Five-leaf generalizations of the D-statistic reveal the directionality of admixture

Kalle Leppälä[12]✉, Flavio Augusto da Silva Coelho[3*], Michaela Richter[3*], Victor A. Albert[3] and Charlotte Lindqvist[3]✉

[1] Research Unit of Mathematical Sciences, University of Oulu, 90014 Oulu, Finland

[2] The Organismal and Evolutionary Biology Research Programme University of Helsinki, 00014 Helsinki, Finland

[3] Department of Biological Sciences, University at Buffalo, 14260 Buffalo NY, USA

February 24, 2024

Abstract

Over the past 15 years, the D-statistic, a four-taxon test for organismal admixture (hybridization, or introgression) that incorporates single nucleotide polymorphism data with allelic patterns ABBA and BABA, has seen considerable use. This statistic seeks to discern significant deviation from either a given species tree assumption, or from the balanced incomplete lineage sorting that could otherwise defy this species tree. However, while the D-statistic can successfully discriminate admixture from incomplete lineage sorting, it is not a simple matter to determine the directionality of admixture using only four-leaf tree models. As such, methods have been developed that use 5 leaves to evaluate admixture trajectories. Among these, the D_{FOIL} method, which tests allelic patterns on the "symmetric" tree $S = (((1,2),(3,4)),5)$, succeeds in finding admixture direction for many five-taxon examples. However, D_{FOIL} does not make full use of all symmetry, nor can D_{FOIL} properly accommodate the use of singleton patterns (such as BAAAA and ABAAA) when ancient samples are included. The latter becomes problematic since "mutational time" elapsed since ancient organism death is unequal to that of modern samples, and because damaged DNA can bias results in unknown ways. Here, we take inspiration from D_{FOIL} to develop a new and completely general family of five-leaf admixture tests, dubbed Δ-statistics, that can either incorporate or exclude the singleton allelic patterns depending on individual taxon and age sampling choices. We describe two new shapes that are also fully testable, namely the "asymmetric" tree $A = ((((1,2),3),4),5)$ and the "quasisymmetric" tree $Q = (((1,2),3),(4,5))$, that can considerably supplement the "symmetric" $S = (((1,2),(3,4)),5)$ model used by D_{FOIL}. We demonstrate the consistency of Δ-statistics under various simulated scenarios, and provide empirical examples using data from black, brown and polar bears, the latter also including two ancient samples from previous studies. Recently D_{FOIL} and one of these ancient samples was used to argue for a dominant polar bear $\rightarrow$ brown bear introgression direction. However, we find, using both this ancient polar bear and our own, that by far the strongest signal using both D_{FOIL} and Δ-statistics on tree S is actually bidirectional gene flow of indistinguishable direction. Further experiments on trees A and Q instead highlight what were likely two phases of admixture: one with stronger brown bear $\rightarrow$ polar bear introgression in ancient times, and a recent phase with predominant polar bear $\rightarrow$ brown bear directionality.

Code and documentation available at `https://github.com/KalleLeppala/Delta-statistics`.

✉ To whom correspondence may be addressed: `kalle.m.leppala@gmail.com`, `cl243@buffalo.edu`

* Equal contributors.

1 Introduction

While phylogenetic tree reconstructions model lineage splits, they do not directly accommodate admixture among populations or species. Demonstrations of introgression among hominin clades (Green et al., 2010; Meyer et al., 2012) were among the first population genomic studies to point out that the evolutionary histories of many organismal groups involve not only lineage bifurcations, but also (sometimes adaptively important) reticulations. Importantly, cross-lineage admixture accommodates the inheritance of traits that might otherwise appear convergently evolved to stem from non-cladistic gene flow instead. Explorations of inter-lineage admixtures at the nucleotide sequence level permit revelation of both the structural and functional impacts of such events. For example, our own research on brown and polar bears has underscored how introgression may have had adaptive consequences following the earlier split of these species into discrete lineages (Lan et al., 2022; Miller et al., 2012). As underscored in that work, alleles introgressed from lower latitude, ecologically generalist brown bear lineages may have influenced exquisitely arctic specific adaptations of polar bears, or vice versa. In this example, it can be appreciated that uncovering the *direction* of admixture events can be of vital importance for understanding both past and future functional impacts of introgression.

Methods development to reveal introgressions based on single nucleotide polymorphism data range from parsimony based to likelihood based, the former taking advantage of simple allelic state patterns on phylogenetic graphs. The classic D-statistic (Huson, Klöpper, Lockhart, & Steel, 2005; Kulathinal, Stevison, & Noor, 2009; Green et al., 2010; Patterson et al., 2012) leverages symmetry in the phylogenetic tree $(((1,2),3),4)$ to detect evidence of introgression between populations. Allelic patterns ABBA (meaning chromosomes from populations 1 and 4 carry allele A while chromosomes from populations 2 and 3 carry allele B at a locus) and BABA are discordant with the underlying phylogeny, but are still sometimes observed due to incomplete lineage sorting (ILS), gene flow, and convergent nucleotide substitutions. Because ILS does not prefer one pattern over the other, its contribution to

$$D = \frac{n(\text{ABBA}) - n(\text{BABA})}{n(\text{ABBA}) + n(\text{BABA})},$$

where $n(\cdot)$ refers to the number of observations, is expected to be zero. At a scale where recurrent substitutions can be seen as rare insignificant events, the deviation of D from zero therefore points to introgression: between populations 2 and 3 or 1 and 4 (or a ghost population stemming from 123 or 1234) when D is positive, and between 1 and 3 or 2 and 4 (or a ghost population stemming from 123 or 1234) when D is negative. A key feature of the D-statistic is that it requires no assumptions on lengths of the branches of the phylogenetic tree, or mutation and coalescence rates among them, and is therefore suitable for the analysis of ancient samples. This is the reason why the D-statistic does not make use of the in principle informative symmetry between the singleton patterns BAAA and ABAA.

Unfortunately, the D-statistic is unable to detect the direction of gene flow. While it is sometimes possible to infer gene flow direction in a sample of only four populations, using f_2-statistic fitting for instance (Leppälä, Nielsen, & Mailund, 2017), the task becomes much less difficult when a fifth population is included in the analysis. This has been a major motivation behind attempts to generalize the classic D-statistic to five-leaf trees, such as the Partitioned D (Eaton & Ree, 2013) and notably D_{FOIL} (Pease & Hahn, 2015) methodologies.

The D_{FOIL} method assumes the phylogenetic tree $S = (((1,2),(3,4)),5)$, which the authors call *symmetric*, (pictured in Figure 1 Panel I). Counts of allelic patterns are compiled into a *signature* $(D_{\text{FO}}\, D_{\text{IL}}\, D_{\text{FI}}\, D_{\text{OL}})$ of statistics, all expected to be zero without admixture. When they deviate from zero, the signs of the signature can be compared with predicted behaviour of various gene flow events, for example, gene flow $1 \to 3$ should produce the signature $(+++0)$. The authors state that the tree $A = ((((1,2),3),4),5)$, which they label *asymmetric* (pictured in Figure 1 Panel II), does not exhibit enough symmetry for construction of a system of statistics analogous to D_{FOIL}. However, in our previous work (Lan et al., 2022, Figure 3) we have already leveraged symmetry within the "asymmetric" tree A; namely the equal probability between gene trees $((1,3),(2,4)),5)$ and $((2,3),(1,4)),5)$ under the null hypothesis of no gene flow. We will demonstrate with proof how allelic pattern counts can be used to infer gene flow direction also in trees of form A.

The current work presents systems of allelic pattern counts that we call Δ-statistics. We consider S, A, and a third tree $Q = (((1,2),3),(4,5))$, which we dub *quasisymmetric* (pictured in Figure 1, Panel III)),

and make full use of the symmetry within each. Together, S, A and Q comprise a complete treatise of trees subject to five-leaf relabeling. Some of these Δ-statistics are based on singleton patterns and are therefore not generally applicable to ancient data (for reasons discussed below), but we nonetheless describe all Δ-statistic subsets as a generalization in the spirit of the classic D-statistic for five leaf trees.

