bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Ghost lineages highly influence the interpretation of introgression tests

**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**

Most species are extinct; those that are not are often unknown. Sequenced and sampled species are often a minority of known ones. Past evolutionary events involving horizontal gene flow, such as horizontal gene transfer, hybridization, introgression and admixture, are therefore likely to involve "ghosts", *i.e.* extinct, unknown or unsampled lineages. The existence of these ghost lineages is widely acknowledged, but their possible impact on the detection of gene flow and on the identification of the species involved is largely overlooked. It is generally considered as a possible source of error that, with reasonable approximation, can be ignored. We explore the possible influence of absent species on an evolutionary study by quantifying the effect of ghost lineages on introgression as detected by the popular D-statistic method. We show from simulated data that under certain frequently encountered conditions, the donors and recipients of horizontal gene flow can be wrongly identified if ghost lineages are not taken into account. In particular, having a distant outgroup, which is usually recommended, leads to an increase in the error probability and to false interpretations in most cases. We conclude that introgression from ghost lineages should be systematically considered as an alternative possible, even probable, scenario.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 1. Introduction

Evolutionary studies are always restricted to a subset of species, populations or individuals. This is by choice, because only a fraction of the data is relevant to the question being addressed, and by necessity, because the approaches used have methodological and technical limitations. Another reason is that most lineages are simply unknown. More than 99.9% of all species that have ever lived are now extinct (Raup 1991) and only a small fraction of extant species have been described. The number of extant eukaryote species that are still uncatalogued is almost an order of magnitude higher than the number of those reported (~1.3 million species have been catalogued, Mora et al. 2011), and is many orders of magnitudes higher if we consider Bacteria and Archaea diversity (Locey and Lennon 2016).

Taking these extinct, unknown or unsampled "ghost" lineages into account is particularly important when studying introgression, *i.e.* the integration of genetic material from one lineage to another *via* hybridization and subsequent backcrossing. This mode of gene flow across species boundaries appears to be common in the Eukaryotic domain and has been shown to be adaptive in some cases (see for example Hedrick 2013 for a review). Introgression has been reported in such diverse lineages 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 (Eaton and Ree 2013; Zhang et al. 2019) and fungi (Zhang et al. 2018; Keuler et al. 2020), to name but a few.

Since ghost lineages are probably massively present around any phylogeny of extant species, many gene flow events that are detectable now are likely to have involved a ghost lineage. This has been repeatedly acknowledged (Maddison 1997; Galtier and Daubin 2008; Green et al. 2010; Eaton and Ree 2013; Szöllősi et al. 2013, 2015), especially in studies of introgression between populations, but it was considered either a source of noise (Pease and Hahn 2015), or a problem that could be resolved by adding new species as they become available, or by combining the results of multiple detection tests (Eaton et al. 2015; Kumar et al. 2017; Barlow et al. 2018). Recently, Hibbins and Hahn (2021) advised bearing ghost lineages "in mind" when investigating gene flow but, as far as we know, the real impact of ghost lineages on the ability of different methods to detect gene flow and correctly identify involved lineages has not been properly evaluated and quantified.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Over the past few years, the ever-growing number of sequenced genomes and the development of new methods have improved the detection of introgression. One of the most widely used methods for inferring introgression is the D-statistic (or Patterson's D), also known as the ABBA-BABA test (Kulathinal et al. 2009; Green et al. 2010; Durand et al. 2011; Patterson et al. 2012). There are many reasons for its success. The D-statistic is easy to understand and implement, quick to compute and easy to interpret. This method is based on phylogenetic discordance and can discriminate incongruence caused by Incomplete Lineage Sorting (ILS) from incongruence caused by gene flow (Kulathinal et al. 2009; Green et al. 2010; Durand et al. 2011; Patterson et al. 2012). The ABBA-BABA test considers four taxa: three ingroup taxa and one outgroup, with a ladder-like phylogenetic relationship (Fig. 1). The test relies on counts of the number of sites that support a discordant topology. Two biallelic SNP patterns are considered, ABBA and BABA, depending on which allele (A: ancestral, B: derived) is present in each taxon. The D-statistic is computed using the classic formula from Durand et al. (2011):

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

The null hypothesis states that under a scenario with no gene flow, both ABBA and BABA patterns can be attributed to ILS and thus should be observed in equal numbers. Significant deviation from this expectation, resulting in a D-statistic significantly different from zero, is usually interpreted as introgression between two of the three lineages forming the ingroup (Fig. 1, left panel). The outgroup should be distant enough from the ingroup such that it is not involved in an introgression with any of the ingroup lineages (Green et al. 2010; Osborne et al. 2016; Irwin et al. 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

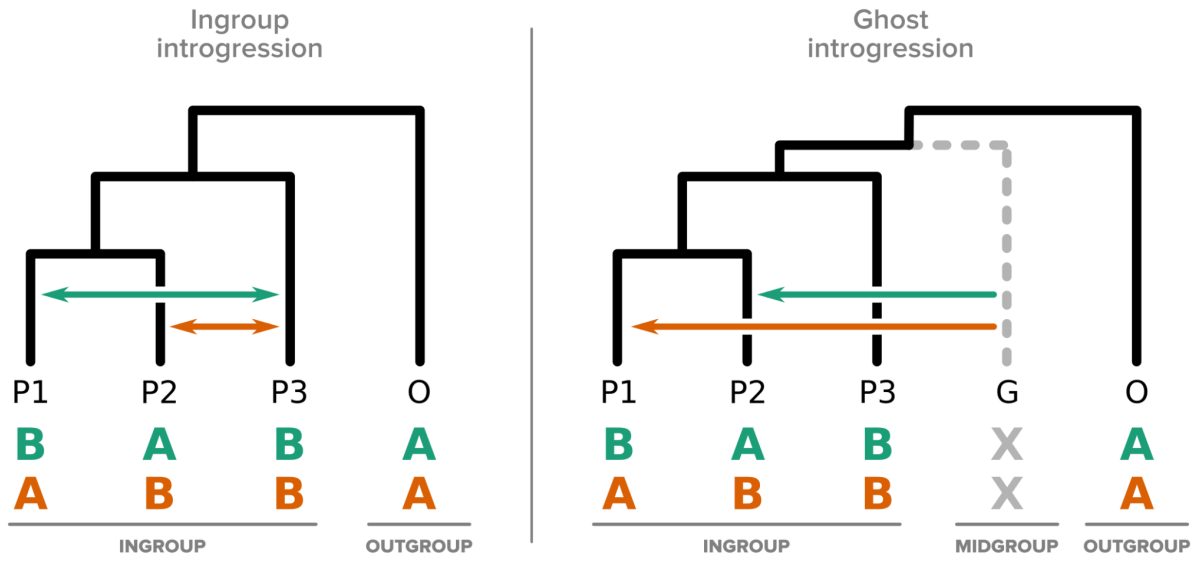


**Figure 1.** Introgression events that can result in a significant excess of ABBA or BABA patterns according to the D-statistic. The usual interpretation of this excess is the hypothesis of "ingroup" introgression (left panel). However "ghost" (or "midgroup") introgression (right panel) from ghost lineages (G) can produce similar patterns.

