2

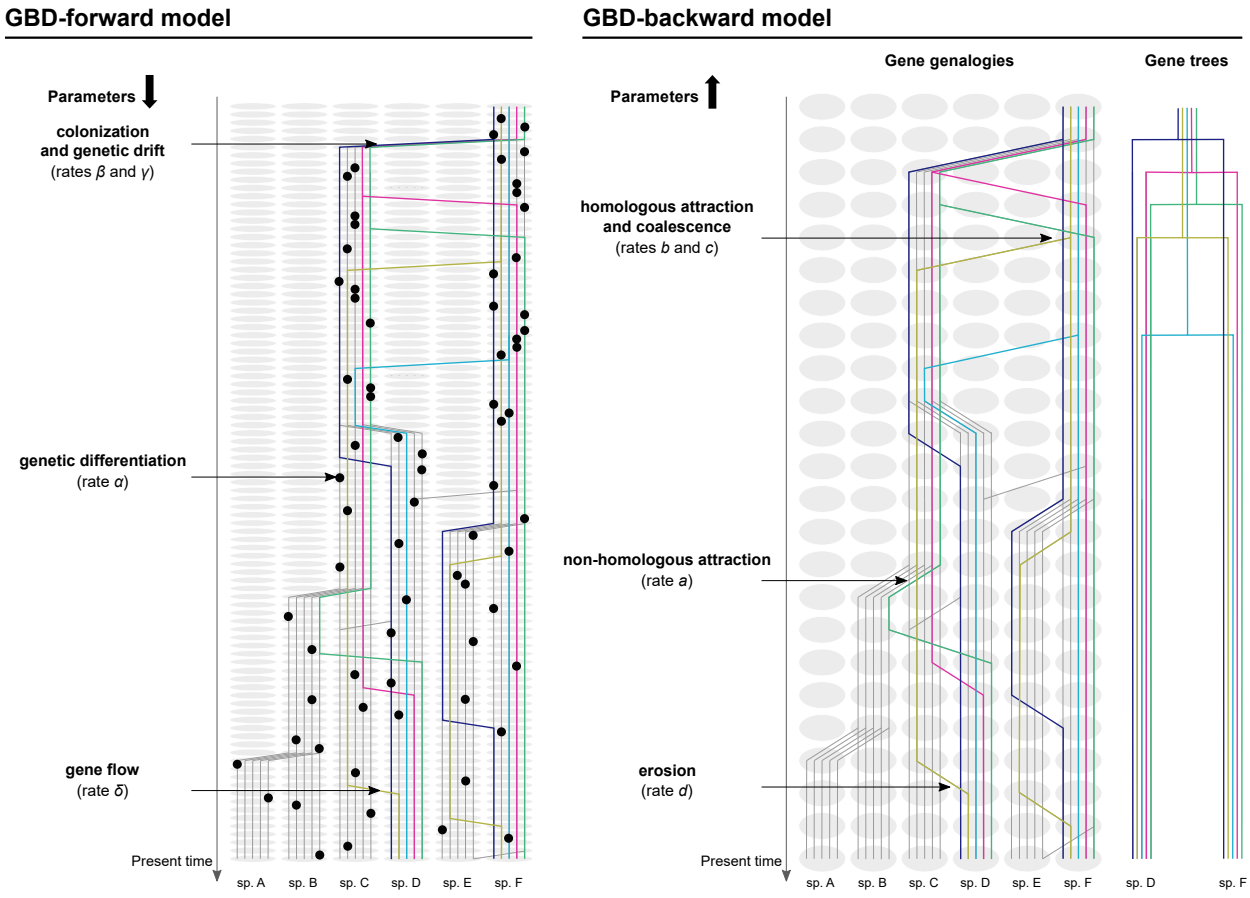
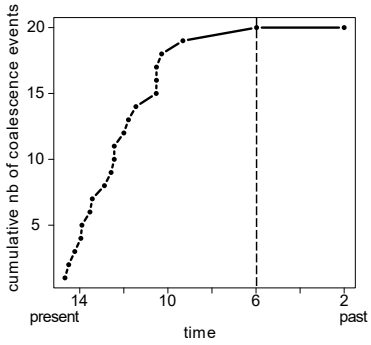

