Article

Ghost lineages deceive introgression tests and call for a new null hypothesis

Théo Tricou[1,*], Eric Tannier[1,2] and Damien M. de Vienne[1,*]

[1] Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Évolutive UMR5558, F-69622 Villeurbanne, France
[2] INRIA Grenoble Rhône-Alpes, F-38334, France

***Corresponding authors: E-mails:** theo.tricou@univ-lyon1.fr, damien.de-vienne@univ-lyon1.fr

---

Abstract

The data that is known and sampled in any evolutionary study is always a small part of what exists, known or not, or what has existed in the past and is extinct. Therefore it is likely that all detected past horizontal gene fluxes, hybridization, introgressions, admixtures or transfers, involve "ghosts", that is, extinct or unsampled lineages. The presence of these ghosts is acknowledged by all scientists, but almost all wish that and make as if their blurring influence would be low, like a background noise that, with a reasonable approximation, can be ignored. We assess this undervalued hypothesis by qualifying and quantifying the effect of ghost lineages on introgression detection by the popular D-statistics method. We use a genomic dataset of bears to illustrate and circumscribe the possibility of misinterpretation and show on simulated data that under certain conditions, far from unrealistic, most results interpreted from D-statistics, concerning the existence of introgression and the identity of donors and recipients of horizontal gene flows, are erroneous. In particular, the use of a distant outgroup, usually given as a solid ground for these tests, leads in fact to an increase in the error probability, and to false interpretations in a vast majority of the cases. We argue for a switch of the null hypothesis: the results of detection methods for gene fluxes should be interpreted with the full and visible participation of the unknown ghosts.

Key words: ghost lineage, D-statistics, ABBA-BABA, introgression, gene-flow, simulation

Introduction

Evolutionary studies are always restricted to a subset of species, populations or individuals. By choice, because the question addressed is only relevant to a fraction of the data, and by necessity, because methods have methodological and technical limits. Another reason is that most lineages are simply unknown. More than 99.9% of all species that ever lived are now extinct (Raup 1991), and out of those that are not, only a small fraction has been described. The number of extant species still uncatalogued is almost an order of magnitude higher than reported ones (around 1.3 Million in 2011, Mora et al. 2011), and many orders of magnitudes higher if considering microbial species (Locey and Lennon 2016).

Accounting for these "ghost lineages" (extinct and unsampled) is particularly relevant when studying horizontal fluxes of genetic material across taxa, because many inferred such events are likely to involve these lineages. This was repeatedly acknowledged (Maddison 1997; Galtier and Daubin 2008; Green et al. 2010; Deren A. R. Eaton and Ree 2013; Szöllősi et al. 2013; Szöllősi et al. 2015), especially in studies of introgression between populations, but was considered in this context either a source of noise (Pease and Hahn 2015), or at best a problem that could be worked out by sampling more or performing multiple tests (Eaton et al. 2015; Kumar et al. 2017; Barlow et al. 2018). Very recently Hibbins and Hahn (2021) advised to "keep in mind" this possibility of misinterpretation but as far as we know, the real impact of ghost lineages on the capacity of gene flow detection methods has not been properly evaluated and quantified.

Introgression depicts the integration of genetic material from one lineage to another, through hybridization and subsequent backcrosses. This gene flow across species boundaries appears to be common in the Eukaryotic Tree of Life, and was proven to be sometimes adaptive (see Hedrick 2013 for a review). Cases of introgression were reported in lineages as diverse as humans (Green et al. 2010; Meyer et al. 2012), boars (Liu et al. 2019), butterflies (Martin et al. 2013; Smith and Kronforst 2013; Massardo et al. 2020), fishes (Schumer et al. 2016; Meier et al. 2017), plants (Deren A. R. Eaton and Ree 2013; Zhang et al. 2019) or fungi (Zhang et al. 2018; Keuler et al. 2020), to only cite a few.

In the last decade, the availability of an increasing number of sequenced genomes and the development of dedicated methods have improved the detection of introgression. One of the most widely used such methods is the D-statistics (or Patterson's D), also known as the ABBA-BABA test (Green et al. 2010; Durand et al. 2011). The reasons for its success are manyfold. D-statistics is easy to understand, easy to implement, fast to compute and easy to interpret, and has the advantage of being a phylogenetic-discordance-based method that can distinguish between incongruence resulting from Incomplete Lineage Sorting (ILS) and incongruence resulting from gene flow. The test considers four species, three species as ingroup and one outgroup, with ladder-like phylogenetic relationships (fig. 1). It consists in counting the number of sites that support a discordant topology. Two SNP patterns are considered, ABBA and BABA, depending on which allele (A: ancestral and B: derived) is found in each species. The null hypothesis of this test states that under a scenario with no gene flow, both patterns can be attributed to ILS and should thus be observed in equal numbers. Significant deviation from this expectation is usually interpreted as introgression between two of the three lineages forming the ingroup (fig. 1, green arrows). The outgroup is chosen distant enough from the ingroup for not being suspected to be involved in introgressions with any of the ingroup lineages.

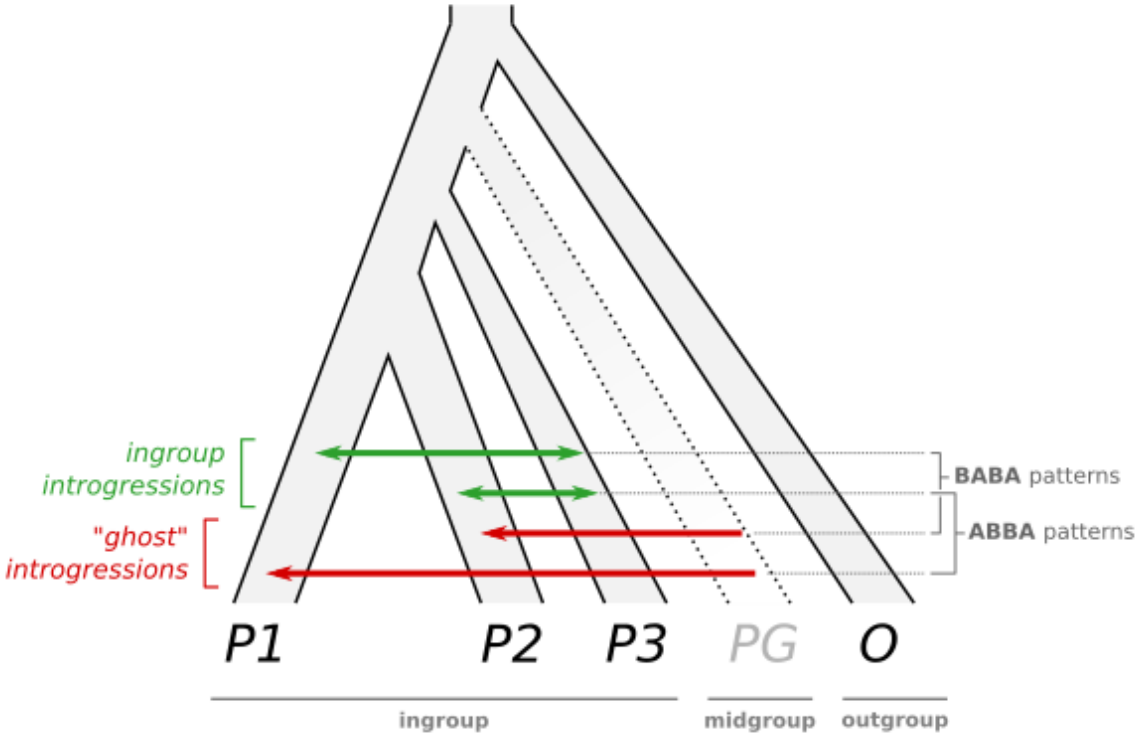

Fig. 1. Introgression events that can result in a detectable excess of ABBA or BABA patterns as identified by the D-statistics test. The common interpretation of this excess is the hypothesis of an "ingroup" introgression (green arrows). However "midgroup" introgressions (red arrows) from ghost lineages can produce similar patterns.