Undersampling is known to be one of the factors that can possibly confound the D-statistic (Martin et al. 2015; Zheng and Janke 2018), and affect the detection of introgression. This is because using a subset logically leads to an underestimation of the true frequency of introgression and thus inflates the role of ILS (Maddison and Knowles 2006). It has been clearly stated that the donor genome could easily be misidentified because introgression from a sampled lineage (*e.g.* P3) or from one sister ghost lineage to the same recipient lineage would produce the same signal and result in indistinguishable D-statistic results (Eaton and Ree 2013; Eaton et al. 2015; Pease and Hahn 2015; Zheng and Janke 2018). Another stronger impact of ghost lineages, however, was foreseen early on in the history of the test (by Durand et al. (2011), in their first description of the test), but has been largely overlooked afterwards: introgression from a ghost lineage 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, right panel). Under this scenario, none of the species thought to be involved in the introgression event are correctly identified. This possible source of error in the interpretation of the D-statistic has often been acknowledged (Durand et al. 2011; Ottenburghs et al. 2017;

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Zheng and Janke 2018; Hibbins and Hahn 2021) but surprisingly, it does not seem to have changed the way the test is commonly interpreted, perhaps because it is thought that the impact of this possibility is low, even though it has not been formally quantified. This is the goal of this study.

We begin by an illustration of the possibility of misinterpreting the ABBA-BABA test when some species are unknown or not included using a previously published bear phylogeny, recurrently used later on to estimate parameters on realistic situations. We then quantify the effect of ghost lineages on the misidentification of the donor and the recipient lineage using simulations. We explore the impact of outgroup choice, number of unsampled species and genetic divergence between introgressed taxa on the probability of misinterpreting introgression events.

We show that under the realistic assumption that there are many ghost lineages branches in the tree, and assuming a simple demographic history of the populations considered, most significant D-statistics are attributable to ghost lineages. This suggests that most of the lineages involved (donors and recipients) are incorrectly identified by the usual interpretation of D-statistics. The error rate increases with the distance between ingroup and outgroup, even though the outgroup is usually chosen so that its distance from the ingroup is sufficient to avoid any introgression between the two (Green et al. 2010; Osborne et al. 2016; Irwin et al. 2018). This observation, that a close outgroup as well as a distant outgroup is a source of interpretation error, hampers the delimitation of a safe zone for the interpretation of the D-statistic.

These results call for a new way of interpreting D-statistics, and more generally call into question established methods of introgression detection. Our results illustrate the recent statement by Ottenburghs (2020) that "the presence of ghost introgression has important consequences for the study of evolutionary processes", and provide a demonstration of this importance.

# 2 Materials and Methods

## 2.1 Bear genomic dataset

We use the dataset from Barlow et al. (2018) to illustrate the possibility of misinterpreting significant results of an ABBA-BABA test. This dataset has the advantage of being easily

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

available and to support, according to the authors, a simple introgression scenario, *i.e.* a documented introgression between polar bears and brown bears from the ABC Islands in Alaska. We downloaded the genome sequences (Barlow et al. 2019) of three brown bears (*Ursus arctos*) from Alaska (id: Adm1), Russia (id: 235) and Slovenia (id: 191Y), one polar bear (*U. maritimus*; id: NB) and one American black bear (*U. americanus*; id: Uamericanus), all aligned against the panda reference genome (Li et al. 2010). Their relationship is ((((Alaska = P1, Russia = P2), Slovenia = P3), Polar bear = P4), Black bear = O). Using scripts available from the GitHub repository of Barlow et al. (2018) https://github.com/jacahill/Admixture, we computed the D-statistic for two quartets: (((Alaska,Russia),Polar bear),Black bear) and (((Alaska,Russia),Slovenia),Black bear). We used the script for all sites (not transversions only like in Barlow et al, 2018) as we do not have any archaic species in the dataset. The weighted block jackknife from the script was used to compute the Z-score in non-overlapping 1 Mb windows with a script available on the GitHub repository mentioned above. We considered the result to be significant if it was more than three standard deviations from zero (Z > 3 or Z < -3), as *per* Green et al. (2010).

### 2.2 Species tree and gene tree simulation

Species trees were simulated using the birth-death simulator implemented in the R function *rphylo* from the *ape* package (Paradis and Schliep 2019). Speciation rate was fixed at 1, extinction rate at 0.9 and the simulation was stopped when N extant lineages, varying in {20, 40, 60, 80, 100}, were present in the species tree (step 1 in Figure 2). Then 20 taxa were uniformly sampled from the N taxa. An introgression event was chosen in the species tree (including unsampled lineages), with a donor and a recipient. The donor branch was selected among the branches of the species tree with a probability proportional to its length, and the time of introgression uniformly at random on this branch. The recipient was randomly sampled among the lineages present at the time of the introgression, with a probability that decreases exponentially with the phylogenetic distance from the donor (step 2 in Figure 2):

$$p_i = e^{-\alpha \underline{D}_i} \qquad (2)$$

Where $\underline{D}_i$ is the distance from the *i*-th recipient, normalized by the distances from all possible recipients, and α is a parameter in {0, 1, 10, 100, 1000}. With $\alpha = 0$, introgressions

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

occur between any contemporaneous branches on a tree with equal probability. When $\alpha$ increases, introgressions are more likely to occur between closely related taxa. If the recipient is a branch with no extant offspring, then the introgression cannot be detected, so only introgressions such that the recipient has descendants among the 20 remaining species were kept. After setting the donor and recipient, a gene tree was generated with subtree pruning and regrafting (SPR) (Bordewich and Semple 2005), simulating the introgression event (step 3 in Figure 2). All unsampled lineages are then pruned from species and gene trees.