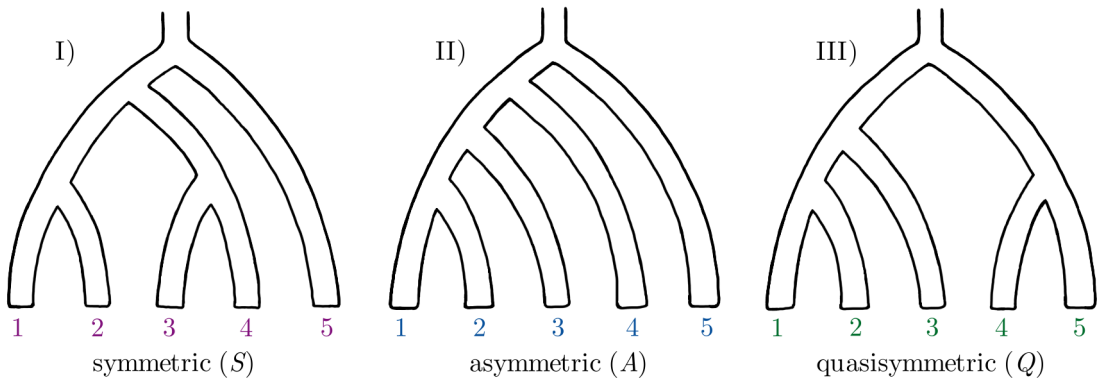

Figure 1: The three possible phylogenetic trees with five leaves.

Our Δ-statistic method is inspired by the D_{FOIL} method, over which it has some advantages:

- Inclusion of trees A and Q.
- Inclusion of several scenarios within tree S that were not considered by Pease and Hahn (2015).
- When not using singleton patterns (applicable to ancient data; as with classic D), our method provides increased power, while retaining the ability to detect the same set of admixture scenarios within tree S as D_{FOIL} does, and additionally gene flow into population 5.
- When resorting to singleton patterns (generally not applicable to ancient data, as with D_{FOIL}), it is able to distinguish more admixture scenarios within tree S than what D_{FOIL} permits.

We test the Δ-statistics with simulations, and apply the method to further analyse the brown bear – polar bear introgression problem.

2 Methods

Definition of Δ-statistics

The *unscaled* Δ-statistics have the form $n(L) - n(R)$, where L and R are two disjoint sets of allelic patterns expected to be observed with equal probability under the null hypothesis of no gene flow, ignoring the possibility of reverse and convergent substitutions. The *scaled* Δ-statistic, decorated with a star, is then $(n(L) - n(R))/(n(L) + n(R))$. In this general framework, D- and D_{FOIL}-statistics are special cases of scaled Δ-statistics.

Because tree Q does not have an outgroup, and because we wish to allow admixture involving the root, or population 5, in trees S and A, we make the *outgroup mutation assumption.* That is, we adopt a notation where allele A is the one observed more often (at least three times out of five) than the other allele B, while remaining agnostic about the ancestral state.

Outgoup mutation assumption: In trees S and A, population 5 is just like any other; both the ancestral and the derived allele state are allowed. (1)

In contrast, the D- and the D_{FOIL}-statistics assume that the ancestral states are known and that the outgroup always carries the ancestral allele, be it A or B. That is an unnecessary complication, and in fact

the allele frequency -based formulae of the D- and D_{FOIL}-statistics, as presented in Harris and DeGiorgio (2017), implicitly make the outgroup mutation assumption (1) as well. In our case, a data preprocessing step of filtering out the patterns where 5 carries the derived allele would not only be unnecessary but actively harmful, although it would only affect some of the Δ-statistics and only when there is gene flow involving the root or the (would-be) outgroup 5.

Consider tree S without gene flow events, and the coalescent process connecting lineages 1–5 at some locus. If we observe the allelic pattern AAABB, caused by a single mutation, then one of two events must have occurred:

- Lineages 4 and 5 first coalesced with each other (first as in closest to the present; in the context of coalescence processes we consider time running backwards). A subsequent mutation occurred before coalescence with any other lineages, only affecting 4 and 5. Allele A is now ancestral. An example of this scenario is depicted in Figure 2 Panel I), where the black box is replaced with the upper of the two bracketed options.

- Lineages 1, 2 and 3 all coalesced with each other in some order before coalescence with lineages 4 or 5. A subsequent mutation occurred, only affecting 1 and 2 and 3. Allele B is now ancestral. An example of this scenario is depicted in Figure 2 Panel II), where the black box is with the upper of the two bracketed options.

In both scenarios, as soon as lineages 3 and 4 are in the same population, their role is symmetrical concerning coalescent history — what happens to 3 could just as well have happened to 4. This is visualized in Figure 2 Panels I) and II), with the black boxes representing two equally likely options, upper and lower. In both scenarios above, the lower option would have resulted in pattern AABAB, which therefore has exactly the same probability as AAABB. The two allelic patterns form an *equal probability set* {AABAB, AAABB} for tree S; we do not know the probability of observing either of these patterns, but we know they are the same by symmetry. Note that if we filter the data so that lineage 5 always carries the ancestral allele, giving up the outgroup mutation assumption (1), we simply censor the latter of the two scenarios listed above, and the set is still an equal probability set.

Symmetry between singleton patterns depends on mutations occurring with the same probability in different populations. The patterns BAAAA and ABAAA have the same probability if the mutation occurred after populations 1 and 2 merged. However, if the mutation occurred in one of the terminal branches, there are many potential reasons for asymmetry, such as different generation times, mutation rates, and in the case of ancient samples, DNA degradation or simply fewer generations. In Figure 2 Panel III) we have depicted such a situation. In the following, equal probability sets of singleton patterns, such as the set {BAAAA, ABAAA}, are burdened with the *synchronization assumption*.

Synchronization assumption: The probability of a mutation in a lineage depends only on the time interval, not the population the lineage is in. (2)

Notice that this assumption concerns the entire species tree, not only the terminal branches. We also point out that, theoretically even ancient samples can satisfy the synchronization assumption when mutations on the terminal branches are so rare that they do not influence the analysis (see simulation in Appendix 5). In fact, methods using the perspective of drift instead of the perspective of coalescence, such as f_2-, f_3- and f_4-statistics (Patterson et al., 2012), typically rely on a simplification that variants are already polymorphic at the root, which is simply a special case of the synchronization assumption. Also in this regard, the D_{FOIL} requirement that leaf 5 always carries the ancestral allele (lack of the outgroup mutation (1)) is problematic, as it emphasizes recent mutations. If an investigator cannot commit to the synchronization assumption, they can simply resort to the subset of Δ-statistics constructed from equal probability sets that do not depend on it.

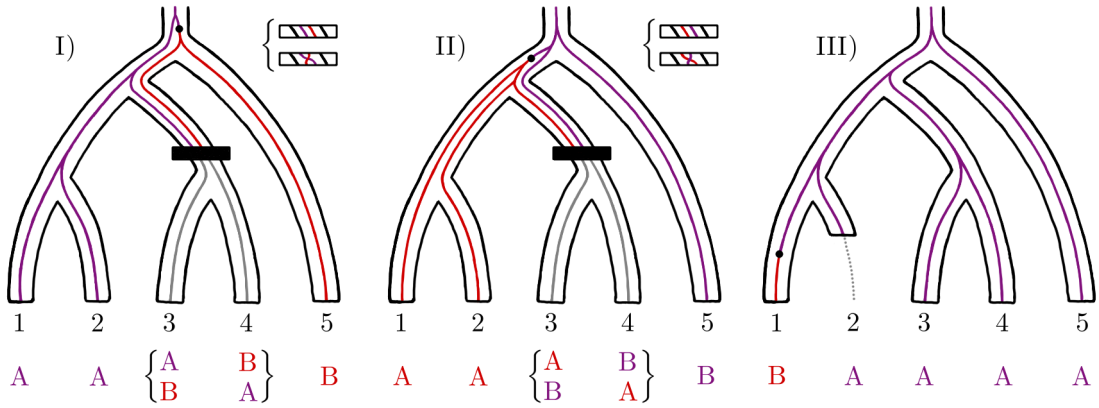

Figure 2: The pattern AAABB results either from a mutation affecting leaves 4 and 5 as in Panel I), or a mutation affecting leaves 1, 2 and 3 as in Panel II). The letter B always denotes the rarer allele, not necessarily the derived allele (red). Symmetry between lineages 3 and 4 is represented as the two black boxes, each hiding two equally likely events that visualize how the probability of pattern AABAB must be the same as the probability of AAABB. We call {AABAB, AAABB} an equal probability set.
In Panel III) {BAAAA, ABAAA} is not an equal probability set because the synchronization assumption (2) is violated. Here, 2 is an ancient sample that was long dead when a mutation occurred in lineage 1.