GBD-forward model
Parameters
colonization
and genetic drift
(rates β and γ)
genetic differentiation
(rate α)
gene flow
(rate δ)
Present time
sp. A
sp. B
sp. C
sp. D
sp. E
sp. F
GBD-backward model
Gene genalogies
Gene trees
Parameters
homologous attraction
and coalescence
(rates b and c)
non-homologous attraction
(rate a)
erosion
(rate d)
Present time
sp. A
sp. B
sp. C
sp. D
sp. E
sp. F
sp. D
sp. F

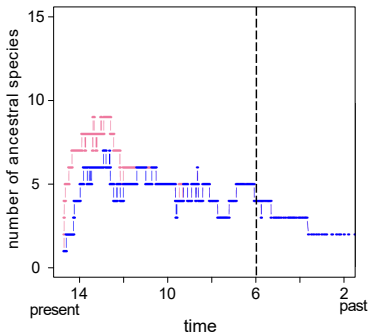
Figure 3

GBD Forward

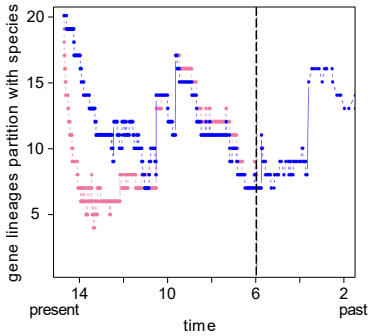
A) Coalescence profile



B) Ancestral species number

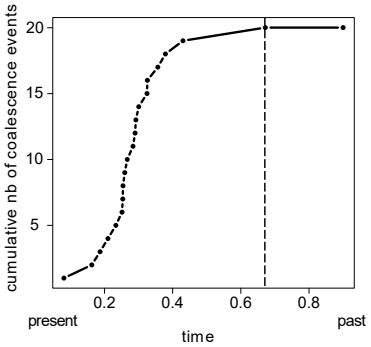


C) Maximal nb. of gene lineages located in one species

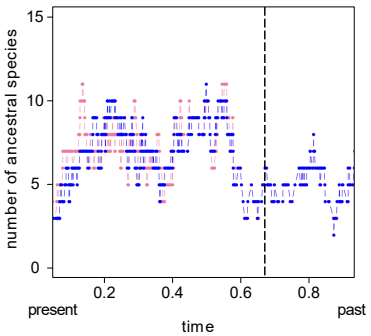


GBD Backward