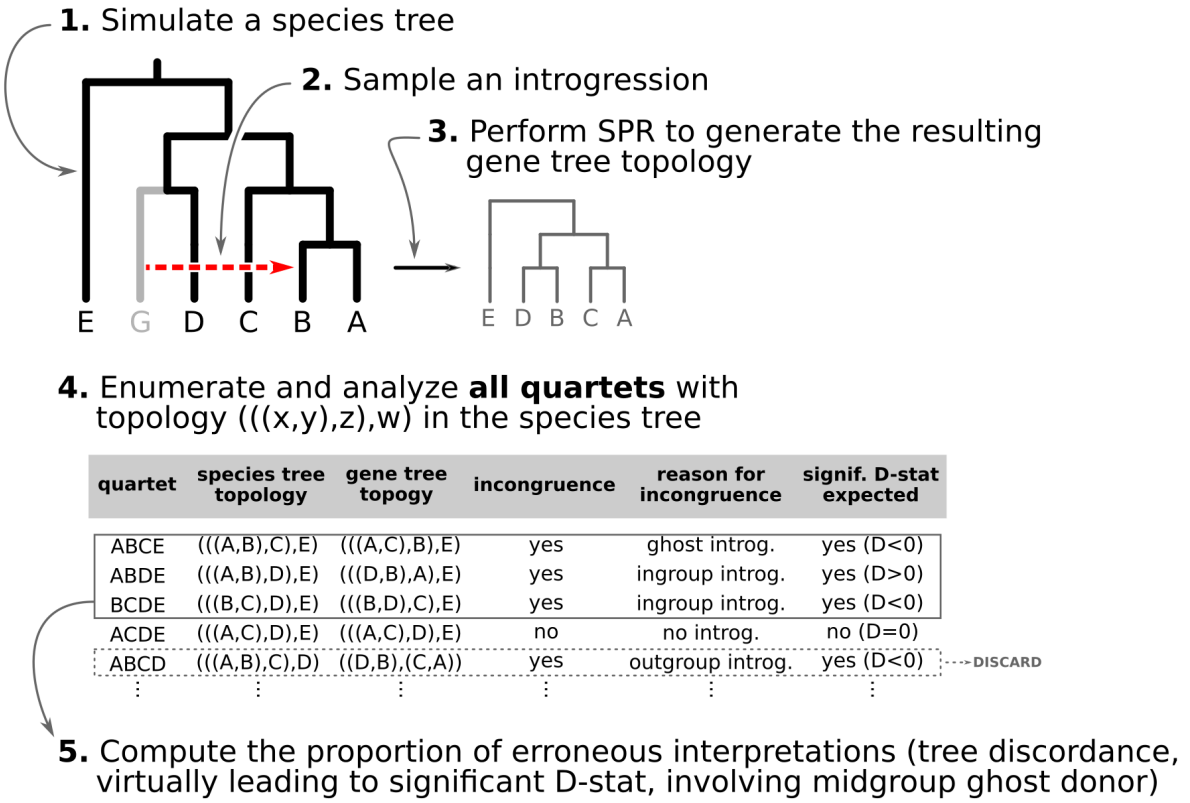


**Figure 2.** Species tree/Gene tree simulation: (1) a species tree is generated under a birth death model and 20 taxa are sampled from it; (2) an introgression event is picked from a random donor and recipient; (3) an introgressed gene tree is constructed from the species tree by SPR; (4) for each quartet with a ladder-like topology (((x,y),z),w) in the species tree, species tree and gene tree topologies are compared to determine if there is an incongruence caused by the introgression; (5) the proportion of erroneous interpretation of the D-statistic across the species tree is computed by the sum of all introgressions with a midgroup ghost donor over all introgressions detected, outgroup introgressions excluded.

### 2.3 Comparison of the species tree and the gene tree as proxy for the D-statistic

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For each gene tree/species tree pair, we counted all species quartets with a ladder-like topology (((P1,P2),P3),P4) in the species tree and (((P1,P3),P2),P4) or (((P2,P3),P1),P4) in the gene tree. These configurations were interpreted as yielding significant D-statistics. This avoids the computational burden of simulating sequences for a high number (250 000, see Section 2.5) of cases and gives reasonably equivalent results (Supplementary Material Section 1).

We then counted the number of situations where the result was correctly interpreted (the simulated introgression is between extant lineages in this quartet) or misinterpreted (the simulated introgression is from ghost lineage) (step 5 in Figure 2).

### 2.4 Measuring the distance to the outgroup

For each species quartet we computed the distance between outgroup and ingroup using the ratio $R=t1/t2$, where $t1$ is the distance (sum of branch lengths) between the most recent common ancestor of the ingroup and the most recent common ancestor of all four taxa (see $t1$ in Fig. 5), and $t2$ is the total height of the four-taxon tree (see $t2$ in Fig. 5). To correlate this distance with the rate of interpretation error of D-statistics, we identified ten intersecting subsets of quartets according to their $R$ value: subsets for which $R$ is higher than a threshold "x", with "x" varying from 0 to 0.9 with a step of 0.1. We computed the rate of interpretation error for each of the 10 subsets.

### 2.5 Simulation dataset

We simulated, for each value of N in {20, 40, 60, 80, 100} and $\alpha$ in {0, 1, 10, 100, 1000}, 100 species trees with N species, and for each species tree, 100 independent gene trees with independent introgression events. For each gene tree/species tree pair, 20 species were uniformly sampled from N (extant species), and the rest were pruned, resulting in 250,000 pairs of trees each with 20 leaves.

## 3 Results

### 3.1 Bear phylogeny exemplifies the problem of interpreting the D-statistic without taking unsampled lineages into account

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Using genomic data, we show how the presence or absence of one lineage, in this case the polar bear, can lead to opposite interpretations of the D-statistic if interpreted without considering ghost lineages. From the bear phylogeny (Material and Methods; phylogeny shown in Figure 3A), we removed either the Slovenian bear (Figure 3B, subset 1), which is not thought to have introgressed with other bear species from the ingroup, or the polar bear (subset 2), which is suspected to have introgressed with brown bears from Alaska (Cahill et al. 2013; Liu et al. 2014; Lan et al. 2016; Kumar et al. 2017; Barlow et al. 2018). In the first subset, we identified 175,413 ABBA patterns and 226,992 BABA patterns resulting in a significant negative D-score (D = -0.128, Z = -11.71), which is congruent with introgression between polar bears and Alaskan bears (Fig. 3B) in the usual interpretation of the test. In the second subset, we identified 266,173 ABBA patterns and 213,830 BABA patterns resulting in a significant positive D-score (D = 0.109, Z = 14.24).

If we were to interpret this second result as evidence of introgression between the lineages sampled here (Alaskan, Russian, Slovenian and black bears), we would conclude that there was introgression between bears from Slovenia and Russia (Figure 3B), even though this significant positive D-statistic could also be attributed to introgression between polar bears (not sampled here) and Alaskan bears. The latter attribution, however, relies on our knowledge of the existence of polar bears, considered a ghost lineage. This hypothesis could similarly be called into question if we knew the existence of another lineage, because we can never assume that we know all the lineages leading to extant or extinct species. Thus, even with good taxonomic sampling, there is a real chance that an interpretation based only on known lineages wrongly infers introgression events.

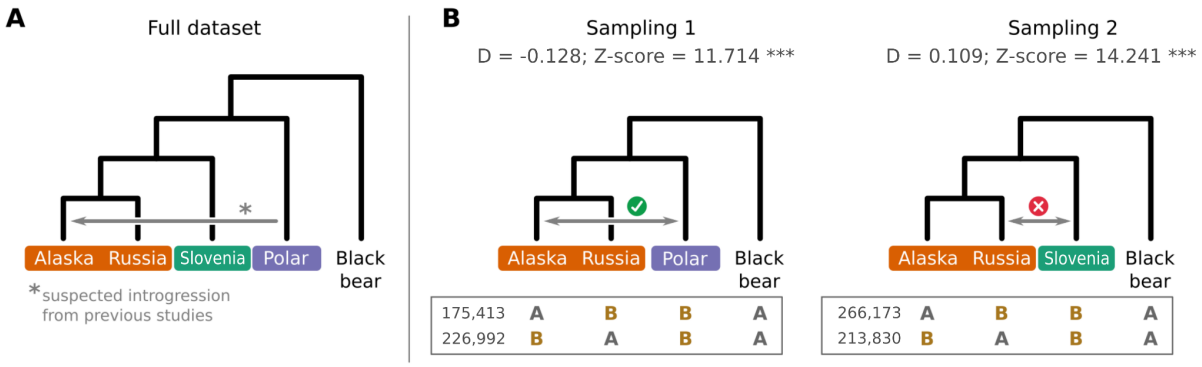


**Figure 3.** The effect of sampling on the interpretation of the D-statistic, using bear genomic

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

data as an example. **A**. Phylogenetic relationship of the five bear taxa sampled. The grey arrow shows the introgression inferred from previous studies. **B**. D-statistic calculated from two four-taxon subsets. The number of ABBA and BABA patterns is given below the trees. In subset 1, the Slovenian bear, a lineage that is not thought to be involved in introgression, is removed. In subset 2, the donor of the introgression shown in 3A (*i.e.* the polar bear) is removed. Introgressions were inferred from the D-statistic (grey arrows), and their congruence with other studies (green tick = congruent, red cross = not congruent) is indicated above the arrow.