Table 1: Equal probability sets and binomial Δ-statistics constructed from them. Rows highlighted grey require the synchronization assumption (2).

tree	equal probability set	binomial Δ-statistics
S	{BABAA, BAABA, ABBAA, ABABA}	${}_S\Delta_1 = n(\text{BABAA}) - n(\text{BAABA})$ ${}_S\Delta_2 = n(\text{ABBAA}) - n(\text{ABABA})$ ${}_S\Delta_3 = n(\text{BABAA}) - n(\text{ABBAA})$ ${}_S\Delta_4 = n(\text{BAABA}) - n(\text{ABABA})$
	{BAAAB, ABAAB}	${}_S\Delta_5 = n(\text{BAAAB}) - n(\text{ABAAB})$
	{AABAB, AAABB}	${}_S\Delta_6 = n(\text{AABAB}) - n(\text{AAABB})$
	{BAAAA, ABAAA}	${}_S\Delta_7 = n(\text{BAAAA}) - n(\text{ABAAA})$
	{AABAA, AAABA}	${}_S\Delta_8 = n(\text{AABAA}) - n(\text{AAABA})$
A	{BABAA, ABBAA}	${}_A\Delta_1 = n(\text{BABAA}) - n(\text{ABBAA})$
	{BAABA, ABABA}	${}_A\Delta_2 = n(\text{BAABA}) - n(\text{ABABA})$
	{BAAAB, ABAAB}	${}_A\Delta_3 = n(\text{BAAAB}) - n(\text{ABAAB})$
	{BAAAA, ABAAA}	${}_A\Delta_4 = n(\text{BAAAA}) - n(\text{ABAAA})$
Q	{BAABA, BAAAB, ABABA, ABAAB}	${}_Q\Delta_1 = n(\text{BAABA}) - n(\text{BAAAB})$ ${}_Q\Delta_2 = n(\text{ABABA}) - n(\text{ABAAB})$ ${}_Q\Delta_3 = n(\text{BAABA}) - n(\text{ABABA})$ ${}_Q\Delta_4 = n(\text{BAAAB}) - n(\text{ABAAB})$
	{BABAA, ABBAA}	${}_Q\Delta_5 = n(\text{BABAA}) - n(\text{ABBAA})$
	{AABBA, AABAB}	${}_Q\Delta_6 = n(\text{AABBA}) - n(\text{AABAB})$
	{BAAAA, ABAAA}	${}_Q\Delta_7 = n(\text{BAAAA}) - n(\text{ABAAA})$
	{AAABA, AAAAB}	${}_Q\Delta_8 = n(\text{AAABA}) - n(\text{AAAAB})$

Given an equal probability set, we pick one element for L and one for R to construct a *binomial* Δ-statistic $n(L) - n(R)$. The binomial Δ-statistics are random variables that have mean zero under the null hypothesis of no admixture. Which element goes to L and which goes to R is arbitrary; our choices and the full collection of equal probability sets for trees S, A and Q can be found in Table 1.

Even though 4 "choose" 2 is $4!/(2! \cdot 2!) = 6$, we construct only four binomial Δ-statistics from the four-element equal probability sets in Table 1. The reason is linear dependency within the full set of six possible differences; the missing two Δ-statistics can be expressed with the four binomial Δ-statistics already defined. For example, in tree S:

$$\begin{cases} n(\text{BABAA}) - n(\text{ABABA}) = {}_S\Delta_1 + {}_S\Delta_4 = {}_S\Delta_2 + {}_S\Delta_3 = {}_S\Delta_{1+4} = {}_S\Delta_{2+3} \\ n(\text{BAABA}) - n(\text{ABBAA}) = {}_S\Delta_3 - {}_S\Delta_1 = {}_S\Delta_4 - {}_S\Delta_2 = {}_S\Delta_{3-1} = {}_S\Delta_{4-2} \end{cases}.$$

This technique and notation can be expanded to construct a large number of Δ-statistics as sums and differences of binomial Δ-statistics, all of which are expected to be zero by the linearity of the mean. For the purpose of scaling and statistical inference, we must understand the sets L and R as disjoint. That is, for ${}_S\Delta_{1+4}$ we have $L = \{\text{BABAA}\} \neq \{\text{BABAA}, \text{BAABA}\}$ and $R = \{\text{ABABA}\} \neq \{\text{BAABA}, \text{ABABA}\}$, because ABABA cannot exist in both sets.

This large family of Δ-statistics, combined from the binomial Δ-statistics, includes as special cases the D_{FOIL}-statistics (Pease & Hahn, 2015) and the Partitioned D (Eaton & Ree, 2013), which can be expressed as:

$$\begin{cases} D_{\text{FO}} = {}_S\Delta^*_{1-2-6-8} = {}_S\Delta^*_{3-4-6-8} \\ D_{\text{IL}} = {}_S\Delta^*_{2-1-6-8} = {}_S\Delta^*_{4-3-6-8} \\ D_{\text{FI}} = {}_S\Delta^*_{1-2-5-7} = {}_S\Delta^*_{3-4-5-7} \\ D_{\text{OL}} = {}_S\Delta^*_{2-1-5-7} = {}_S\Delta^*_{4-3-5-7} \end{cases} \qquad \begin{cases} D_{12} = {}_S\Delta^*_5 \\ D_1 = {}_S\Delta^*_{-3} \\ D_2 = {}_S\Delta^*_{-4} \end{cases}$$

The star stands for the scaling $(n(L) - n(R))/(n(L) + n(R))$. There also exist some Δ-statistics such as $n(\text{AABAA}) + n(\text{BABAA}) + n(\text{AABBA}) + n(\text{ABAAB}) - n(\text{ABAAA}) - n(\text{BBAAA}) - n(\text{ABABA}) - n(\text{AABAB})$ in tree S that cannot be constructed from the binomial statistics.

We are now prepared to present our preferred twenty Δ-statistics.