Among several possible confounding effects of the D-statistics (Simon H. Martin et al. 2015; Yichen Zheng and Janke 2018), the impact of ghost lineages on introgression detection was foreseen. Working on a subset of the species logically leads to an underestimation of the true frequency of introgression and thus an inflated role of ILS (Maddison and Knowles 2006). In this line it was clearly stated that the donor genome could easily be misidentified because introgression from its sister ghost lineage to the same recipient genome would produce the same signal (Deren A. R. Eaton and Ree 2013; Eaton et al. 2015; Pease and Hahn 2015; Yichen Zheng and Janke 2018). Another impact of those ghost lineages, however, was anticipated early in the history of the test (in the article that formalizes it, Durand et al. 2011), but was largely overlooked: the fact that some introgressions from ghost lineages branching between the ingroup and the outgroup (the "midgroup", see fig. 1) could lead to the wrong identification of both the donor *and* the recipient genomes (fig. 1, red arrows). In other

words, in this scenario, none of the species usually considered as taking part in the introgression are correctly identified. The existence of this possible source of error in the interpretation of the D-statistics was acknowledged by some scientists as a possible confounding signal (Durand et al. 2011; Jente Ottenburghs et al. 2017; Yichen Zheng and Janke 2018; Hibbins and Hahn 2021) but surprisingly its impact has never been quantified.

In this study, we first analyse a genomic dataset from a phylogenetic study of bears to show how the interpretation of introgression tests can depend on the presence or absence of relevant lineages in the sampled data. We then quantify the effect of ghost lineages on the misidentification of donors and recipients lineages using simulations. In particular we explore the effect of the choice of the outgroup, the size of the tree and the effect of preferential transfers to closely related species on this misidentification.

We show that under the realistic assumption of many extinct or unknown lineages in the vicinity of the lineages analyzed, and under a simple demographic history of the populations considered, most significant D-statistics tests are attributable to ghost lineages. This leads to the wrong identification of the lineages involved (donors and recipients) in the majority of the cases, notably when the distance between the ingroup and the outgroup is big. This is particularly problematic since the outgroup is usually chosen distant enough from the ingroups to avoid the influence of introgression between the two. This hampers the delimitation of a safe zone for the D-statistics.

These results call for a new way of interpreting D-statistics, and questions introgression detection methods in general. Our results are illustrative of the recent statement by Ottenburghs (2020) that "the presence of ghost introgression has important consequences for the study of evolutionary processes", and proposes a first quantification of this importance.

2. Results

2.1. The problem of interpreting significant D-statistics by introgression between sampled lineages

D-statistics results, because they are computed on species present in the study, are usually interpreted as introgression between those known lineages. We show here by reanalysing a

genomic dataset that this interpretation highly depends on the taxonomic sampling and the prior knowledge we have on related lineages.

We studied five species of bears with the known topology presented in figure 2A (data from Barlow et al. 2018), from which we deliberately removed either the Swedish bears (fig 2B, Sampling 1), not suspected to be involved in introgressions with other bear species, or the polar bears (Sampling 2) that are suspected from previous studies to be involved in an introgression with the bears from Alaska (Cahill et al. 2013; Liu et al. 2014; Lan et al. 2016; Kumar et al. 2017). Using a multiple sequence alignment of the full genomes, we then counted for each sampling the number of ABBA and BABA patterns and computed the D-statistics. We then applied a weighted block jackknife and calculated a Z-score (distance to standard deviation) to test for significance (see material and methods for more details on the applied protocol). For the first sampling, we identified 175,413 ABBA patterns and 226,992 BABA patterns resulting in a significantly negative D-score (D=-0.128, Z=-21.4), whose interpretation is in accordance with the introgression previously suspected (fig. 2). For the second sampling, we identified 265,220 ABBA patterns and 215,443 BABA patterns resulting in a significantly positive D-score (D = 0.104, Z-score = 20.3).

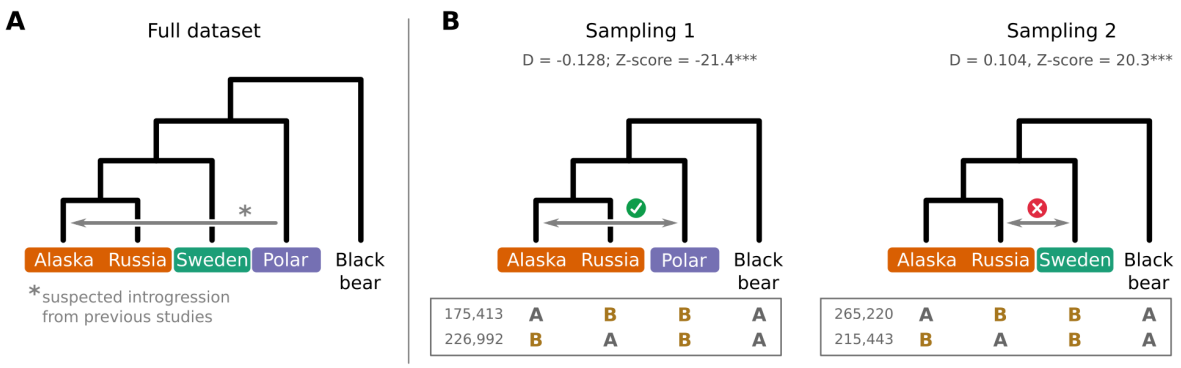

Fig. 2. Illustration of the effect of different samplings on the interpretation of the D-statistics using a genomic dataset of bears. **A**. The five bear species used in this experiment and their phylogenetic relationships. The grey arrow shows the introgression that was inferred from previous studies. **B**. Two samplings realized prior to computing the D-statistics test on the 4 remaining species. The number of ABBA and BABA patterns are displayed below the trees. In Sampling 1 a lineage that is not suspected to be involved in an introgression is removed. In Sampling 2, the donor of the suspected introgression is excluded. The introgressions inferred based on the usual way of interpreting the D-stats (grey arrows), and their compatibility with the introgression described in other studies (green tick and red cross) are indicated in the trees.

If we were to interpret this second result as introgression between lineages present in the test (Alaskan, Russian, Swedish and Black bears), we would conclude to an introgression between bears from Sweden and Russia (fig. 2B), while this significant score can be attributed to the introgression between Polar bears (absent from the sampling) and Alaskan bears. This alternative hypothesis, however, relies on our knowledge of the existence of polar bears. It could thus also be questioned if we knew the existence of another lineage, because we can never consider that we know all lineages leading to extant or extinct species. As such, any interpretation by known lineages, even with a good taxonomic sampling, has real chances to infer an nonexistent introgression. It is unknown if there is a point to which such interpretation is safe. In the sequel of this article we thus quantify this effect. We show that it can be highly confusing, and argue for a way of interpreting the D-statistics including unknown lineages.