### 3.2 Significant D-statistics are often due to introgressions from ghost lineages

Using simulated datasets (Material and Methods), we estimated the frequency of misinterpreting introgression events. We counted the number of D-statistics due to midgroup ghost introgressions (corresponding to the *proportion of erroneous interpretations*). We observed between 15% and 100% of erroneous interpretations, the frequency of which increased with (i) the proportion of unsampled lineages, (ii) the distance between ingroup (P1, P2, P3) and outgroup O, and (iii) the probability of introgression between distantly related lineages (Supplementary Material Section 2 for a complete summary). We describe these three trends in detail in the following sections and relate the range of each parameter we used to biological data such as the bear genomes described above.

#### *3.2.1. The proportion of unsampled species*

The effect of absent lineages on the interpretation of the D-statistic was investigated using simulated species trees with N = {20, 40, 60, 80, 100} from which 20 species were randomly sampled. This corresponds to a sampling effort ranging from 100% (20 species out of 20) to 20% (20 species out of 100). We observed that low sampling contributes to an increase in the number of misinterpreted D-statistics due to ghost introgression (Fig. 4). While the mean proportion of erroneous interpretations is ~25% when 100% of extant lineages are sampled, it is close to 60% when only 20% of extant lineages are sampled.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

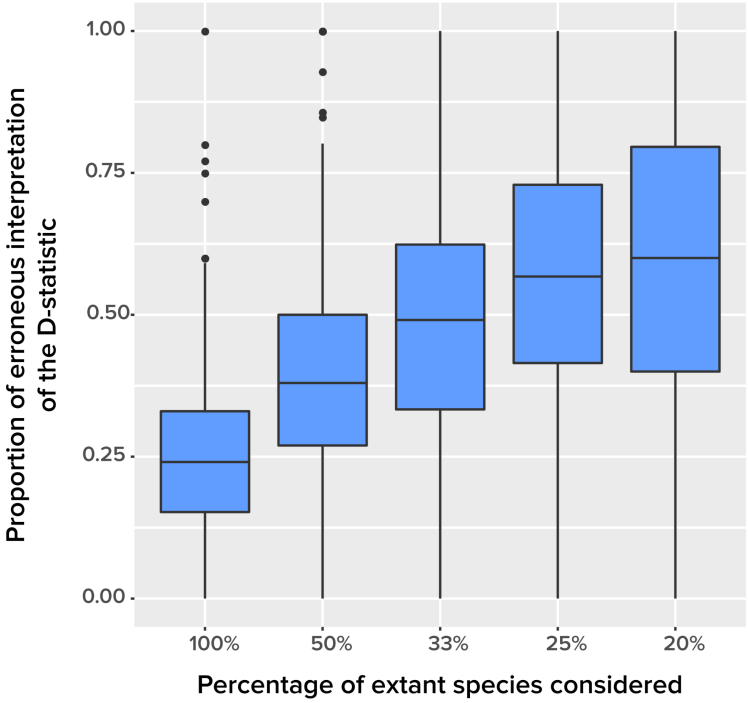


**Figure 4.** The effect of taxonomic sampling (x-axis) on the proportion of erroneous interpretations of the D-statistic (y-axis). The error rate is increasing with the amount of unknown.

For example, to study introgression in bears, Barlow et al. (2018) sampled 13 *Ursus* species and one *Ailuropoda* species, both members of the family Ursidae, based on the availability of genomic data. However, the Global Biodiversity Information Facility (gbif.org) reports 140 species in Ursidae, 100 of them belonging to the genus *Ursus*, which is close to the highest error rate in the simulations.

By contrast, and unlike we initially expected, we found no correlation between the number of extinct lineages in the tree and the proportion of erroneous interpretations of the D-statistic (Supplementary Material Section 3). Our interpretation is that increasing the number of extinct lineages is achieved by increasing the probability of extinction in the birth-death process, which also increases the probability that mid-group lineages, the possible source of ghost introgressions, become extinct before having the opportunity to introgress. Further investigations are needed to better characterize this effect.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 3.2.2 The distance between outgroup and ingroup

In ABBA-BABA tests, the outgroup is usually chosen so that its distance from the ingroup is sufficient to minimize the chance of introgression between the two (Green et al. 2010; Osborne et al. 2016; Irwin et al. 2018). Zheng and Janke (2018) stated that the distance between outgroup and ingroup had little to no impact on the sensitivity of the D-statistic. However, they focused on evaluating the effect of saturation of sequence substitutions in the outgroup and did not consider possible introgressions from mid- or outgroups.

From our simulations, we observed that the proportion of ghost introgressions (leading to erroneous interpretations) increased with *R*, the relative distance to the outgroup (see Materials and Methods). On average, when $R > 0.3$, more than 50% of the significant D-statistics are associated with ghost introgressions (Fig. 5). We found that, when $R > 0.7$, a median of 100% of D-statistics resulted from ghost introgressions.