For the symmetric tree, S, we define

$$\begin{aligned}
{}_S\Delta^*_{1-6} &= \frac{\big(n(\text{BABAA})+n(\text{AAABB})\big)-\big(n(\text{BAABA})+n(\text{AABAB})\big)}{\big(n(\text{BABAA})+n(\text{AAABB})\big)+\big(n(\text{BAABA})+n(\text{AABAB})\big)}, \\
{}_S\Delta^*_{2-6} &= \frac{\big(n(\text{ABBAA})+n(\text{AAABB})\big)-\big(n(\text{ABABA})+n(\text{AABAB})\big)}{\big(n(\text{ABBAA})+n(\text{AAABB})\big)+\big(n(\text{ABABA})+n(\text{AABAB})\big)}, \\
{}_S\Delta^*_{3-5} &= \frac{\big(n(\text{BABAA})+n(\text{ABAAB})\big)-\big(n(\text{ABBAA})+n(\text{BAAAB})\big)}{\big(n(\text{BABAA})+n(\text{ABAAB})\big)+\big(n(\text{ABBAA})+n(\text{BAAAB})\big)}, \\
{}_S\Delta^*_{4-5} &= \frac{\big(n(\text{BAABA})+n(\text{ABAAB})\big)-\big(n(\text{ABABA})+n(\text{BAAAB})\big)}{\big(n(\text{BAABA})+n(\text{ABAAB})\big)+\big(n(\text{ABABA})+n(\text{BAAAB})\big)}, \\
\hline
{}_S\Delta^*_{5+7} &= \frac{\big(n(\text{BAAAB})+n(\text{BAAAA})\big)-\big(n(\text{ABAAB})+n(\text{ABAAA})\big)}{\big(n(\text{BAAAB})+n(\text{BAAAA})\big)+\big(n(\text{ABAAB})+n(\text{ABAAA})\big)}, \\
{}_S\Delta^*_{6+8} &= \frac{\big(n(\text{AABAB})+n(\text{AABAA})\big)-\big(n(\text{AAABB})+n(\text{AAABA})\big)}{\big(n(\text{AABAB})+n(\text{AABAA})\big)+\big(n(\text{AAABB})+n(\text{AAABA})\big)}, \\
{}_S\Delta^*_{3+4-5+7} &= \frac{\big(n(\text{BABAA})+n(\text{BAABA})+n(\text{ABAAB})+n(\text{BAAAA})\big)-\big(n(\text{ABBAA})+n(\text{ABABA})+n(\text{BAAAB})+n(\text{ABAAA})\big)}{\big(n(\text{BABAA})+n(\text{BAABA})+n(\text{ABAAB})+n(\text{BAAAA})\big)+\big(n(\text{ABBAA})+n(\text{ABABA})+n(\text{BAAAB})+n(\text{ABAAA})\big)}, \\
{}_S\Delta^*_{1+2-6+8} &= \frac{\big(n(\text{BABAA})+n(\text{ABBAA})+n(\text{AAABB})+n(\text{AABAA})\big)-\big(n(\text{BAABA})+n(\text{ABABA})+n(\text{AABAB})+n(\text{AAABA})\big)}{\big(n(\text{BABAA})+n(\text{ABBAA})+n(\text{AAABB})+n(\text{AABAA})\big)+\big(n(\text{BAABA})+n(\text{ABABA})+n(\text{AABAB})+n(\text{AAABA})\big)},
\end{aligned} \tag{3}$$

for the asymmetric tree, A, we define

$$
\begin{aligned}
{}_A\Delta^*_{1-2} &= \frac{\big(n(\text{BABAA})+n(\text{ABABA})\big)-\big(n(\text{ABBAA})+n(\text{BAABA})\big)}{\big(n(\text{BABAA})+n(\text{ABABA})\big)+\big(n(\text{ABBAA})+n(\text{BAABA})\big)}\,, \\
{}_A\Delta^*_{1-3} &= \frac{\big(n(\text{BABAA})+n(\text{ABAAB})\big)-\big(n(\text{ABBAA})+n(\text{BAAAB})\big)}{\big(n(\text{BABAA})+n(\text{ABAAB})\big)+\big(n(\text{ABBAA})+n(\text{BAAAB})\big)}\,, \\
{}_A\Delta^*_{2-3} &= \frac{\big(n(\text{BAABA})+n(\text{ABAAB})\big)-\big(n(\text{ABABA})+n(\text{BAAAB})\big)}{\big(n(\text{BAABA})+n(\text{ABAAB})\big)+\big(n(\text{ABABA})+n(\text{BAAAB})\big)}\,, \\
\hline
{}_A\Delta^*_{1+2-3+4} &= \frac{\big(n(\text{BABAA})+n(\text{BAABA})+n(\text{ABAAB})+n(\text{BAAAA})\big)-\big(n(\text{ABBAA})+n(\text{ABABA})+n(\text{BAAAB})+n(\text{ABAAA})\big)}{\big(n(\text{BABAA})+n(\text{BAABA})+n(\text{ABAAB})+n(\text{BAAAA})\big)+\big(n(\text{ABBAA})+n(\text{ABABA})+n(\text{BAAAB})+n(\text{ABAAA})\big)}\,,
\end{aligned}
\tag{4}
$$

and for the quasisymmetric tree, Q, we define

$$
\begin{aligned}
{}_Q\Delta^*_{1-6} &= \frac{\big(n(\text{BAABA})+n(\text{AABAB})\big)-\big(n(\text{BAAAB})+n(\text{AABBA})\big)}{\big(n(\text{BAABA})+n(\text{AABAB})\big)+\big(n(\text{BAAAB})+n(\text{AABBA})\big)}\,, \\
{}_Q\Delta^*_{2-6} &= \frac{\big(n(\text{ABABA})+n(\text{AABAB})\big)-\big(n(\text{ABAAB})+n(\text{AABBA})\big)}{\big(n(\text{ABABA})+n(\text{AABAB})\big)+\big(n(\text{ABAAB})+n(\text{AABBA})\big)}\,, \\
{}_Q\Delta^*_{3-5} &= \frac{\big(n(\text{BAABA})+n(\text{ABBAA})\big)-\big(n(\text{ABABA})+n(\text{BABAA})\big)}{\big(n(\text{BAABA})+n(\text{ABBAA})\big)+\big(n(\text{ABABA})+n(\text{BABAA})\big)}\,, \\
{}_Q\Delta^*_{4-5} &= \frac{\big(n(\text{BAAAB})+n(\text{ABBAA})\big)-\big(n(\text{ABAAB})+n(\text{BABAA})\big)}{\big(n(\text{BAAAB})+n(\text{ABBAA})\big)+\big(n(\text{ABAAB})+n(\text{BABAA})\big)}\,, \\
\hline
{}_Q\Delta^*_{5+7} &= \frac{\big(n(\text{BABAA})+n(\text{BAAAA})\big)-\big(n(\text{ABBAA})+n(\text{ABAAA})\big)}{\big(n(\text{BABAA})+n(\text{BAAAA})\big)+\big(n(\text{ABBAA})+n(\text{ABAAA})\big)}\,, \\
{}_Q\Delta^*_{6+8} &= \frac{\big(n(\text{AABBA})+n(\text{AAABA})\big)-\big(n(\text{AABAB})+n(\text{AAAAB})\big)}{\big(n(\text{AABBA})+n(\text{AAABA})\big)+\big(n(\text{AABAB})+n(\text{AAAAB})\big)}\,, \\
{}_Q\Delta^*_{3+4-5+7} &= \frac{\big(n(\text{BAABA})+n(\text{BAAAB})+n(\text{ABBAA})+n(\text{BAAAA})\big)-\big(n(\text{ABABA})+n(\text{ABAAB})+n(\text{BABAA})+n(\text{ABAAA})\big)}{\big(n(\text{BAABA})+n(\text{BAAAB})+n(\text{ABBAA})+n(\text{BAAAA})\big)+\big(n(\text{ABABA})+n(\text{ABAAB})+n(\text{BABAA})+n(\text{ABAAA})\big)}\,, \\
{}_Q\Delta^*_{1+2-6+8} &= \frac{\big(n(\text{BABAA})+n(\text{ABBAA})+n(\text{AABAB})+n(\text{AAABA})\big)-\big(n(\text{BAAAB})+n(\text{ABAAB})+n(\text{AABBA})+n(\text{AAAAB})\big)}{\big(n(\text{BAABA})+n(\text{ABABA})+n(\text{AABAB})+n(\text{AAABA})\big)+\big(n(\text{BAAAB})+n(\text{ABAAB})+n(\text{AABBA})+n(\text{AAAAB})\big)}\,.
\end{aligned}
\tag{5}
$$

The horizontal lines remind us that the binomial statistics ${}_S\Delta_7, {}_S\Delta_8, {}_A\Delta_4, {}_Q\Delta_7$ and ${}_Q\Delta_8$ require the synchronization assumption (2). See Appendix 1 for allele frequency -based formulations of (3), (4) and (5).