2.2. Quantification of the interpretation problem by simulations

2.2.1 Significant D-statistics are often due to ghost lineages

To evaluate the impact that ghost lineages can have on the detection of introgressions with the D-statistics, we performed various simulations following these steps: (i) Simulate a species tree under a simple birth-death model, (ii) sample independent introgression events occurring between contemporaneous lineages in the simulated species tree, (iii) for every possible combination of 4 lineages (*i.e.* quartet) with the topology (((P1,P2),P3),O), count the number of the sampled introgressions that would lead to a significant D-statistics test (see Material and Methods for details).

Significant D-statistics were due either to ingroup introgressions (the classical interpretation), or to midgroup introgressions (see fig. 1). The case of introgression from the outgroup or from outside the quartet were discarded to conform to the standard situation of the use of D-statistics. The mean proportion of midgroup introgressions over all significant D-statistics (hereafter referred to as the *proportion of erroneous interpretations*) was recorded and its value was evaluated for different values of the parameters tested. Overall, we observed between 15% and 100% of erroneous interpretations (see the results hereafter) depending on (i) the choice of the outgroup, (ii) the existence of 'preferential' introgressions between close

relatives and (iii) the proportion of unsampled lineages considered (supplementary fig. S1). These results are detailed in the following sections.

2.2.2. Distant outgroups lead to an increase in the erroneous interpretation of the test

In D-statistics tests, outgroups are usually chosen distant enough from the ingroup so that the chance that they are involved in the detected introgression is minimized. Zheng and Janke (2018) have stated that the distance between the outgroup and the ingroup had no to little impact on the D-statistics power. However this study focused on the effect of saturation of sequence substitutions in the outgroup and did not consider the possible role of extinct and unsampled lineages. Here we investigate how choosing a distant outgroup increases the chances for significant D-statistics to be due to midgroup introgressions.

Using our simulations, we tested the effect of the distance to the outgroup on the proportion of erroneous interpretations. As a measure of the distance to the outgroup we used the ratio $R=t1/t2$ between the distance separating the most recent common ancestor of the three ingroup taxa and the most recent common ancestor of all four taxa (t1 in fig. 3), and the total height of the 4-taxa tree (t2 in fig. 3). We then computed the proportion of erroneous interpretations for all quartets (R = 0) and for nine subsets of quartets with distances to the outgroup (R) higher than x, x varying from 0 to 0.9 with a step of 0.1. We observe that the proportion of midgroup introgressions increases with the distance to the outgroup, leading to wrong identification of the donor and the recipient of the introgression. On average, when R is over 0.3, more than 50% of the significant D-statistics are imputable to ghost introgressions (fig. 3).

For comparison, among many other examples, in the first study that used the D-statistics (Green et al. 2010), the ratio R was equal to 0.873 (825,000 years separate modern humans to Neanderthal while 6.5 million years separate humans and chimps). For the bear example mentioned in Section 2.1, articles that identified the introgression between polar and Alaskan bears (Cahill et al. 2013; Liu et al. 2014; Lan et al. 2016; Kumar et al. 2017; Barlow et al. 2018) used two different outgroups, black bears and pandas, with R values of 0.4 and 0.9, respectively. These values all fall into a range of high probability of erroneous interpretation in our study.

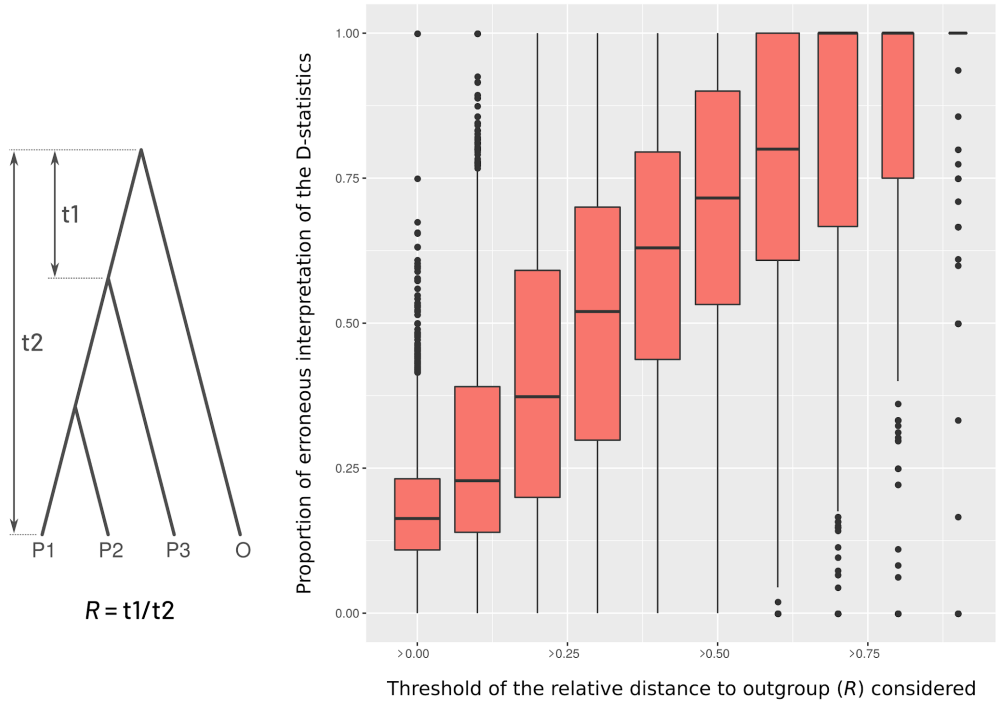

Fig. 3. Effect of the outgroup distance on the proportion of erroneous interpretation of the D-statistics. Left: distances used to compute the relative distance of the outgroup (*R*). Right: proportion of erroneous interpretations observed (*y*-axis), as a function of the threshold for *R* above which quartets are considered (*x*-axis).

2.2.3. Preferential introgression can mitigate the impact of ghost introgression on the amount of erroneous interpretation of the test

Introgression results from hybridization, whose increased probability when decreasing genetic distance between species was repeatedly demonstrated (Edmands 2002; Chapman and Burke 2007; Montanari et al. 2014). The rationale behind this is that more closely related species are more genetically compatible and therefore produce more viable hybrids than more distant (and thus genetically incompatible) ones (Mallet 2005).

To test the effect of these preferential introgressions on the proportion of erroneous interpretation of the D-statistics, we incorporated in the simulations a probability of introgression that was dependent on the phylogenetic distance between the species. The strength of this "phylogenetic effect" was controlled by the parameter α of an exponential law (see material and methods). The effect of α is illustrated in figure 4A. When $\alpha = 0$, introgressions occur at random between species; when $\alpha = 1000$, introgressions occur almost

exclusively between sister species. Intermediate values (α = 1, 10, 100) produce intermediate patterns.

We observed that the impact of the outgroup distance on the proportion of erroneous interpretation decreased with increasing values of α (fig. 4B). In simulations where alpha is maximum (α = 1000), the proportion of significant D-statistics due to ghost introgression is around 0.25%, regardless of the distance separating the ingroup from the outgroup. This can be explained by the fact that most introgressions occur between lineages that are so close that they are rarely outside the ingroup. Taken together, these findings demonstrate that when preferential introgressions are common in a clade, the amount of ghost introgression is decreased and so is the proportion of misinterpretation of the D-statistics. Still, we note that even under conditions of highly preferential introgressions, the error remains quite high (it does not decrease below 25% in our setting, on average).