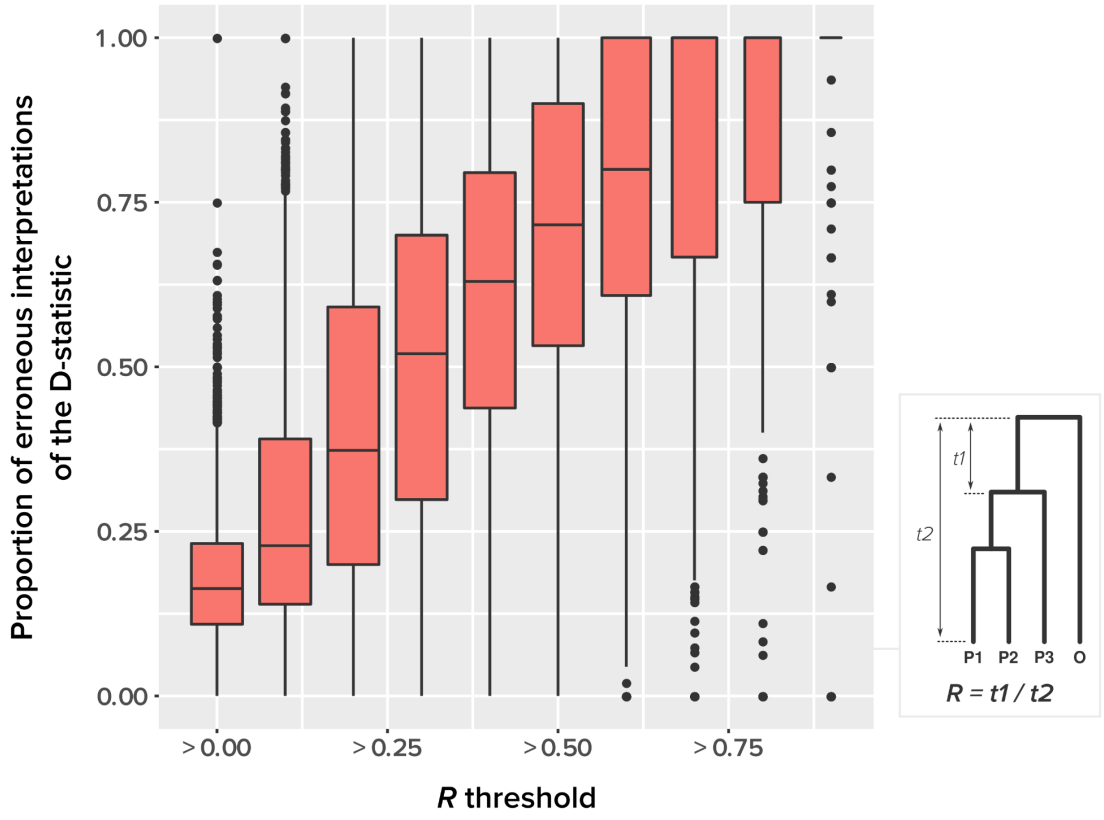


**Figure 5.** The relationship between outgroup distance (*R*) and the proportion of erroneous interpretations of the D-statistic for different thresholds of *R* (x-axis). The distances *t1* and *t2* used to calculate the relative distance to the outgroup (*R*) are described on the right box.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To relate our findings to different biological data, Green et al. 2010 used the D-statistic to detect introgression between Neanderthal and modern humans and had a *R* value equal to 0.873: 825,000 years separated modern humans from Neanderthal, while 6.5 million years separated humans from chimps (the outgroup) (Green et al. 2010). The study of bears of Barlow et al. (2018) used two different outgroups, black bears and pandas, with *R* values *ca*. 0.4 and 0.9, respectively. According to our simulations, all these values fall within the range of high probability of erroneous interpretation.

### 3.2.3 The distance between donor and recipient

Species that are genetically close have a higher chance to introgress, which could mitigate the previous result. Indeed, if the distance between outgroup and ingroup is sufficient to prevent introgression between the two, then putative midgroups ghost lineages may also be too distant. It is well known that the probability of hybridization, and consequently of introgression, decreases as genetic distance between species increases (Edmands 2002; Mallet 2005; Chapman and Burke 2007; Montanari et al. 2014).

To test whether this observation mitigates the importance of ghost lineages when detecting introgression, we used different values of $\alpha$, a parameter that lets the probability of introgression vary with the phylogenetic distance between donor and recipient (see Material and Methods). We used $\alpha = \{0, 1, 10, 100, 1000\}$. When $\alpha = 0$, introgressions occur uniformly at random; when $\alpha = 1000$, introgressions occur almost exclusively between sister taxa (Figure 6A).

We observed that the impact of outgroup distance on the proportion of erroneous interpretations decreased with increasing values of $\alpha$ (Figure 6B). As expected, in simulations where alpha is maximum ($\alpha = 1000$), the proportion of significant D-statistics due to ghost introgressions is not affected by the distance separating ingroup and outgroup. Nevertheless, this proportion remains quite high, and its median value does not fall below 25% under our settings. For other values of $\alpha$, this proportion is higher and increases with distance to the outgroup.

We investigated what could be a realistic value of alpha in biological data. We used phylogenetic trees from several studies where one or several introgression paths were identified. We counted the number of internal nodes between donor and recipient. Then, we randomly simulated the same number of introgressions with different values of alpha and

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

calculated the average number of internal nodes between donor and recipient (removing introgressions between sister branches as this cannot be observed when using D-statistic). We retained the value of alpha giving the average number of internal nodes that was closest to what is observed in the biological data (Supplementary Material Section 4). We analyzed the bear phylogeny of Hailer et al. (2012) and found that $\alpha = 0$ gives the closest result (actual number of nodes = 5; with $\alpha = 0$, the average number of nodes from the simulations was 4.5). In the phylogeny of the *Bos* species complex of Wu et al. (2018), the actual number of nodes is higher than in our simulations even with $\alpha = 0$. The same result was found with the phylogeny of the *Anopheles gambiae* species complex of Fontaine et al. (2015). By contrast, for the woodcreeper phylogeny of Pulido-Santacruz et al. (2020), we found that the value closest to the biological dataset was obtained with $\alpha = 100$. Lastly, it was estimated that values of $\alpha$ between 100 and 1000 best fit the spider phylogeny of Leduc-Robert and Maddison (2018). These results are described in full in Supplementary Materials section 3. These examples, taken from very diverse organisms, tend to show that the higher probability of introgression between closely related species is not sufficient to secure a safe zone for the ABBA-BABA test.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

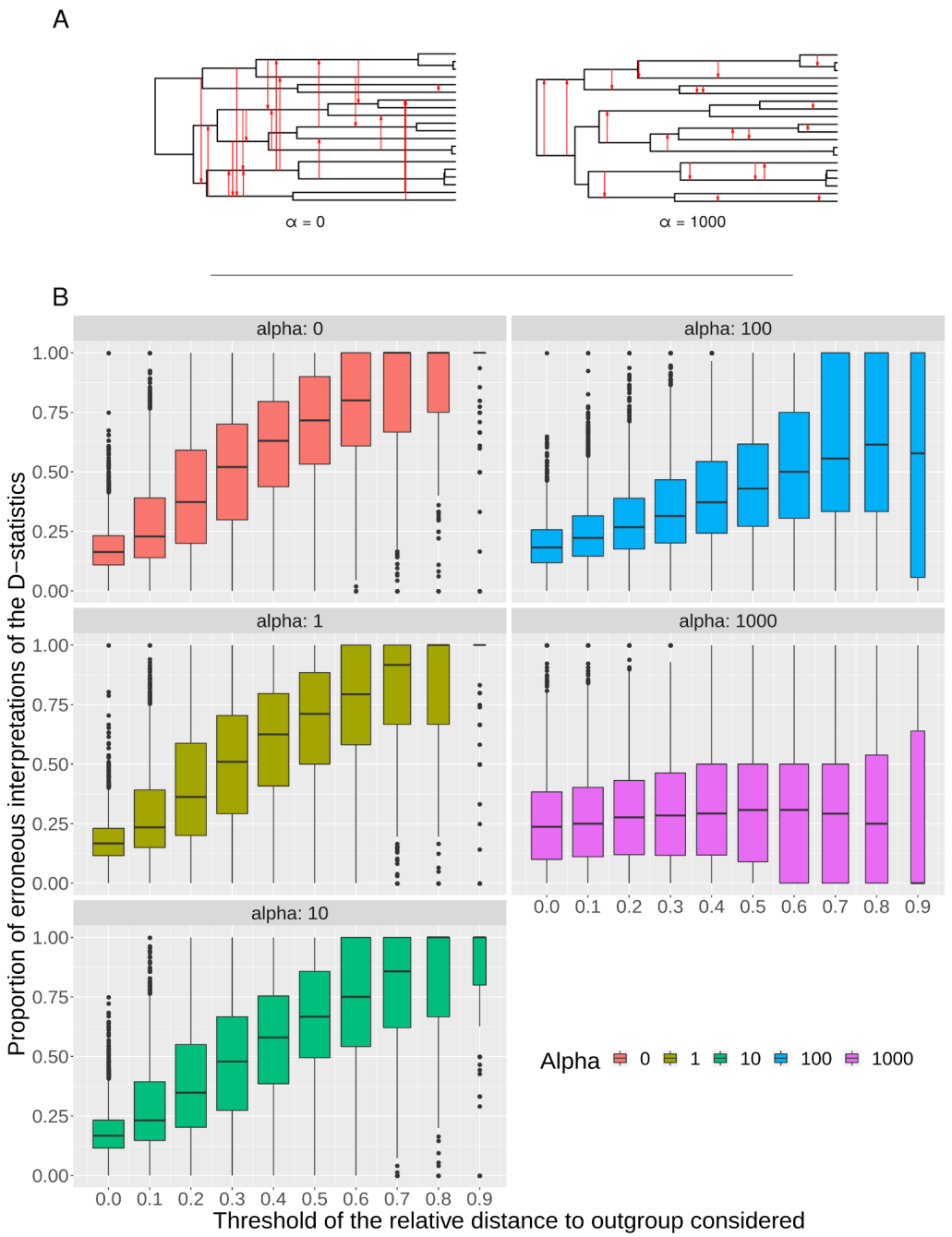


**Figure 6.** The effect of the probability of introgression on the proportion of erroneous interpretations of the D-statistic. **A.** Illustration of the effect of the $\alpha$ parameter, which imposes constraints on introgression in relation to phylogenetic distance. **B**. Relationship between relative outgroup distance ($x$-axis) and the proportion of erroneous interpretations ($y$-axis) for different levels of constraint on introgression related to phylogenetic distance, determined by $\alpha = \{0, 1, 10, 100, 1000\}$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 4 Discussion