Classification of expected Δ-statistics

The types of alternative hypotheses considered in this work are *unidirectional gene flow events*. The species tree becomes equipped with a new branch making it an admixture graph. We denote, for example, $3 \to 12$, to indicate that the new branch stems from the terminal branch 3 and then admixes into the branch that is ancestral to populations 1 and 2, as shown in Figure 2 Panel I). From the coalescence point of view, lineages reaching the admixture node have some positive probability of entering the new branch instead. The new branch carrying the gene flow is allowed to be of any length within the constraints of the tree; $1234 \to 1$ in S cannot have zero length, for example. There is reason to think that introgression may typically be mediated by some unsampled *ghost population* (Tricou, Tannier, & de Vienne, 2022), hence we do not assume that the length of the admixture branch is zero even when it could be. The bidirectional arrow $3 \leftrightarrow 12$ is used as a shorthand notation that means either $3 \to 12$ or $12 \to 3$ (although truly bidirectional gene flow actually behaves similarly whenever the symbol $\leftrightarrow$ is invoked).

One difficulty is ensuring that the expected behaviour of Δ-statistics is *consistent* under admixture events of interest. Consider gene flow $3 \to 2$ in tree S. If lineage 2 is introgressed and coalesces with lineage 3 before the populations 3 and 4 merge (Figure 2 Panel II)), then we expect ${}_S\Delta_4$ to be positive as the pattern ABABA cannot be observed at all. On the other hand, if lineage 2 is introgressed but does not coalesce with

lineage 3 by the time populations 3 and 4 merge (Figure 2 Panel III)), pattern ABABA becomes more likely than pattern BAABA (see Lemma 6 in Appendix 2) and ${}_S\Delta_4$ is expected to be negative. This is an example of inconsistency, where the expected sign depends on graph parameters (the probability of coalescence within the terminal branch 3) that we are committed to making no assumptions about. The twenty Δ-statistics (3), (4) and (5) highlighted in the previous section are chosen in such a way that the expected behaviour is always consistent and the ability to identify different gene flow events is maximized, while avoiding the use of singleton patterns to the extent it is possible.

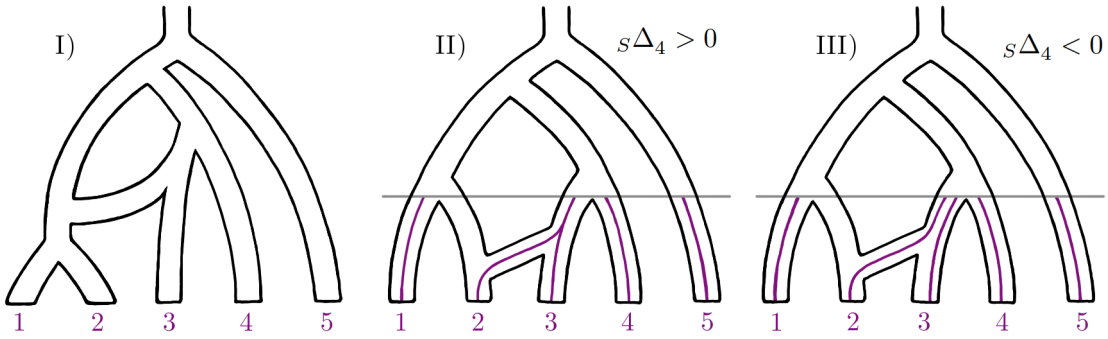

Figure 3: Panel I) depicts the admixture graph meant by $3 \to 12$.
Panels II) and III) demonstrate inconsistency of ${}_S\Delta_4$ under the gene flow scenario $3 \to 2$. In Panel II), 2 and 3 coalesce before meeting 4, and so ${}_S\Delta_4 > 0$. In Panel III), 2 and 3 do not coalesce before meeting 4, and so ${}_S\Delta_4 < 0$.

The expected classifications of Δ-statistics (3), (4) and (5) under every unidirectional gene flow event on trees S, A and Q are collected in Tables 2, 3 and 4. The entry $+$ (resp. $-$) signifies that the expectation is some unknown positive (resp. negative) number, and the entry 0 means it is zero. The proof can be found in Appendix 2. If a gene flow event is not listed in the table, the expected classification of each Δ-statistic is then 0. The vertical dividers indicate that the right sides of the tables depend on the synchronization assumption (2). (See Appendix 3 for collapsed tables that only use the left sides, that is, the Δ-statistics that are independent of (2).) We will not provide the proof here, but the collapsed tables for S and A also work if the data have been filtered so that lineage 5 always carries the ancestral allele; the outgroup mutation assumption (1) is only strictly needed because of admixture involving the root or population 5, and even in that case only the right sides of the tables are affected.

Table 2: Predicted effects on tree $S = (((1,2),(3,4)),5)$. The last four columns assume (2).

	${}_S\Delta^*_{1-6}$	${}_S\Delta^*_{2-6}$	${}_S\Delta^*_{3-5}$	${}_S\Delta^*_{4-5}$	${}_S\Delta^*_{5+7}$	${}_S\Delta^*_{6+8}$	${}_S\Delta^*_{3+4-5+7}$	${}_S\Delta^*_{1+2-6+8}$
$1 \to 3$	$+$	$+$	$+$	0	$-$	$-$	0	0
$3 \to 1$	$+$	0	$+$	$+$	$-$	$-$	0	0
$1 \to 4$	$-$	$-$	0	$+$	$-$	$+$	0	0
$4 \to 1$	$-$	0	$+$	$+$	$-$	$+$	0	0
$2 \to 3$	$+$	$+$	$-$	0	$+$	$-$	0	0
$3 \to 2$	0	$+$	$-$	$-$	$+$	$-$	0	0
$2 \to 4$	$-$	$-$	0	$-$	$+$	$+$	0	0
$4 \to 2$	0	$-$	$-$	$-$	$+$	$+$	0	0
$12345 \to 1$	0	0	$-$	$-$	$+$	0	$+$	0
$34 \leftrightarrow 2$ $1234 \to 1$	0	0	$-$	$-$	$+$	0	0	0
$5 \to 1$	0	0	$-$	$-$	$+$	0	$-$	0
$1 \to 5$	0	0	$-$	$-$	0	0	$-$	0
$12345 \to 2$	0	0	$+$	$+$	$-$	0	$-$	0
$34 \leftrightarrow 1$ $1234 \to 2$	0	0	$+$	$+$	$-$	0	0	0
$5 \to 2$	0	0	$+$	$+$	$-$	0	$+$	0
$2 \to 5$	0	0	$+$	$+$	0	0	$+$	0
$12345 \to 3$	$-$	$-$	0	0	0	$+$	0	$+$
$12 \leftrightarrow 4$ $1234 \to 3$	$-$	$-$	0	0	0	$+$	0	0
$5 \to 3$	$-$	$-$	0	0	0	$+$	0	$-$
$3 \to 5$	$-$	$-$	0	0	0	0	0	$-$
$12345 \to 4$	$+$	$+$	0	0	0	$-$	0	$-$
$12 \leftrightarrow 3$ $1234 \to 4$	$+$	$+$	0	0	0	$-$	0	0
$5 \to 4$	$+$	$+$	0	0	0	$-$	0	$+$
$4 \to 5$	$+$	$+$	0	0	0	0	0	$+$

Table 3: Predicted effects on tree $A = ((((1,2),3),4),5)$. The last column assumes (2).