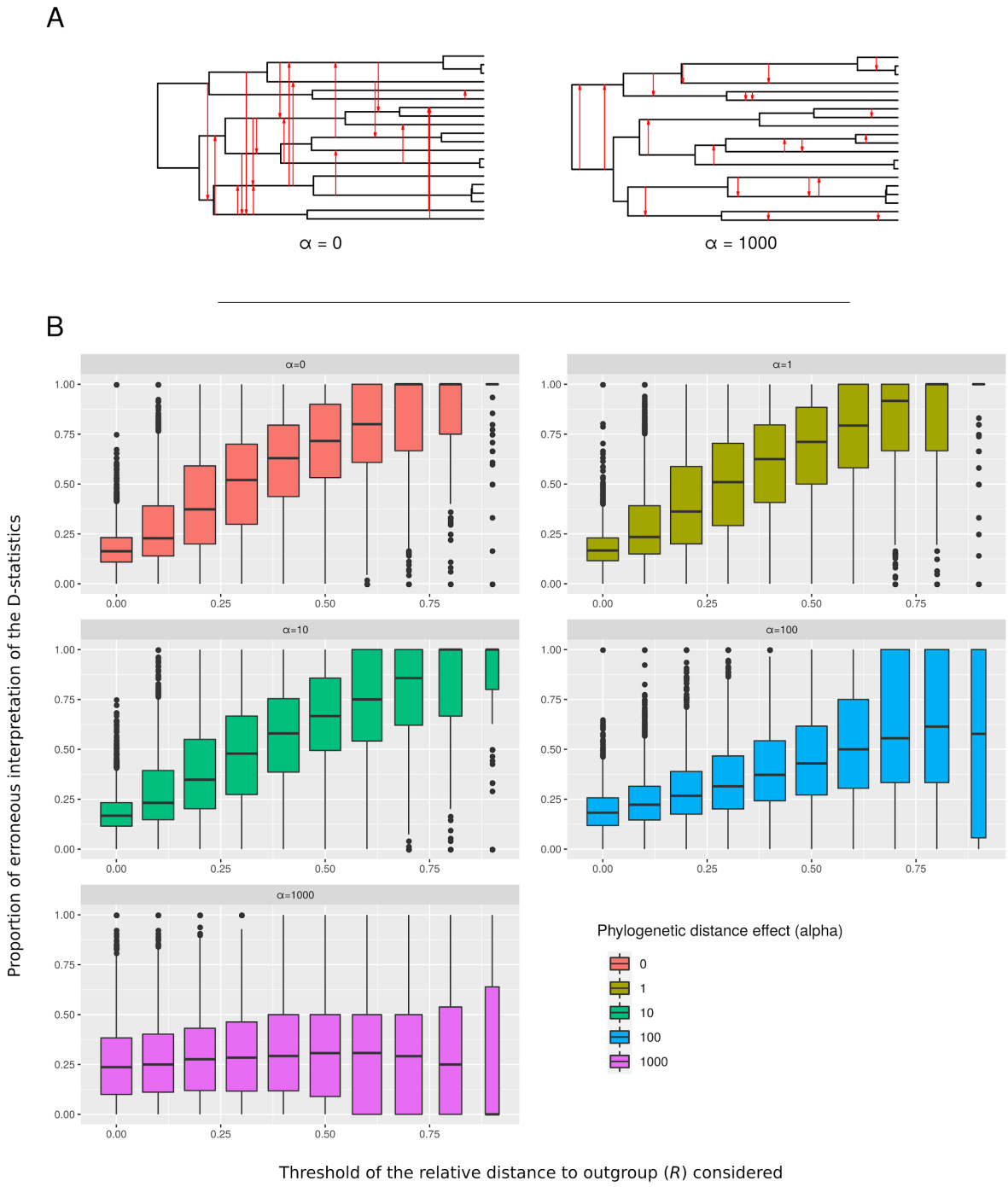

Fig. 4. Effect of preferential introgressions with closest relatives on the proportion of erroneous interpretation of D-statistics. **A.** Illustration of the effect of the α parameter, controlling the phylogenetic distance effect on introgressions. **B**. Relationship between the outgroup distance (x-axis) and the proportion of erroneous interpretations (y-axis) for different strengths of the phylogenetic distance effect as controlled by α (α = 0 -no effect-, 1, 10, 100 and 1000). For each parameter α, 100 species trees with 20 extant species were simulated and 100 introgressions were sampled.

2.2.4. The error probability increases with the number of unsampled taxa but not with the number of extinct ones

The proportion of erroneous interpretations of the D-statistics, computed in previous sections, can be seen as a probability, for a given quartet, to display a significant D-statistics not because of ingroup introgressions but because of ghost introgression. This probability is thus affected by the proportion of the extant species that are considered in the analysis. Until now, we have been conservative by simulating species trees with 20 extant species and considering all of them for building the quartets and computing the proportion of erroneous interpretations of the D-statistics. But what if these 20 tips represent only a fraction of the species that are really extant? Answering this question allows mimicking the actual knowledge that we have on the species we study, with many (or most) of them being still undescribed (Chapman et al. 2009).

We simulated species trees with N = 20, 40, 60, 80 and 100 species in total, and randomly sampled 20 species each time. This allowed simulating sampling efforts between 100% (20 species out of 20) and 20% (20 species out of 100). We observe that low sampling contributes to an increase in the amount of misinterpreted D-statistics due to ghost introgression (fig. 5). While the mean proportion of erroneous interpretation is around 25% when 100% of extant lineages are sampled, this error is close to 60% when only 20% of the lineages are known.

The effect of an increase of the proportion of erroneous interpretation when decreasing the fraction of extant species sampled is observed whatever the value of α (the phylogenetic distance effect) and whatever the distance to the outgroup considered (supplementary fig. S1).