### 4.1 Ghost lineages: An important factor affecting introgression tests

Different parameters are known to affect the robustness or sensitivity of the D-statistic. For instance, variations in population size (Eriksson and Manica 2012; Lohse and Frantz 2014; Martin et al. 2015; Zheng and Janke 2018) and/or ancestral population structure (Durand et al. 2011; Lohse and Frantz 2014; Martin et al. 2015) have been shown to produce significant D-statistics in the absence of introgression. Applying the D-statistic to smaller genomic windows, rather than over the entire genome, gives very variable estimates of D (Martin et al. 2015). Complex introgression scenarios, with more than one introgression in the quartet, are another source of error (Rogers and Bohlender 2015; Elworth et al. 2018). Our findings suggest that, in addition to the variables listed above, the interpretation of the D-statistic should systematically, and maybe primarily, take into account ghost lineages.

Recently, Hibbins and Hahn (2021) published results that are in line with our findings. Their simulation study confirms that introgression from a midgroup ghost lineage can result in a significant D-statistic, which may lead to the misidentification of the identity of the lineages involved in the introgression if only known lineages are considered. Similar results have been observed with the D3 statistic (Hahn and Hibbins 2019), a test for detecting introgression that uses only three lineages and the branch lengths of the phylogeny. While the study of Hibbins and Hahn (2021) confirms that ghost introgressions may lead to erroneous interpretations, they do not quantify the extent to which this factor affects the interpretation of the D-statistic.

### 4.2. An intractable incompleteness

Although some families are believed to be extensively described (Chapman and Burke 2007), it is not possible and probably will never be possible to assume that we work with an exhaustive taxon set. A study from 2011 estimated that 8.7 million of eukaryote species are alive today (Mora et al. 2011), and a study from 2016 estimated that there are 1 trillion species on Earth (Locey and Lennon 2016). By contrast, 2.5 million species have been described and catalogued in The Catalogue of Life (CoL). The kingdom Animalia has the most descriptions with 1.4M catalogued species, while Plantae, Fungi and other kingdoms have 375K, 145K and 81K catalogued species, respectively. This means that, at best, we know 25% of the biodiversity that is alive today. In practice, there is a strong disparity in the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

percentage of undescribed species among taxonomic groups. Chapman estimated in 2009 that less than 20% of the phylum Chordata is yet to be described while 80% of the class Insecta is still unknown to us (Chapman et al. 2009). The proportion of undescribed species is even higher in prokaryotes with less than 1% of species described for viruses, Archaea and Bacteria, and because microbial biodiversity is harder to estimate than that of macro-organisms, these numbers could be orders of magnitude lower. Thus, the effects shown here cannot be circumvented by adding or expecting more species and improving computational techniques to handle larger datasets. We are bound to work with a very small fraction of what exists. According to our simulations, with 25% of species sampled, on average more than 50% of introgressions could be due to ghost lineages and subsequently be misinterpreted by a D-statistic. This implies that, with the exception of some well described eukaryote groups such as the genus *Homo*, our lack of taxonomic knowledge will greatly impact the reliability of the D-statistic.

### 4.3 Other introgression detection methods

Several other methods have been developed to mitigate some of the limitations of the D-statistic, but their robustness to ghost lineages has not yet been explored.

It is possible to apply the D-statistic test in datasets with more than four species by performing multiple tests on different quartets. The D-statistics are then analyzed together, using the interpretation of each individual test as a constraint for the interpretation of the other tests. This enables a finer detection of introgression, the identification of donors and recipients (while a single test cannot distinguish the donor from the recipient), and possibly assigning introgression events to groups of taxa instead of single taxa (Pease et al. 2016; Rouard et al. 2018; Suvorov et al. 2020). However, if each individual test is interpreted, as it is usually done, without considering the possibility of ghost introgressions, the joint interpretation of multiple tests will miss a high number of scenarios. Moreover, there is no method that formalizes the constraints from multiple tests and no guarantee that the result is correct or unique, and that the order in which single tests are analysed does not matter.

Extensions of the D-statistic, namely the partitioned-D (Eaton and Ree 2013) and $D_{FOIL}$ (Pease and Hahn 2015) tests, have been proposed to infer introgression in 5-taxon phylogenies (instead of four) and to polarize (in some cases a direction to the introgression can be assessed) the introgression. Although $D_{FOIL}$ can detect more introgression patterns than

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ABBA-BABA, it is still blind to ghost lineages, and thus presents similar theoretical possibilities of misinterpretation as the ABBA-BABA test (see Supplementary Material Section 5 for a listing of the patterns that lead to misinterpretation). This possibility is mentioned by Pease and Hahn (2015) but the proportion of misinterpretations remains to be quantified, and alternative interpretations handling ghost lineages has not been written.

Soraggi et al. (2018) proposed an extension of the D-statistic ($D_{ext}$) to study introgression events among non-African human populations using Africans as the outgroup. This test is robust to introgression from an external group which is not part of the analysed populations. However, the use of this version of D-statistic is restricted to extinct clades for which introgression events with extant species have been already identified as the Neanderthal introgression. This precludes its use in cases where there is no *a priori* knowledge on the existence of ghost lineages.

Other more global methods, such as STRUCTURE/ADMIXTURE (Pritchard et al. 2000; Tang et al. 2005), Treemix (Pickrell and Pritchard 2012) and Phylonet (Than et al. 2008; Wen et al. 2018), have been designed to detect introgression across an entire phylogeny. These tools do not consider ghost lineages and their potential effect on the detected signal. Thus, introgression events are only inferred between known lineages and branches in a tree. It is interesting to note that two of these tools, STRUCTURE and ADMIXTURE, which are popular choices for reconstructing genetic history and testing admixture scenarios, have recently been shown to be subject to misinterpretation due to ghost introgressions (Lawson et al. 2018). The impact of ghost lineages on the detection of introgression is therefore not just a question of using the right tool.

### 4.4. Application of the D-statistic in the light of ghost introgressions.

Now that the importance of the possibility of ghost introgression is recognized, and that no current method is able to handle the effects of ghost lineages, it is time to adapt all methods or develop new methods to take this factor into account.

For the single D-statistic test (four-taxon quartet), the solution is simply to take this uncertainty into account by considering alternative scenarios with at least equal probability. In phylogenies with more than four taxa, the D-statistics from all quartets could be analyzed using an algorithmic method that would combine a set of scenarios. This will require formalizing the objective (*i.e.* minimizing the number of incoherences between quartet

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

results), choosing a set of quartets according to this objective and devising a combinatorial algorithm that handles, for each quartet, information from several possible scenarios including ghost lineages.