	${}_A\Delta^*_{1-2}$	${}_A\Delta^*_{1-3}$	${}_A\Delta^*_{2-3}$	${}_A\Delta^*_{1+2-3+4}$
$1 \leftrightarrow 3$ $123 \to 2$	$+$	$+$	0	0
$2 \leftrightarrow 3$ $123 \to 1$	$-$	$-$	0	0
$1 \to 4$	$-$	0	$+$	0
$4 \to 1$	$-$	$-$	$+$	0
$2 \to 4$	$+$	0	$-$	0
$4 \to 2$	$+$	$+$	$-$	0
$1 \to 5$	0	$-$	$-$	$-$
$5 \to 1$	$-$	$-$	$-$	$-$
$1234 \to 1$	$-$	$-$	$-$	0
$12345 \to 1$	$-$	$-$	$-$	$+$
$2 \to 5$	0	$+$	$+$	$+$
$5 \to 2$	$+$	$+$	$+$	$+$
$1234 \to 2$	$+$	$+$	$+$	0
$12345 \to 2$	$+$	$+$	$+$	$-$

Table 4: Predicted effects on tree $Q = (((1,2),3),(4,5))$. The last four columns assume (2).

	${}_Q\Delta^*_{1-6}$	${}_Q\Delta^*_{2-6}$	${}_Q\Delta^*_{3-5}$	${}_Q\Delta^*_{4-5}$	${}_Q\Delta^*_{5+7}$	${}_Q\Delta^*_{6+8}$	${}_Q\Delta^*_{3+4-5+7}$	${}_Q\Delta^*_{1+2-6+8}$
$1 \to 4$	$+$	$+$	$+$	0	$-$	$-$	0	$-$
$4 \to 1$	$+$	0	$+$	$+$	$-$	$-$	$+$	0
$1 \to 5$	$-$	$-$	0	$+$	$-$	$+$	0	$+$
$5 \to 1$	$-$	0	$+$	$+$	$-$	$+$	$+$	0
$2 \to 4$	$+$	$+$	$-$	0	$+$	$-$	0	$-$
$4 \to 2$	0	$+$	$-$	$-$	$+$	$-$	$-$	0
$2 \to 5$	$-$	$-$	0	$-$	$+$	$+$	0	$+$
$5 \to 2$	0	$-$	$-$	$-$	$+$	$+$	$-$	0
$2 \to 45$	0	0	$-$	$-$	$+$	0	0	0
$45 \to 2$	0	0	$-$	$-$	$+$	0	$-$	0
$1 \leftrightarrow 3$ $123 \to 2$	0	0	$-$	$-$	0	0	$-$	0
$12345 \to 2$	0	0	$-$	$-$	$-$	0	$-$	0
$1 \to 45$	0	0	$+$	$+$	$-$	0	0	0
$45 \to 1$	0	0	$+$	$+$	$-$	0	$+$	0
$2 \leftrightarrow 3$ $123 \to 1$	0	0	$+$	$+$	0	0	$+$	0
$12345 \to 1$	0	0	$+$	$+$	$+$	0	$+$	0
$5 \to 12$	$-$	$-$	0	0	0	$+$	0	0
$12 \to 5$	$-$	$-$	0	0	0	$+$	0	$+$
$4 \to 3$	$-$	$-$	0	0	0	0	0	$-$
$3 \to 4$	$-$	$-$	0	0	0	$-$	0	$-$
$4 \to 12$	$+$	$+$	0	0	0	$-$	0	0
$12 \to 4$	$+$	$+$	0	0	0	$-$	0	$-$
$5 \to 3$	$+$	$+$	0	0	0	0	0	$+$
$3 \to 5$	$+$	$+$	0	0	0	$+$	0	$+$
$123 \leftrightarrow 5$ $12345 \to 4$	0	0	0	0	0	$+$	0	$+$
$123 \leftrightarrow 4$ $12345 \to 5$	0	0	0	0	0	$-$	0	$-$

Classification of observed Δ-statistics

With the null hypothesis of no admixture events, and assuming independent loci, the number of observations in L follows the binomial distribution $n(L) \sim B(n(L)+n(R), 0.5)$. Using normal approximation, the number

$$Z = \frac{n(L) - n(R)}{\sqrt{n(L) + n(R)}}$$

has nearly the standard Gaussian distribution. Let α be the maximum tolerated probability of classifications $-$ or $+$ when the null hypothesis holds (type I error). We classify the observed statistic as

$$\begin{cases} + & \text{if} \quad Z \geq \quad \Phi^{-1}(1 - 0.5\alpha)\,, \\ - & \text{if} \quad Z \leq -\Phi^{-1}(1 - 0.5\alpha)\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{6}$$

and as a convention the case $n(L) = n(R) = 0$ as 0. Following D_{FOIL} (Pease & Hahn, 2015), we use the value $\alpha = 0.01$ throughout this paper. A stricter value, $\alpha = 0.0027$, could also be considered because the cutoff $Z = \pm 3$ is essentially a community standard for the classic D-statistic (Green et al., 2010).

In practice, the assumption of independent loci is invalid and linkage disequilibrium will inflate the Z-scores in absolute value. This can be alleviated with a block jackknife procedure (Busing, Meijer, & Leeden, 1999). In the context of D-statistics applied to modern humans and Neanderthals, the block jackknife has been experimentally found to be adequate so long as the blocks are at least 2 Mb in length (Green et al., 2010, Supplementary Online Material 15).

Local ancestry

Introgression happens on the level of complete chromosomes, consequently broken down by recombination. By zooming into smaller windows we can prevent multiple different gene flow events from influencing an analysis. For example, bidirectional gene flow event $2 \leftrightarrow 3$ in tree S produces the signal $(+\,+\,-\,-\,|\,+\,-\,0\,0)$ by Table 2, rows 5 and 6. But so would the mixture of $34 \rightarrow 2$ and $12 \rightarrow 3$ by rows 10 and 22. These two scenarios can be told apart by studying the local ancestry; in the first case we expect some windows with signature $(+\,+\,-\,0\,|\,+\,-\,0\,0)$ and some with signature $(0\,+\,-\,-\,|\,+\,-\,0\,0)$, whereas in the latter case we expect signatures $(0\,0\,-\,-\,|\,+\,0\,0\,0)$ and $(+\,+\,0\,0\,|\,0\,-\,0\,0)$. Because the alternative hypotheses we consider are unidirectional gene flow events, the size of windows to zoom into should be so small that it is reasonable to suppose each window is affected by at most one of the events. We dub this requirement the *single event assumption*.

Single event assumption: The genetic segment the Δ-statistic analysis is performed on only has traces of at most one gene flow event. (7)

Shorter windows are more likely to satisfy the single event assumption (7), but eventually face issues with stochasticity and statistical power. We again turn to D_{FOIL} (Pease & Hahn, 2015), which also adopts a window approach. Pease and Hahn experimentally found a minimum window size recommendation of $L \geq 1/(N_e r)$ kb, where N_e is the effective population size and r the recombination rate per nucleotide and generation. As this length can already be less than 2 Mb, there is certainly no room for a block jackknife, and linkage disequilibrium will influence the results. This poses a challenge for interpretation when the window size is small.

3 Results

Admixture simulations

We simulated gene trees in a coalescence process, where the time until coalescence between two given lineages in the same randomly mating population follows an exponential distribution with parameter λ. A unit of time is then about $2\lambda N_e$ generations, where N_e is the effective population size. Number of mutations on a

(mutation rate weighted) simulated gene tree is then sampled from a Poisson distribution and the location of mutations within it from a uniform distribution; resulting in biallelic patterns. Compared to simulating genetic regions under recombination, the simple coalescent approach is a compromise because it does not inform us on the window size necessary to infer gene flow events. For this question we rely on Pease and Hahn (2015), who conclude that in order to D_{FOIL} to work well, $N_e rL$ should be at least one thousand, where r is the recombination rate per site per generation and L is the window size.