D) Coalescence profile



E) Ancestral species number



F) Maximal nb. of gene lineages located in one species

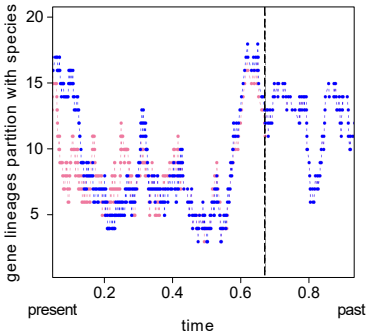


Figure 4

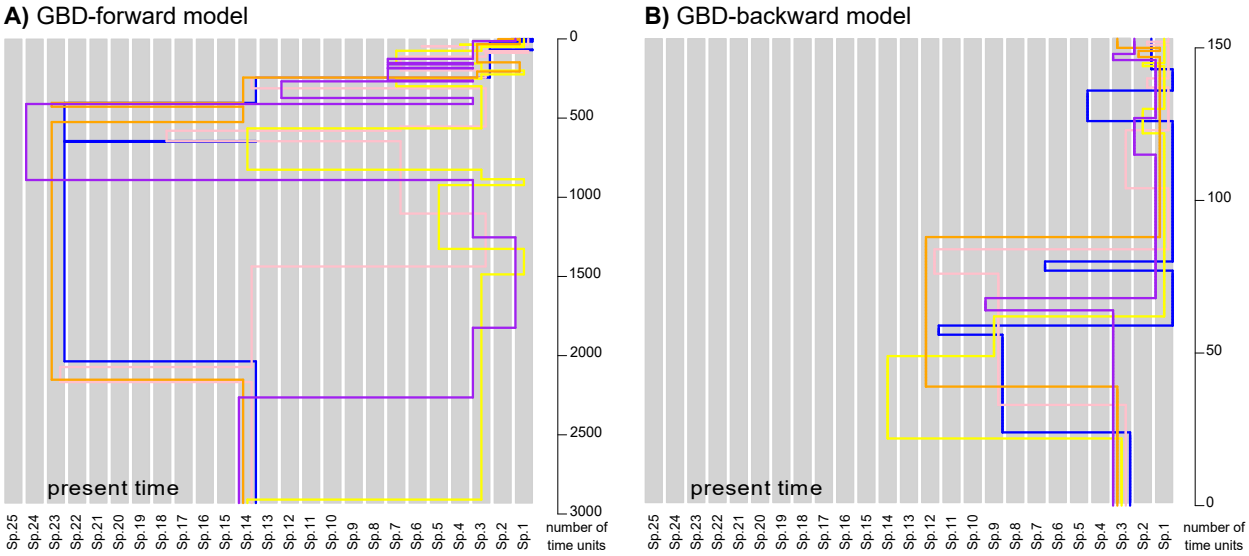

A) GBD-forward model
present time
0
500
1000
1500
2000
2500
3000
number of time units
Sp.25 Sp.24 Sp.23 Sp.22 Sp.21 Sp.20 Sp.19 Sp.18 Sp.17 Sp.16 Sp.15 Sp.14 Sp.13 Sp.12 Sp.11 Sp.10 Sp.9 Sp.8 Sp.7 Sp.6 Sp.5 Sp.4 Sp.3 Sp.2 Sp.1
B) GBD-backward model
present time
150
100
50
0
number of time units
Sp.25 Sp.24 Sp.23 Sp.22 Sp.21 Sp.20 Sp.19 Sp.18 Sp.17 Sp.16 Sp.15 Sp.14 Sp.13 Sp.12 Sp.11 Sp.10 Sp.9 Sp.8 Sp.7 Sp.6 Sp.5 Sp.4 Sp.3 Sp.2 Sp.1

Figure 5

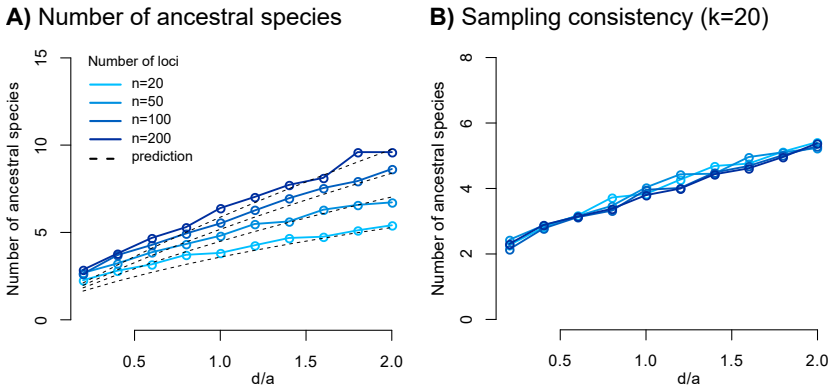
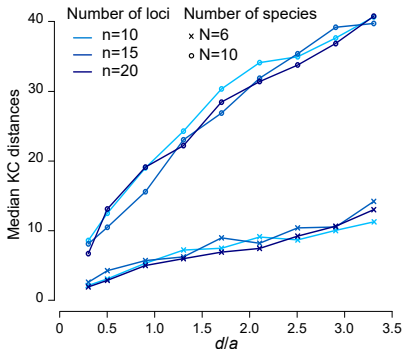