Note that this approach will not only avoid interpretation errors, but will possibly point to the existence of unknown lineages that have contributed through introgression to the genomes of known lineages. Therefore, this approach would combine the detection of introgression and the detection of unsampled or extinct taxa. This has already been achieved with *ad hoc* methods for human (Prüfer et al. 2014; Dannemann and Racimo 2018) and whale lineages (Foote et al. 2019) and could be generalized to enrich the phylogeny of known species with unknown species, for which we have no trace other than the genes that have lived, for a while, in them. This is a promising route for future work.

# 5. Conclusion

The D-statistic is a key tool for studying introgression as it provides, in specific cases, a robust test for detecting gene flow. However, our results show that one important caveat of this test is its lack of consideration of ghost lineages, which can lead to the misinterpretation of a significant result. Thus, the *bona fide* interpretation of a single significant D-statistic should be a set of possible scenarios that include the possibility of ghost introgressions, which are equally likely in the absence of other information. Based on our simulations, we have suggested that ghost introgressions are often the most likely scenario. It is possible that in the future, the usual interpretation of a significant D-statistic , *i.e.* ingroup introgression, becomes the exception rather than the rule.

**Acknowledgments**

This work was supported by the French National Research Agency (Grants ANR-18-CE02-0007-01 and ANR-19-CE45-0010). Simulations were performed using the computing facilities of the CC LBBE/PRABI. We thank Gergely Szöllosi for useful discussions

**Software Availability**

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

## References

Barlow A., Cahill J.A., Hartmann S., Theunert C., Xenikoudakis G., Fortes G.G., Paijmans J.L.A., Rabeder G., Frischauf C., Grandal-d'Anglade A., García-Vázquez A., Murtskhvaladze M., Saarma U., Anijalg P., Skrbinšek T., Bertorelle G., Gasparian B., Bar-Oz G., Pinhasi R., Slatkin M., Dalén L., Shapiro B., Hofreiter M. 2018. Partial genomic survival of cave bears in living brown bears. Nat. Ecol. Evol. 2:1563–1570.

Barlow A., Cahill J.A., Hartmann S., Theunert C., Xenikoudakis G., Fortes G.G., Paijmans J.L.A., Rabeder G., Frischauf C., Grandal-D'Anglade A., García-Vázquez A., Murtskhvaladze M., Saarma U., Anijalg P., Skrbinšek T., Bertorelle G., Gasparian B., Bar-Oz G., Pinhasi R., Slatkin M., Dalén L., Shapiro B., Hofreiter M. 2019. Data from: Partial genomic survival of cave bears in living brown bears. :6331386741 bytes.

Bordewich M., Semple C. 2005. On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance. Ann. Comb. 8:409–423.

Cahill J.A., Green R.E., Fulton T.L., Stiller M., Jay F., Ovsyanikov N., Salamzade R., John J.S., Stirling I., Slatkin M., Shapiro B. 2013. Genomic Evidence for Island Population Conversion Resolves Conflicting Theories of Polar Bear Evolution. PLOS Genet. 9:e1003345.

Chapman A.D., 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 M.A., Burke J.M. 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.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for Ancient Admixture between Closely Related Populations. Mol. Biol. Evol. 28:2239–2252.

Eaton D.A.R., Hipp A.L., 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.

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

Edmands S. 2002. Does parental divergence predict reproductive compatibility? Trends Ecol. Evol. 17:520–527.

Elworth R.A.L., Allen C., Benedict T., Dulworth P., Nakhleh L. 2018. DGEN: A Test Statistic for Detection of General Introgression Scenarios. bioRxiv.:348649.

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.

Fontaine M.C., Pease J.B., Steele A., Waterhouse R.M., Neafsey D.E., Sharakhov I.V., Jiang X., Hall A.B., Catteruccia F., Kakani E., Mitchell S.N., Wu Y.-C., Smith H.A., Love R.R., Lawniczak M.K., Slotman M.A., Emrich S.J., Hahn M.W., Besansky N.J. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 347:1258524.

Foote A.D., Martin M.D., Louis M., Pacheco G., Robertson K.M., Sinding M.-H.S., Amaral A.R., Baird R.W., Baker C.S., Ballance L., Barlow J., Brownlow A., Collins T., Constantine R., Dabin W., Rosa L.D., Davison N.J., Durban J.W., Esteban R., Ferguson S.H., Gerrodette T., Guinet C., Hanson M.B., Hoggard W., Matthews C.J.D., Samarra F.I.P., Stephanis R. de, Tavares S.B., Tixier P., Totterdell J.A., Wade P., Excoffier L., Gilbert M.T.P., Wolf J.B.W., Morin P.A. 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 R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C., Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L.V.,

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lalueza-Fox C., Rasilla M. de la, Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.

Hahn M.W., Hibbins M.S. 2019. A Three-Sample Test for Introgression. Mol. Biol. Evol. 36:2878–2882.

Hailer F., Kutschera V.E., Hallström B.M., Klassert D., Fain S.R., Leonard J.A., Arnason U., Janke A. 2012. Nuclear Genomic Sequences Reveal that Polar Bears Are an Old and Distinct Bear Lineage. Science. 336:344–347.

Hedrick P.W. 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. .

Irwin D.E., Milá B., Toews D.P.L., Brelsford A., Kenyon H.L., Porter A.N., Grossen C., Delmore K.E., Alcaide M., Irwin J.H. 2018. A comparison of genomic islands of differentiation across three young avian species pairs. Mol. Ecol. 27:4839–4855.

Keuler R., Garretson A., Saunders T., Erickson R.J., Andre N.S., Grewe F., Smith H., Lumbsch H.T., Huang J.-P., Clair L.L.S., Leavitt S.D. 2020. Genome-scale data reveal the role of hybridization in lichen-forming fungi. Sci. Rep. 10:1497.

Kulathinal R.J., Stevison L.S., Noor M.A.F. 2009. The Genomics of Speciation in Drosophila: Diversity, Divergence, and Introgression Estimated Using Low-Coverage Genome Sequencing. PLOS Genet. 5:e1000550.

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. Sci. Rep. 7:46487.

Lan T., Cheng J., Ratan A., Miller W., Schuster S.C., Farley S., Shideler R.T., Mailund T., Lindqvist C. 2016. Genome-wide evidence for a hybrid origin of modern polar bears. bioRxiv.:047498.

Lawson D.J., van Dorp L., Falush D. 2018. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. Nat. Commun. 9:3258.

Leduc-Robert G., Maddison W.P. 2018. Phylogeny with introgression in Habronattus jumping spiders (Araneae: Salticidae). BMC Evol. Biol. 18:24.