We tested all unidirectional gene flow scenarios that have non-zero expected signatures in trees S, A and Q, for a total of 32, 18 and 34 simulations, respectively. The number of patterns simulated for each scenario was 1,000,000 (presented here), 100,000 or 10,000 (presented in Appendix 4). We used $\lambda = 1$, with the last common ancestor of all five populations living at time point 3 (for example, if generation time is 10 years and $N_e = 10,000$, this translates into 600,000 years). For mutations the Poisson parameter per unit of time was 0.0001 (corresponding to per-year-per-locus mutation rate of 0.5×10^{-9} in the example given above). The other nodes of the species tree were placed roughly evenly, and the admixture events were instantaneous when possible and mediated by a ghost population otherwise, with admixture proportion at 10%. For details, see documentation on our GitHub page (`https://github.com/KalleLeppala/Delta-statistics`).

In all but one scenarios with 1,000,000 patterns the observed Δ-statistics matched the predictions in Tables 2, 3 and 4 (results plotted in Figures 4, 5 and 6). With 100,000 independent samples, the Δ-statistics relying on the synchronization assumption (2) became unreliable (Appendix 4), while the left hand sides of the Tables 2, 3 and 4 (collapsed versions can be found in Appendix 3) remained accurate. With only 10,000 independent samples, the Δ-statistics no longer functioned well (Appendix 4).

Compared with the D_{FOIL} method applied to the same simulated patterns on the symmetric tree S (Appendix 5), the Δ-statistics had similar accuracy with all three sample sizes, while the Δ-statistics limited to Table S1 in Appendix 3 had considerable advantage in power.

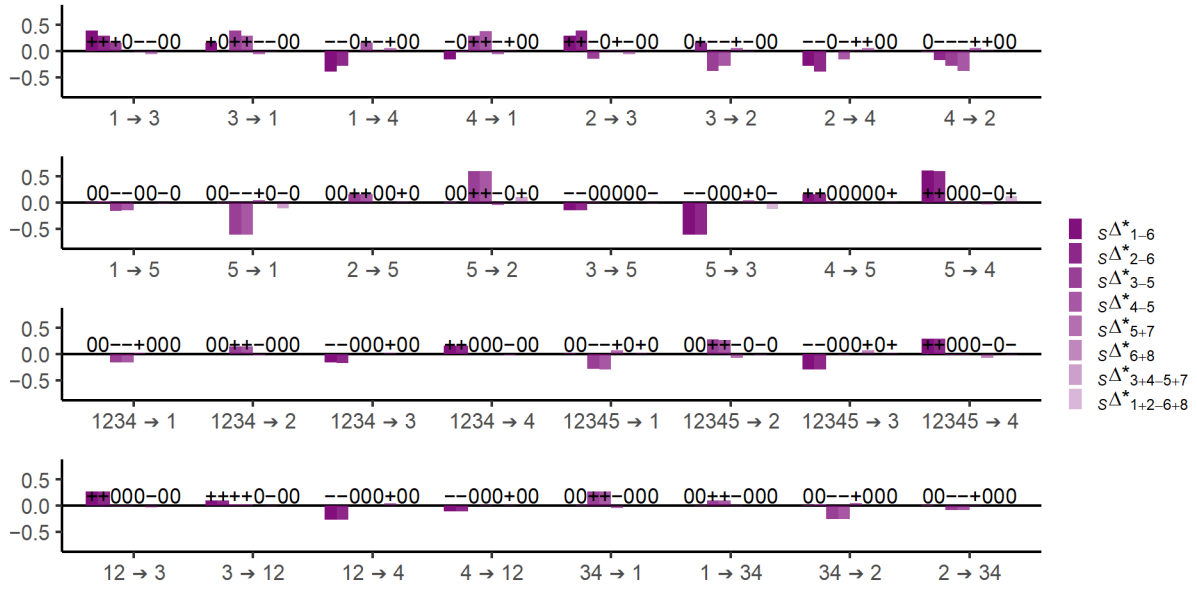

Figure 4: Simulation results for the symmetric tree S. All results match the predictions in Table 2 except for the signature of $3 \to 12$ where (dependent) statistics ${}_S\Delta_{3-5}$ and ${}_S\Delta_{4-5}$ were supposed to be zero — unsurprising given the total number of true null hypotheses we test using the probability $\alpha = 0.01$ of wrong classification.

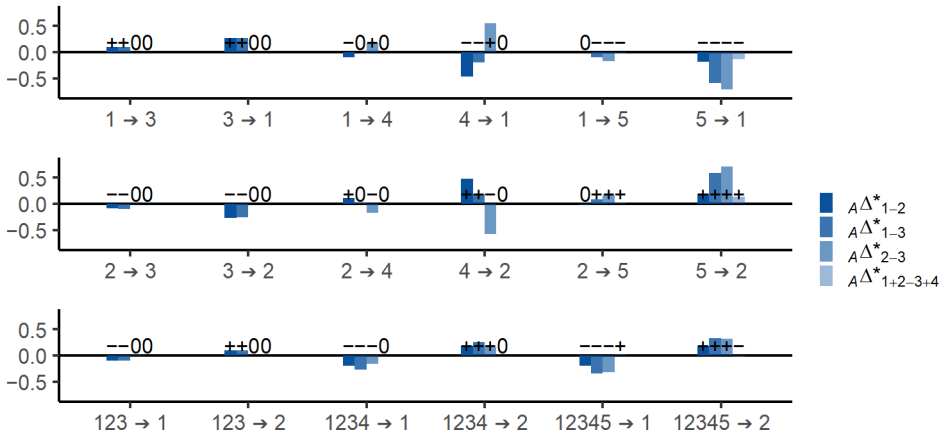

Figure 5: Simulation results for the asymmetric tree A. All results match the predictions in Table 3.

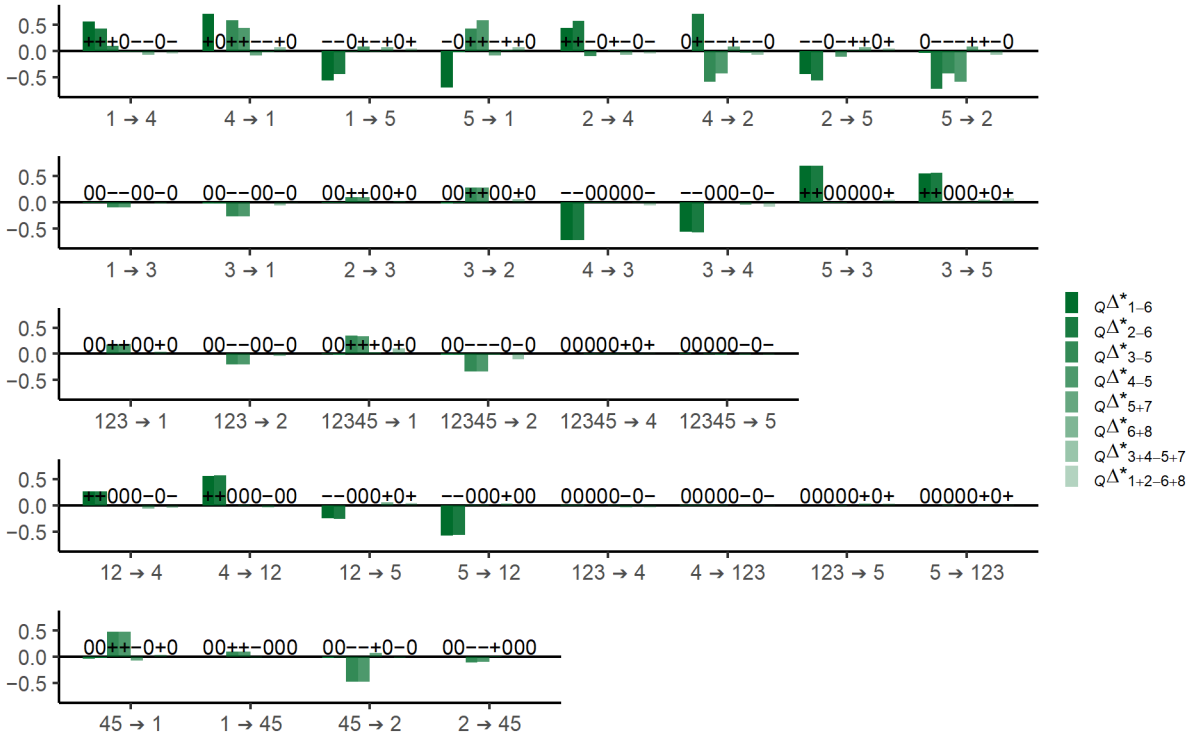

Figure 6: Simulation results for the quasisymmetric tree Q. All results match the predictions in Table 4.