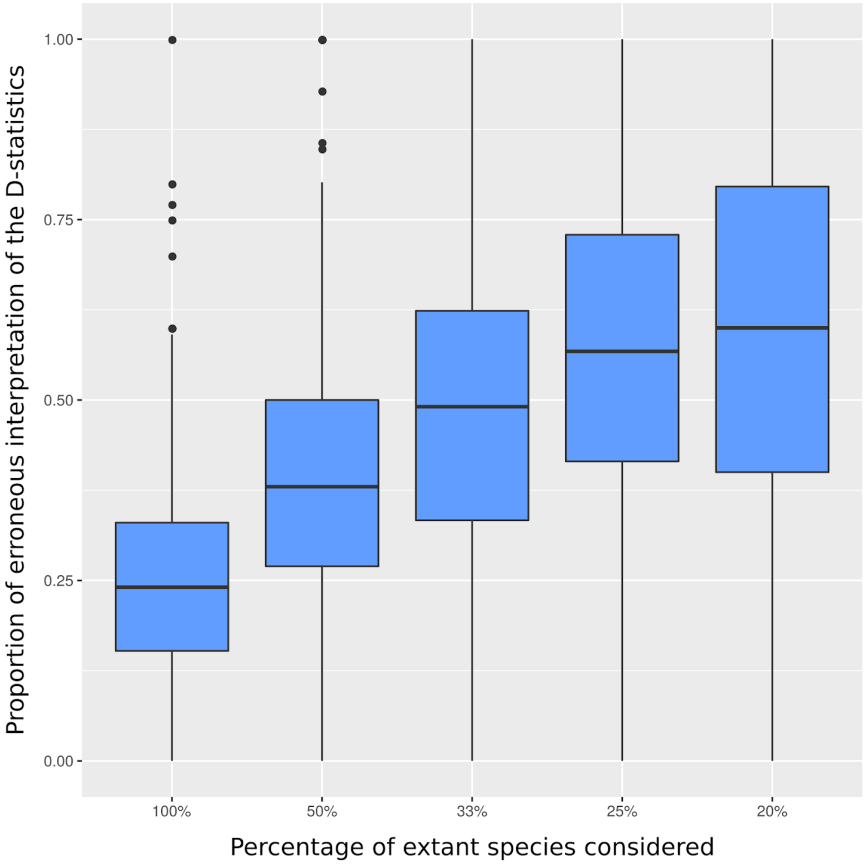

Fig. 5. Effect of taxonomic sampling on the proportion of erroneous interpretation of D-statistics. Mean proportion of erroneous interpretations observed (y-axis) as a function of the taxonomic sampling effort (*x*-axis). For this simulation, the phylogenetic distance effect (α) was null and no filter was applied on the outgroup distance.

When looking at the relationship between the number of extinct lineages in the tree and the proportion of erroneous interpretation of the D-statistics (supplementary fig. S2), we found no correlation, unlike initially expected. An intuitive explanation can be proposed. By definition, under a simple birth-death model, an extinct lineage branching in the midgroup (and thus being potentially responsible for a midgroup introgression) originates earlier in the tree than an extinct lineage branching in the ingroup (and thus being potentially responsible for an ingroup introgression), and it also gets extinct earlier. Consequently, increasing the number of extinct lineages in the tree (by increasing the probability of extinction in the birth-death process) increases the number of possible extinct ingroup donors more rapidly than the number of possible midgroup donors. This may lead to the absence of a detectable signal. Further investigations are needed to better characterize this effect in the future.

3. Discussion

Previous studies have explored different parameters affecting the robustness or sensitivity of the D-statistics. Demographic aspects, such as population size and variation thereof (Eriksson and Manica 2012; Simon H. Martin et al. 2015; Yichen Zheng and Janke 2018), or ancestral population structure (Durand et al. 2011; Lohse and Frantz 2014; Simon H. Martin et al. 2015) were shown to decrease its sensitivity. Authors also noticed that the use of the D-statistics to locate introgressed loci by applying it to small windows across the genome could give stochastic estimates of D (Simon H. Martin et al. 2015), and that complex introgression scenarios, with more than one introgression recipient in the quartet, was a source of error of the D-statistics (the test being designed to identify a unique introgression event, Leo Elworth et al. 2018).

Our findings strongly suggest that, in addition to the effects listed above, D-statistics can be strongly impacted by ghost lineages, which changes the way the test should be used and interpreted by default.

Recently, Hibbins and Hahn (2021) used simulations to illustrate how robust different methods are to different introgression scenarios. As initially proposed by Durand et al (2011), and confirmed here, they showed that a ghost introgression from a midgroup lineage (scenario called "Outgroup ghost to P1" in their article) could result in a significant D-statistics, in which case both donor and recipient lineages could be misidentified if the statistics was interpreted by introgression between known lineages. The authors observed comparable results with the D3 method (Hahn and Hibbins 2019), a D-statistics-like method that uses only three lineages and tree branch lengths instead of SNP pattern to detect introgression. While this study confirms that ghost introgressions are a source of possible erroneous interpretation, it does not quantify the extent to which this error impacts the interpretation of the D-statistics.

Several methods have been developed to alleviate some of the limitations of the D-statistics, but their robustness to ghost lineages has not been explored. Soraggi et al. (2018) proposed an extension of the D-statistics (D_{ext}) to study introgression events among non-African populations while using Africans as the outgroup. This test anticipates (and deals with) the

expected signal in the D-statistics due to the known introgression from Neanderthal to European populations. The use of this "extinct-aware" version of D-statistics is however restricted to documented extinct clades for which introgression events with extant species have been not only identified but also quantified. This precludes its use in cases when there is no *a priori* on the existence of ghost lineages.

Other extensions of the D-statistics, namely the partitioned-D (Deren A. R. Eaton and Ree 2013) and the D_{FOIL} (Pease and Hahn 2015), were proposed to detect introgression on 5-taxon phylogenies (instead of 4) and to be able to polarize (in some cases) the introgressions. Ghost introgression was mentioned to be a possible source of noise in D_{FOIL} (Pease and Hahn 2015), but just as for the D-statistics the amount of errors due to this alternative scenario had not been quantified. We tested the D_{FOIL} approach on our simulations (supplementary material) and observed that it was not affected by ghost introgressions in the case where the test was revealing a polarized introgression, but that the proportion of such cases decreased drastically when increasing the number of ghost lineages (supplementary fig. S3). So the use of such method does not seem to alleviate the ghost lineage issue that we brought out in this study.
Another approach for dealing with more than 4 or 5 taxa with the D-statistics consists in performing multiple D-statistics on a single dataset, changing each time some of the ingroup lineages that form the quartet. Joint analyses of all the D-statistics are done afterwards, using the interpretation of each individual test as a constraint for the interpretation of the others. These constraints permit a finer detection of introgression, by allowing to polarize the events and to assign introgressions to groups of taxa instead of single taxa (Pease et al. 2016; Rouard et al. 2018; Suvorov et al. 2020). Unfortunately, the interpretation of each individual test is usually performed without the possibility of ghost introgression in mind. The joint interpretation of the multiple tests is done intuitively without a formal method, and only a discretionary fraction of all possible combinations of 4-taxa is considered. Altogether, it makes it hard to anticipate the effect of ghost lineages with these multiple tests.

We think that there is room for approaches combining the results of all possible D-statistics for a dataset, recognizing that there are elements that can influence the interpretation but that we cannot know *a priori* and considering ghost introgression as the default hypothesis. This approach could eventually yield automatic methods to predict, without other knowledge than

introgressed genes, the existence of archaic species, as it was for example achieved with ad-hoc methods in particular lineages of human (Prüfer et al. 2014; Dannemann and Racimo 2018) and whales (Foote et al. 2019). This is a promising route for future work.

A last solution to limit the effect of ghost lineages on the interpretation of the D-statistics is to improve the species sampling. Indeed, we show in our results that the proportion of erroneous interpretation decreases when increasing the sampling effort. But although some families are believed to be extensively described (Chapman et al. 2009), it is not reasonable to assume that the whole extinct and extant diversity is known. Reaching exhaustivity of past and present biodiversity cannot be the way to cope with biases introduced by the unknown.

Beyond D-statistics, more global methods are designed to detect introgression with a complete set of species such as STRUCTURE/ADMIXTURE (Pritchard et al. 2000; Tang et al. 2005), Treemix (Pickrell and Pritchard 2012) or Phylonet (Than et al. 2008; Wen et al. 2018). These tools do not consider or mention ghost lineages either and their potential blurring effect on the signal detected. Thereby introgressions events are only inferred between known lineages and branches in a species tree. It is interesting to note that two of these methods, STRUCTURE and ADMIXTURE, that are popular choices for reconstructing genetic history and testing for admixture scenarios, were shown recently to be subject to misinterpretation because of ghost introgression as well (Lawson et al. 2018). The impact of ghost lineages on the detection of introgression thus goes much beyond the simple question of the tool used. It is a problem that we, as a community, are only starting to recognize.