A) Number of ancestral species
B) Sampling consistency (k=20)
Number of loci
n=20
n=50
n=100
n=200
prediction
Number of ancestral species
d/a
0
5
10
15
0.5
1.0
1.5
2.0
0
2
4
6
8

Figure 6

A) Influence of the number of loci (*n*) and of the number of species (*N*)



B) Influence of the homologous attraction (rate *b*)

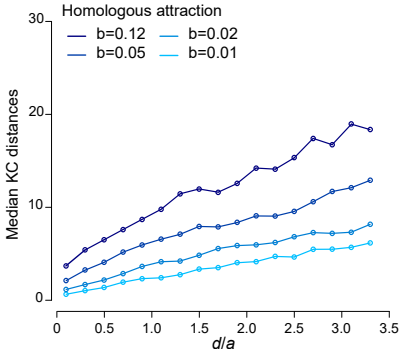


Figure 7

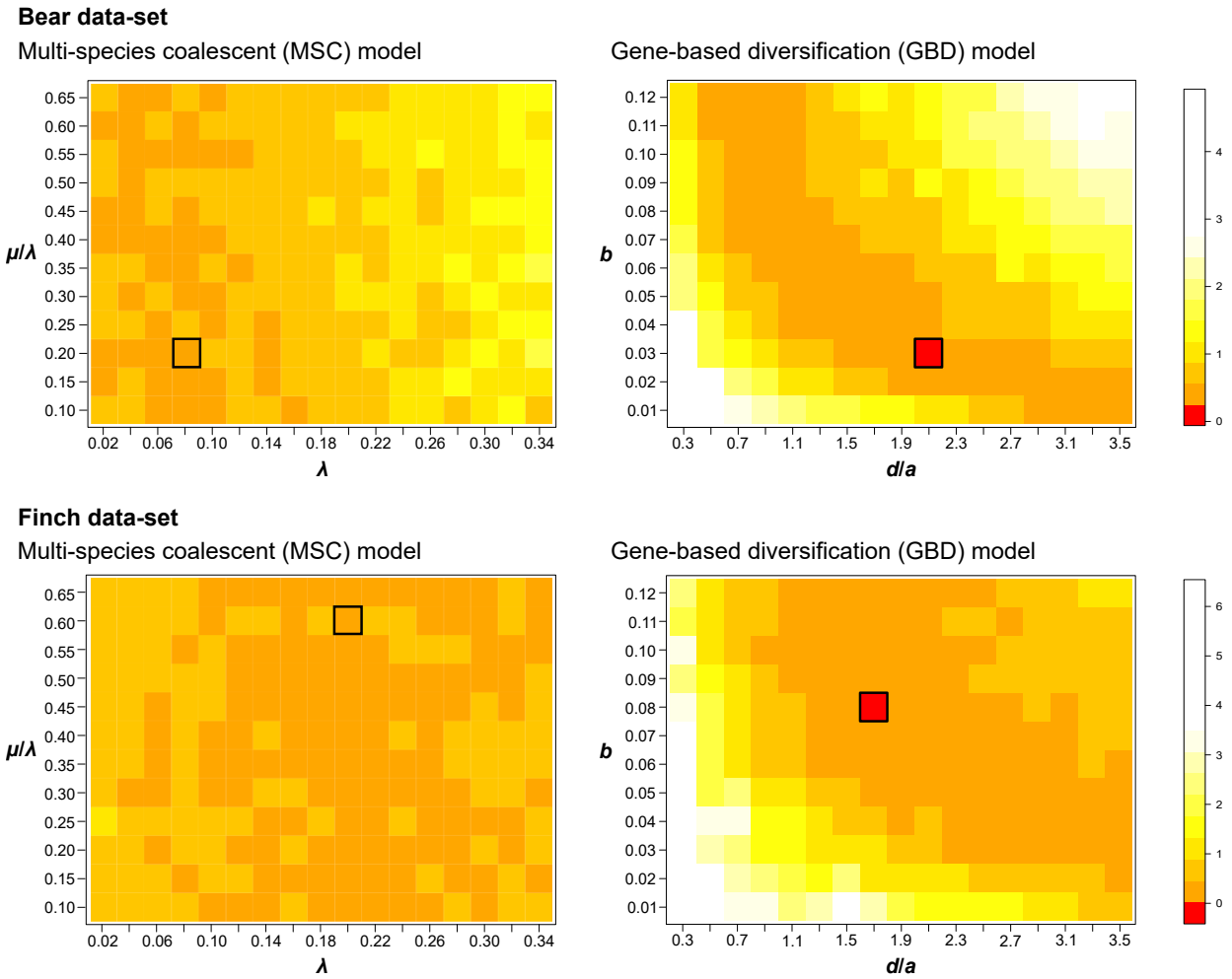
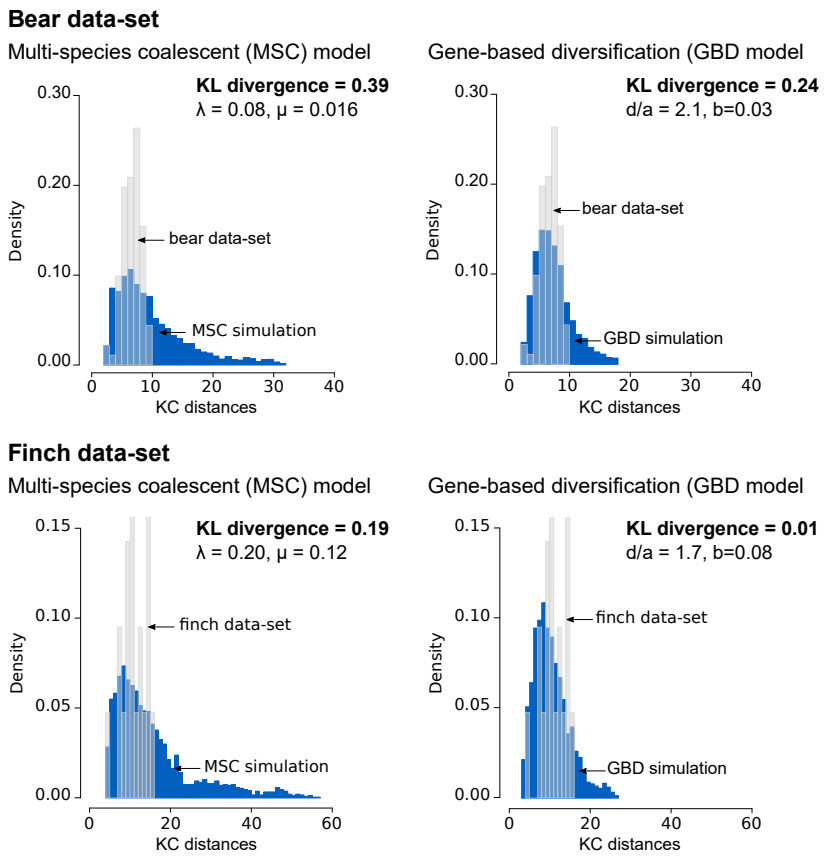

Bear data-set
Multi-species coalescent (MSC) model
Gene-based diversification (GBD) model
μ/λ
λ
b
d/a
Finch data-set
Multi-species coalescent (MSC) model
Gene-based diversification (GBD) model
μ/λ
λ
b
d/a

Figure 8


Bear data-set
Multi-species coalescent (MSC) model
KL divergence = 0.39
λ = 0.08, μ = 0.016
bear data-set
MSC simulation
Density
KC distances
Gene-based diversification (GBD model
KL divergence = 0.24
d/a = 2.1, b=0.03
bear data-set
GBD simulation
Density
KC distances
Finch data-set
Multi-species coalescent (MSC) model
KL divergence = 0.19
λ = 0.20, μ = 0.12
finch data-set
MSC simulation
Density
KC distances
Gene-based diversification (GBD model
KL divergence = 0.01
d/a = 1.7, b=0.08
finch data-set
GBD simulation
Density
KC distances