Li R., Fan W., Tian G., Zhu H., He L., Cai J., Huang Q., Cai Q., Li B., Bai Y., Zhang Z., Zhang Y., Wang W., Li J., Wei F., Li H., Jian M., Li J., Zhang Z., Nielsen R., Li D., Gu W., Yang Z., Xuan Z., Ryder O.A., Leung F.C.-C., Zhou Y., Cao J., Sun X., Fu Y., Fang X., Guo X., Wang B., Hou R., Shen F., Mu B., Ni P., Lin R., Qian W., Wang G., Yu C., Nie W., Wang J., Wu Z., Liang H., Min J., Wu Q., Cheng S., Ruan J., Wang M., Shi Z., Wen M., Liu B., Ren X., Zheng H., Dong D., Cook K., Shan G., Zhang H., Kosiol C., Xie X., Lu Z., Zheng H., Li Y., Steiner C.C., Lam T.T.-Y., Lin S., Zhang Q., Li G., Tian J., Gong T., Liu H., Zhang D., Fang L., Ye C., Zhang J., Hu W., Xu A., Ren Y., Zhang G., Bruford M.W., Li Q., Ma L., Guo Y., An N., Hu Y., Zheng Y., Shi Y., Li Z., Liu Q., Chen Y., Zhao J., Qu N., Zhao S., Tian F., Wang X., Wang H., Xu L., Liu X., Vinar T., Wang Y., Lam T.-W., Yiu S.-M., Liu S., Zhang H., Li D., Huang Y., Wang X., Yang G., Jiang Z., Wang J., Qin N., Li L., Li J., Bolund L., Kristiansen K., Wong G.K.-S., Olson M., Zhang X., Li S., Yang H., Wang J., Wang J. 2010. The sequence and de novo assembly of the giant panda genome. Nature. 463:311–317.

Liu L., Bosse M., Megens H.-J., Frantz L.A.F., Lee Y.-L., Irving-Pease E.K., Narayan G., Groenen M.A.M., Madsen O. 2019. Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion. Nat. Commun. 10:1992.

Liu S., Lorenzen E.D., Fumagalli M., Li B., Harris K., Xiong Z., Zhou L., Korneliussen T.S., Somel M., Babbitt C., Wray G., Li J., He W., Wang Z., Fu W., Xiang X., Morgan C.C., Doherty A., O'Connell M.J., McInerney J.O., Born E.W., Dalén L., Dietz R., Orlando L., Sonne C., Zhang G., Nielsen R., Willerslev E., Wang J. 2014. Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears. Cell. 157:785–794.

Locey K.J., Lennon J.T. 2016. Scaling laws predict global microbial diversity. Proc. Natl. Acad. Sci. 113:5970–5975.

Lohse K., Frantz L.A.F. 2014. Neandertal Admixture in Eurasia Confirmed by Maximum-Likelihood Analysis of Three Genomes. Genetics. 196:1241–1251.

Maddison W.P. 1997. Gene Trees in Species Trees. Syst. Biol. 46:523–536.

Maddison W.P., Knowles L.L. 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 S.H., Dasmahapatra K.K., Nadeau N.J., Salazar C., Walters J.R., Simpson F., Blaxter M., Manica A., Mallet J., Jiggins C.D. 2013. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 23:1817–1828.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Martin S.H., Davey J.W., Jiggins C.D. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. Mol. Biol. Evol. 32:244–257.

Massardo D., VanKuren N.W., Nallu S., Ramos R.R., Ribeiro P.G., Silva-Brandão K.L., Brandão M.M., Lion M.B., Freitas A.V.L., Cardoso M.Z., Kronforst M.R. 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 J.I., Marques D.A., Mwaiko S., Wagner C.E., 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 J.G., Jay F., Prüfer K., Filippo C. de, Sudmant P.H., Alkan C., Fu Q., Do R., Rohland N., Tandon A., Siebauer M., Green R.E., Bryc K., Briggs A.W., Stenzel U., Dabney J., Shendure J., Kitzman J., Hammer M.F., Shunkov M.V., Derevianko A.P., Patterson N., Andrés A.M., Eichler E.E., Slatkin M., Reich D., Kelso J., Pääbo S. 2012. A High-Coverage Genome Sequence from an Archaic Denisovan Individual. Science. 338:222–226.

Montanari S.R., Hobbs J.-P.A., Pratchett M.S., Bay L.K., 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 D.P., Adl S., Simpson A.G.B., Worm B. 2011. How Many Species Are There on Earth and in the Ocean? PLoS Biol. 9:e1001127.

Osborne O.G., Chapman M.A., Nevado B., Filatov D.A. 2016. Maintenance of Species Boundaries Despite Ongoing Gene Flow in Ragworts. Genome Biol. Evol. 8:1038–1047.

Ottenburghs J. 2020. Ghost Introgression: Spooky Gene Flow in the Distant Past. BioEssays. 42:2000012.

Ottenburghs J., Kraus R.H.S., van Hooft P., van Wieren S.E., Ydenberg R.C., Prins H.H.T. 2017. Avian introgression in the genomic era. Avian Res. 8:30.

Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 35:526–528.

Patterson N., Moorjani P., Luo Y., Mallick S., Rohland N., Zhan Y., Genschoreck T., Webster T., Reich D. 2012. Ancient Admixture in Human History. Genetics. 192:1065–1093.

Pease J.B., Haak D.C., Hahn M.W., Moyle L.C. 2016. Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. PLOS Biol. 14:e1002379.

Pease J.B., Hahn M.W. 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. Syst. Biol. 64:651–662.

Pickrell J.K., Pritchard J.K. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genet. 8:e1002967.

Pritchard J.K., 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 P.H., de Filippo C., Li H., Mallick S., Dannemann M., Fu Q., Kircher M., Kuhlwilm M., Lachmann M., Meyer M., Ongyerth M., Siebauer M., Theunert C., Tandon A., Moorjani P., Pickrell J., Mullikin J.C., Vohr S.H., Green R.E., Hellmann I., Johnson P.L.F., Blanche H., Cann H., Kitzman J.O., Shendure J., Eichler E.E., Lein E.S., Bakken T.E., Golovanova L.V., Doronichev V.B., Shunkov M.V., Derevianko A.P., Viola B., Slatkin M., Reich D., Kelso J., Pääbo S. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature. 505:43–49.

Pulido-Santacruz P., Aleixo A., Weir J.T. 2020. Genomic data reveal a protracted window of introgression during the diversification of a neotropical woodcreeper radiation*. Evolution. 74:842–858.

Raup D.M. 1991. Extinction: bad genes or bad luck? New York: W.W. Norton.

Rogers A.R., Bohlender R.J. 2015. Bias in estimators of archaic admixture. Theor. Popul. Biol. 100:63–78.

Rouard M., Droc G., Martin G., Sardos J., Hueber Y., Guignon V., Cenci A., Geigle B., Hibbins M.S., Yahiaoui N., Baurens F.-C., Berry V., Hahn M.W., D'Hont A., Roux N. 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 D.L., Rosenthal G.G., Andolfatto P. 2016. Ancient hybridization and genomic stabilization in a swordtail fish. Mol. Ecol. 25:2661–2679.

Smith J., Kronforst M.R. 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.

Suvorov A., Kim B.Y., Wang J., Armstrong E.E., Peede D., D'Agostino E.R.R., Price D.K., Wadell P., Lang M., Courtier-Orgogozo V., David J.R., Petrov D., Matute D.R., Schrider D.R., Comeault A.A. 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.30.437672; this version posted October 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Widespread introgression across a phylogeny of 155 Drosophila genomes. bioRxiv.:2020.12.14.422758.

Szöllősi G.J., Davín A.A., 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 G.J., 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 N.J. 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.

Wu D.-D., Ding X.-D., Wang S., Wójcik J.M., Zhang Y., Tokarska M., Li Y., Wang M.-S., Faruque O., Nielsen R., Zhang Q., Zhang Y.-P. 2018. Pervasive introgression facilitated domestication and adaptation in the Bos species complex. Nat. Ecol. Evol. 2:1139–1145.

Zhang B.-W., Xu L.-L., Li N., Yan P.-C., Jiang X.-H., Woeste K.E., Lin K., Renner S.S., 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.

Zheng Y., Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. BMC Bioinformatics. 19:10.