4. Conclusion

The D-statistics is a key tool for studying introgression as it grants, in specific cases, a robust test for gene flow. However our results show that one important caveat of the test is the lack of consideration of ghost lineages, which can represent a major flaw when it comes to interpreting a significant result. Therefore the real message given by a single significant D-statistics is that, among the quartet of lineages, one lineage is the recipient of an introgression. But determining the actual pair of donor and recipient is not possible with a single test as both ingroup introgression and ghost introgression are valid hypothesis, the latter being possibly predominant as shown here.

As such we advocate a change of the null hypothesis for the D-statistics to integrate ghost introgression as full-fledge alternative hypothesis. It results that the common interpretation of an ingroup introgression for a significant D-statistics becomes the exception rather than the rule, particularly when the outgroup is selected distant from the ingroup, which is common practice.

4. Material and methods

4.1. Brown bear introgression

We used publicly available data from Barlow et al. (2018, 2019) to compute D-statistics. We did the test on two different quartets of species that differed by only one species (see trees in fig. 2B). The first one (Sampling 1) was composed of two brown bears, one from Alaska (also named ABC; with the identification Adm1 in Barlow et al. 2019) and one from Russia (id: 235 in Barlow et al. 2019), one Polar bear from Canada (id: NB in Barlow et al. 2019), and one American black bear (id: Uamericanus in Barlow et al. 2019) used as outgroup, following (Barlow et al. 2018). For the second sample (Sampling 2), the polar bear was replaced by a Swedish bear (id: Swe in Barlow et al. 2019) while the other species remained unchanged. In both cases we used the genomes of individuals mapped on the Giant Panda (Barlow et al. 2019) to generate multiple sequence alignments for the two quartets of species. D-statistics tests were then performed after counting the biallelic SNP patterns (ABBA and BABA) in those two alignments and applying the classical formula from Durand et al. (2011):

$$D = \frac{\Sigma ABBA - \Sigma BABA}{\Sigma ABBA + \Sigma BABA} \quad (1)$$

Using a custom R script, we computed the D-statistics for the pectinate trees presented in figure 2B: (((Alaska,Russia)Sweden)Black) and (((Alaska,Russia)Polar)Black). To test for significance, we applied the weighted block jackknife as proposed in (Green et al. 2010), using 1 Mb non-overlapping blocks. We considered the result significant if it was more than three standard errors from zero (Z > 3 or Z < 3), following (Green et al. 2010).

4.2. Simulations

4.2.1. Simulating species trees and introgressions

To test the effect of ghost lineages on the proportion of erroneous interpretation of the D-statistics, we chose the following approach, summarized in figure 6.

We simulated species trees with *N* extant taxa (N = 20, 40, 60, 80 and 100), using a simple birth-death process as implemented in the *rphylo* function of the *ape* package in R (Paradis and Schliep 2019) and keeping track of all the extinct taxa. The speciation rate was fixed to 1 and extinction rates (p_{ex}) varied between 0 and 0.8, with steps of 0.2. Out of the *N* extant taxa, 20 were sampled at the end, thus simulating a "sampling effect". This led to a proportion of known extant taxa between 0.2 (20 out of 100) and 1.0 (20 out of 20).

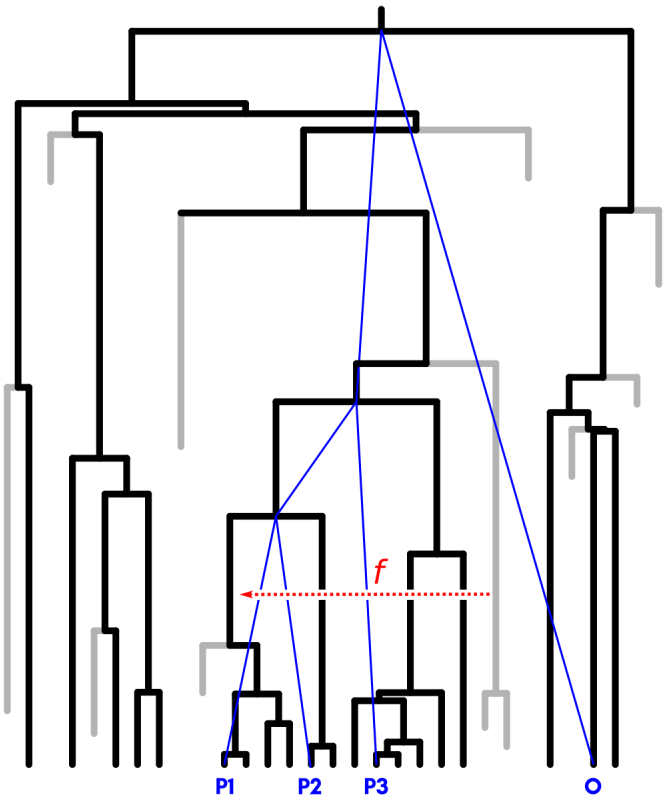

Fig. 6. Illustration of the data simulated for testing the impact of ghost lineages on the detection of introgression with the D-statistics. A species tree is simulated with a birth-death process (black and grey lines), an introgression is sampled (red dashed arrow), and for each possible quartet of species with topology (((P1,P2),P3),O) (blue lines and labels) the D-statistics is either computed (after simulation of loci, Approach 1), or inferred (Approach 2). If the D-statistics is not null, the type of introgression involved is characterized, either as *ingroup* if the introgression involves P1, P2, P3 or their close relatives as donor and receiver, or as *midgroup* if the introgression involves lineages

branching outside the ingroup as in this example. Introgression involving the *outgroup* and its close relatives and the lineages branching outside the 4-taxa subtree are simply discarded (see text).

For each set of parameters 100 species trees were simulated. Then for each complete species tree (containing all taxa, extinct and extant), 100 introgression events were simulated. The timing and the identity of the donor branch was sampled with a probability proportional to the number of branches available at each time, and the recipient of the transfer was subsequently sampled from all other branches available at the same time with probability

$$p_i = e^{-\alpha \overline{D}_i} \tag{2}$$

where $\overline{D}_i$ is the distance to the *i*-th recipient, normalized by the distances to all the possible recipients. When $\alpha = 0$, there is no effect of the phylogenetic distance on the choice of the recipient of the introgression. When α increases, introgressions are more likely to occur to closely related taxa, thus simulating a "preferential transfer effect". Values of α tested were 0, 1, 10, 100 and 1000. For each value of *N*, α and p_{ex}, 100 repetitions were performed.

4.2.2. Computing or inferring the D-statistics

Using the species trees and the introgressions simulated in the previous step (fig. 6), we devised two different strategies for evaluating the impact of ghost introgressions on the estimation of D-statistics. The first one is more accurate, as it simulates loci evolving in populations along the branches of the species tree but is very time and resource consuming. The second one is more lightweight as it relies solely on the topology and the branch lengths of the locus trees induced by the introgressions, and allows thus exploring a wider range of parameters.

Approach 1: We simulated species trees with a birth-death model using Zombi (Davín et al. 2020), with a custom python script, we converted species tree branch lengths in units of generation and fixed the age of the root of the trees to 10^6 generations. For each species tree and each of the 100 introgressions sampled (see previous step), we used *ms* (Hudson 2002) to simulate 10^6 independent loci, with a single mutation, evolving in populations of fixed size

(Ne) of 100,000 individuals. To simulate the introgression, a single event of migration was imposed over 1 generation for a fraction f = 50% of the donor population to invade the recipient population. The R package *Coala* (Staab and Metzler 2016) was used to execute *ms* and to analyse the resulting data.

For each quartet with topology (((P1,P2),P3),O) in the species tree and for each introgression simulated, we counted the number of loci with ABBA and BABA patterns. We then computed the D-statistics following equation (1) above and we performed a binomial test to evaluate whether the difference in the number of ABBA and BABA patterns was significant.

Approach 2: For each quartet with topology (((P1,P2),P3),O) in the species tree, we estimated whether its topology resulting from the simulated introgression would lead to a D-statistics significantly different from 0. These topologies are (((P3,P1),P2),O) and (((P3,P2),P1),O). Doing this we consider that every introgression that could theoretically be detected with the D-statistics is indeed detected.

For one set of parameters (N=40, $\alpha = 0$, $pex = 0.5$), the same species tree and the same 500 introgression events were used to compare the estimations of the proportion of erroneous interpretations for each quartet with the two approaches detailed above. These proportions were highly correlated (fig. 7), with a slight overestimation of the proportion of erroneous interpretations with Approach 1. We thus decided to use the more conservative Approach 2 for all the subsequent analyses, which was also much less time and resource consuming.

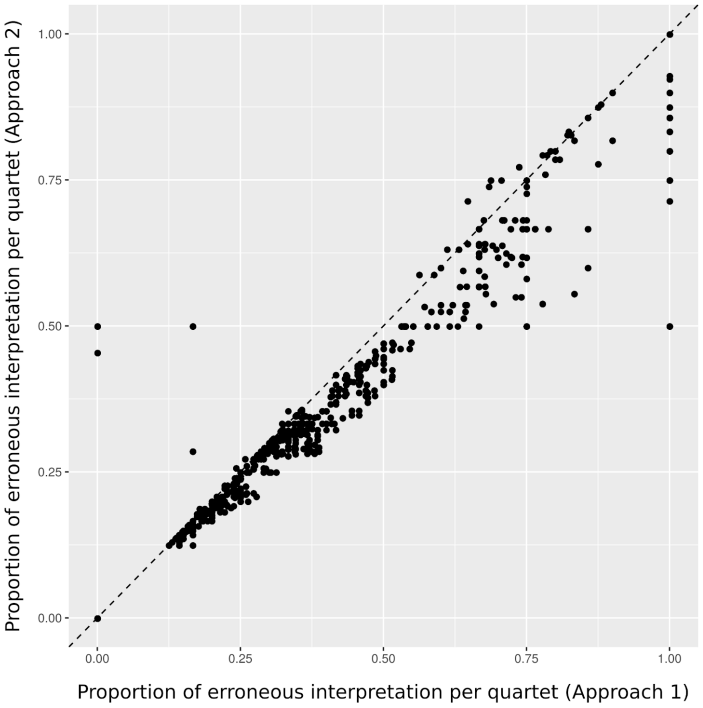

Fig. 7. Comparison of two possible approaches for estimating the proportion of erroneous interpretation of the D-statistics in this study : computation of the D-statistics based on simulated SNPs (Approach 1, x-axis) or simple inference based on the topology of the trees after introgression (approach 2, y-axis). The proportion of erroneous interpretation over 500 introgressions for a unique species tree (N=40, α=0, pex=0.5) was computed for each quartet (black dots). The dashed line represents the first diagonal.

4.2.3. *Evaluating the proportion of "ingroup"* vs *"midgroup" introgressions and estimating the proportion of erroneous interpretation*

For both approaches, we stored the quartets that were significantly different from 0 (or considered as such) and we looked in each case whether the introgression associated to this result was involving a lineage branching in the midgroup (referred to as a ghost introgression from now on) or a lineage within the ingroup. For those quartet we compute the distance to the outgroup using the ratio $R=t1/t2$ (see fig. 3).

Quartets for which the outgroup was involved in the introgression event were discarded on the basis that the outgroup is usually selected distant enough so as to not be involved in gene flows with the ingroup. For the same reason we discarded quartets for which the donor was

outside the four selected taxon. Note that such introgressions can lead to significant D-statistics that would also be erroneously interpreted (supplementary fig. S4). Ignoring them is thus conservative. For each simulation we computed the proportion of erroneous interpretations, i.e. the number of significant D-statistics as result of ghost introgressions over the sum of significant D-statistics (considering only ingroup and ghost introgressions).

Acknowledgments

This work was supported by the French National Research Agency (Grant ANR-18-CE02-0007-01). Simulations were performed using the computing facilities of the CC LBBE/PRABI.

Software Availability

All codes used to generate and analyze simulations performed in this study are available at: https://github.com/theotricou/Ghost_abba_baba.

References

Barlow A, Cahill JA, Hartmann S, Theunert C, Xenikoudakis G, Fortes GG, Paijmans JLA, Rabeder G, Frischauf C, Grandal-d'Anglade A, et al. 2018. Partial genomic survival of cave bears in living brown bears. *Nat. Ecol. Evol.* 2:1563–1570.

Barlow A, Cahill JA, Hartmann S, Theunert C, Xenikoudakis G, Fortes GG, Paijmans JLA, Rabeder G, Frischauf C, Grandal-D'Anglade A, et al. 2019. Data from: Partial genomic survival of cave bears in living brown bears. :6331386741 bytes. Available from: http://datadryad.org/stash/dataset/doi:10.5061/dryad.cr1496b

Cahill JA, Green RE, Fulton TL, Stiller M, Jay F, Ovsyanikov N, Salamzade R, John JS, Stirling I, Slatkin M, et al. 2013. Genomic Evidence for Island Population Conversion Resolves Conflicting Theories of Polar Bear Evolution. *PLOS Genet.* 9:e1003345.

Chapman AD, Australia, Department of the Environment W Heritage, and the Arts, Australian Biological Resources Study. 2009. Numbers of living species in Australia and the world. Canberra, A.C.T.: Department of the Environment, Water, Heritage and the Arts

Chapman MA, Burke JM. 2007. Genetic divergence and hybrid speciation. *Evol. Int. J. Org. Evol.* 61:1773–1780.

Dannemann M, Racimo F. 2018. Something old, something borrowed: admixture and adaptation in human evolution. *Curr. Opin. Genet. Dev.* 53:1–8.

Davín AA, Tricou T, Tannier E, de Vienne DM, Szöllősi GJ. 2020. Zombi: a phylogenetic simulator of trees, genomes and sequences that accounts for dead linages.Schwartz R, editor. *Bioinformatics* 36:1286–1288.

Deren A. R. Eaton, Ree RH. 2013. Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Syst. Biol.* 62:689–706.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. *Mol. Biol. Evol.* 28:2239–2252.
Eaton DAR, Hipp AL, González‐Rodríguez A, Cavender‐Bares J. 2015. Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution* 69:2587–2601.
Edmands S. 2002. Does parental divergence predict reproductive compatibility? *Trends Ecol. Evol.* 17:520–527.
Eriksson A, Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl. Acad. Sci.* 109:13956–13960.
Foote AD, Martin MD, Louis M, Pacheco G, Robertson KM, Sinding M-HS, Amaral AR, Baird RW, Baker CS, Ballance L, et al. 2019. Killer whale genomes reveal a complex history of recurrent admixture and vicariance. *Mol. Ecol.* 28:3427–3444.
Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. *Philos. Trans. R. Soc. B Biol. Sci.* 363:4023–4029.
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MHY, et al. 2010. A Draft Sequence of the Neandertal Genome. *Science* 328:710–722.
Hahn MW, Hibbins MS. 2019. A Three-Sample Test for Introgression. :5.
Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* 22:4606–4618.
Hibbins M, Hahn M. 2021. Phylogenomic approaches to detecting and characterizing introgression. Available from: https://ecoevorxiv.org/uahd8/
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
Jente Ottenburghs, Megens H-J, Kraus RHS, van Hooft P, van Wieren SE, Crooijmans RPMA, Ydenberg RC, Groenen MAM, Prins HHT. 2017. A history of hybrids? Genomic patterns of introgression in the True Geese. *BMC Evol. Biol.* 17:201.
Keuler R, Garretson A, Saunders T, Erickson RJ, Andre NS, Grewe F, Smith H, Lumbsch HT, Huang J-P, Clair LLS, et al. 2020. Genome-scale data reveal the role of hybridization in lichen-forming fungi. *Sci. Rep.* 10:1497.
Kumar V, Lammers F, Bidon T, Pfenninger M, Kolter L, Nilsson MA, Janke A. 2017. The evolutionary history of bears is characterized by gene flow across species. *Sci. Rep.* 7:46487.
Lan T, Cheng J, Ratan A, Miller W, Schuster SC, Farley S, Shideler RT, Mailund T, Lindqvist C. 2016. Genome-wide evidence for a hybrid origin of modern polar bears. *bioRxiv*:047498.
Lawson DJ, van Dorp L, Falush D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* 9:3258.
Leo Elworth RA, Allen C, Benedict T, Dulworth P, Nakhleh L. 2018. Dgen: A Test Statistic for Detection of General Introgression Scenarios. Evolutionary Biology Available from: http://biorxiv.org/lookup/doi/10.1101/348649
Liu L, Bosse M, Megens H-J, Frantz LAF, Lee Y-L, Irving-Pease EK, Narayan G, Groenen MAM, Madsen O. 2019. Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. *Nat. Commun.* 10:1992.
Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen TS, Somel M, Babbitt C, et al. 2014. Population Genomics Reveal Recent Speciation and

Rapid Evolutionary Adaptation in Polar Bears. *Cell* 157:785–794.
Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci.* 113:5970–5975.
Lohse K, Frantz LAF. 2014. Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes. *Genetics* 196:1241–1251.
Maddison WP. 1997. Gene Trees in Species Trees. *Syst. Biol.* 46:523–536.
Maddison WP, Knowles LL. 2006. Inferring Phylogeny Despite Incomplete Lineage Sorting. *Syst. Biol.* 55:21–30.
Mallet J. 2005. Hybridization as an invasion of the genome. *Trends Ecol. Evol.* 20:229–237.
Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.* 23:1817–1828.
Massardo D, VanKuren NW, Nallu S, Ramos RR, Ribeiro PG, Silva-Brandão KL, Brandão MM, Lion MB, Freitas AVL, Cardoso MZ, et al. 2020. The roles of hybridization and habitat fragmentation in the evolution of Brazil's enigmatic longwing butterflies, Heliconius nattereri and H. hermathena. *BMC Biol.* 18:84.
Meier JI, Marques DA, Mwaiko S, Wagner CE, Excoffier L, Seehausen O. 2017. Ancient hybridization fuels rapid cichlid fish adaptive radiations. *Nat. Commun.* 8:14363.
Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prüfer K, Filippo C de, et al. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science* 338:222–226.
Montanari SR, Hobbs J-PA, Pratchett MS, Bay LK, Van Herwerden L. 2014. Does genetic distance between parental species influence outcomes of hybridization among coral reef butterflyfishes? *Mol. Ecol.* 23:2757–2770.
Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. 2011. How Many Species Are There on Earth and in the Ocean?Mace GM, editor. *PLoS Biol.* 9:e1001127.
Ottenburghs J. 2020. Ghost Introgression: Spooky Gene Flow in the Distant Past. *BioEssays* n/a:2000012.
Paradis E, Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35:526–528.
Pease JB, Haak DC, Hahn MW, Moyle LC. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLOS Biol.* 14:e1002379.
Pease JB, Hahn MW. 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Syst. Biol.* 64:651–662.
Pickrell JK, Pritchard JK. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genet.* 8:e1002967.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155:945–959.
Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, de Filippo C, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505:43–49.
Raup DM. 1991. Extinction: bad genes or bad luck? New York: W.W. Norton
Rouard M, Droc G, Martin G, Sardos J, Hueber Y, Guignon V, Cenci A, Geigle B, Hibbins MS, Yahiaoui N, et al. 2018. Three New Genome Assemblies Support a Rapid Radiation in Musa acuminata (Wild Banana). *Genome Biol. Evol.* 10:3129–3140.
Schumer M, Cui R, Powell DL, Rosenthal GG, Andolfatto P. 2016. Ancient hybridization and genomic stabilization in a swordtail fish. *Mol. Ecol.* 25:2661–2679.
Simon H. Martin, Davey JW, Jiggins CD. 2015. Evaluating the Use of ABBA–BABA

Statistics to Locate Introgressed Loci. *Mol. Biol. Evol.* 32:244–257.
Smith J, Kronforst MR. 2013. Do Heliconius butterfly species exchange mimicry alleles? *Biol. Lett.* 9:20130503.
Soraggi S, Wiuf C, Albrechtsen A. 2018. Powerful Inference with the D-Statistic on Low-Coverage Whole-Genome Data. *G3 GenesGenomesGenetics* 8:551–566.
Staab PR, Metzler D. 2016. Coala: an R framework for coalescent simulation. *Bioinformatics* 32:1903–1904.
Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Wadell P, Lang M, Courtier-Orgogozo V, et al. 2020. Widespread introgression across a phylogeny of 155 Drosophila genomes. *bioRxiv*:2020.12.14.422758.
Szöllősi GJ, Davín AA, Tannier E, Daubin V, Boussau B. 2015. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140335.
Szöllősi GJ, Tannier E, Lartillot N, Daubin V. 2013. Lateral Gene Transfer from the Dead. *Syst. Biol.* 62:386–397.
Tang H, Peng J, Wang P, Risch NJ. 2005. Estimation of individual admixture: Analytical and study design considerations. *Genet. Epidemiol.* 28:289–301.
Than C, Ruths D, Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9:322.
Wen D, Yu Y, Zhu J, Nakhleh L. 2018. Inferring Phylogenetic Networks Using PhyloNet. *Syst. Biol.* 67:735–740.
Yichen Zheng, Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinformatics* 19:10.
Zhang B-W, Xu L-L, Li N, Yan P-C, Jiang X-H, Woeste KE, Lin K, Renner SS, Zhang D-Y, Bai W-N. 2019. Phylogenomics Reveals an Ancient Hybrid Origin of the Persian Walnut. *Mol. Biol. Evol.* 36:2451–2461.
Zhang W, Zhang X, Li K, Wang C, Cai L, Zhuang W, Xiang M, Liu X. 2018. Introgression and gene family contraction drive the evolution of lifestyle and host shifts of hypocrealean fungi. *Mycology* 9:176–188.