An empirical example: Brown bear – polar bear gene flow

In previous research, we investigated the problem of potential inheritance of a brown bear mitochondrial genome (mtDNA), specifically through brown bears from the Admiralty, Chichagof and Baranof islands of Southeast Alaska (ABC brown bears), by the ancestor of all modern polar bears. As recognized earlier (Talbot & Shields, 1996), all polar bear mtDNAs, also including that of an ancient 120,000 year old sample from Svalbard, was nested inside brown bear mtDNA haplotype diversity (Lindqvist et al., 2010, Figure 3), which suggested either a recent divergence of polar bears from brown bears around 150,000 years ago, or that they acquired the brown bear mitochondrion via ILS or gene flow. Sequencing and analysis of the ancient

polar bear's nuclear genome resolved an older divergence date and that admixture between the species had indeed occurred (Miller et al., 2012). Studies by other investigators came to similar conclusions, with the specification that the direction of admixture was principally or even exclusively from polar into brown bear, an assertion that has been predominant since (Cahill et al., 2013, 2015, 2018). Our own later research (Lan et al., 2022), using five- and six-leaf admixture graphs and data from the ancient Svalbard sample, strongly supported bidirectional admixture between these bear species, but with brown bear $\rightarrow$ polar bear being the predominant pattern. Almost concurrently, another investigator group published a genome of their own ancient polar bear sample, named "Bruno", and examined admixture between polar and brown bears using the D_{FOIL} method (Wang et al., 2022). Using the tree S, which in fact principally provided an indecisive result when it came to gene flow direction (see their Fig. 3, which is in log scale), they preferred a mostly unidirectional polar bear $\rightarrow$ brown bear admixture scenario.

Here, we reevaluate their findings and our own earlier results on brown bear – polar bear gene flow using Δ-statistics. For this purpose, we started with the same Variant Call Format (VCF) file used in our 2022 work (Lan et al., 2022), adding to it a newly constructed genotype VCF for the Bruno sample (obtaining its genome data from the Short Read Archive of the National Center for Biotechnology Information Sequence Reads Archive). Data processing, variant calling, and VCF construction and filtering (using the same parameters as for `Dataset02`) followed the methodology outlined by Lan et al., 2022. The final VCF was converted into a variant-major additive component file (`--recode A-transpose`) using PLINK 2.0 (Chang et al., 2015).

Using PCA and TreeMix analyses, we confirmed a similar ancestry of the Bruno sample as our ancient polar bear (dubbed APB) (Appendix 6). From those results, we could conclude that APB and Bruno were both slightly (and almost equivalently) more closely related to brown bears than modern polar bears are. As such, we were also able to infer that our incorporation of the Bruno data was successful. We conducted analyses of Δ-statistics using the following populations: brown bears from Admiralty, Baranof, and Chichagof Islands (ABC), European brown bears (EBB), North American mainland brown bears (BB), polar bears (PB from Svalbard and AK from Alaska), black bears (BLK), and both ancient individuals APB and Bruno. The following combinations were tested:

- Symmetric: $S = (((1 = \mathrm{ABC}, 2 = \mathrm{EBB/BB}), (3 = \mathrm{APB/AK/Bruno}, 4 = \mathrm{PB})), 5 = \mathrm{BLK})$
- Asymmetric: $A = ((((1 = \mathrm{ABC}, 2 = \mathrm{BB}), 3 = \mathrm{EBB}), 4 = \mathrm{APB/Bruno/PB}), 5 = \mathrm{BLK})$
- Quasisymmetric: $Q = (((1 = \mathrm{ABC}, 2 = \mathrm{BB}), 3 = \mathrm{EBB}), (4 = \mathrm{APB/AK/Bruno}, 5 = \mathrm{PB}))$

These combinations can be seen as various five-leaf perspectives, together giving a better picture of the full phylogenetic tree $((((\mathrm{ABC}, \mathrm{BB}), \mathrm{EBB}), (\mathrm{APB/Bruno}, (\mathrm{AK}, \mathrm{PB}))), \mathrm{BLK})$ than separately. The analyses involved non-overlapping windows of varying sizes (100 kb, 500 kb, 1 Mb, 5 Mb, and 10 Mb) (Supplementary Tables S5–S10). To obtain genome-wide values, each combination was also computed using jackknife blocks of the same intervals (Supplementary Tables S11–S13). For results from combinations not presented in Figures 7 and 8 see Appendix 6 and Supplementary Tables S5–S13.

Starting with the symmetric tree, $S = (((1 = \mathrm{ABC}, 2 = \mathrm{EBB}), (3 = \mathrm{APB/AK}, 4 = \mathrm{PB})), 5 = \mathrm{BLK})$ (Figures 7 and 8, top rows), we find strong support for the event $34 \leftrightarrow 1^*$ (the asterisk emphasizing that there are other possible explanations besides the chosen representative $34 \leftrightarrow 1$). Some of this signal could be explained by gene flow between EBB and BLK (Kumar et al., 2017), but not all, as can be seen by comparing the two top left plots. Most of the signal is therefore either due to ancient admixture between brown bears and polar bears, or to ghost relatives of all brown bears and polar bears (such as cave bears (Barlow et al., 2018)) introgressing into EBB ($1234 \rightarrow 2$). There is indeed a clear signal of modern gene flow from PB into ABC when leaf $3 = \mathrm{APB}$ (Figure 7). However, this signal dissolves when it is replaced with a modern polar bear individual (Figure 8), presumably because the split between populations 3 and 4 defining "modernity" is pulled towards the present time. We also wish to address the apparent conversion of $34 \leftrightarrow 1^*$ into $5 \rightarrow 2$ as the window size is increased in Figure 8. We interpret this as a violation of the single event assumption (7). As it happens, the signature of $34 \leftrightarrow 1^*$ is $(0\,0 + + \,|\, - 0\,0\,0)$ and the signature of $5 \rightarrow 2$ is $(0\,0 + + \,|\, - 0 + 0)$. A window containing segments from both events will produce the mixture signature $(0\,0 + + \,|\, - 0 + 0)$ and be classified as $5 \rightarrow 2$, even if $34 \leftrightarrow 1^*$ was dominant in terms of admixture proportion. On the same note, the mixture of signals $4 \rightarrow 1$ and $34 \leftrightarrow 1^*$ in Figure 7 is dominated by $4 \rightarrow 1$, as can be observed comparing the two window sizes. Overall, tree S provides evidence for modern polar

bear $\rightarrow$ brown bear gene flow, and for admixture between European brown bears and black bears, but the strongest signal remains ambiguous in nature. This result is consistent when replacing APB with Bruno (Figure S16 in Appendix 6), and with D_{FOIL} results achieved by Wang et al., 2022 and ourselves (Figures S19, S20 and discussion in Appendix 6). The issue with interpreting the strongest signal is in the tree S itself; even if we are looking at brown bear – polar bear admixture, the direction of gene flow can only be inferred if it is recent enough to affect only one of the two polar bear samples. As we currently do not have unadmixed ancient polar bear specimens available, we overcome this methodological problem by studying the asymmetric tree A, which does not require an unadmixed sister sample.

The asymmetric tree, $A = ((((1 = \mathrm{ABC}, 2 = \mathrm{BB}), 3 = \mathrm{EBB}), 4 = \mathrm{APB/PB}), 5 = \mathrm{BLK})$ (Figures 7 and 8 middle rows), provides evidence for gene flow in both directions between ABC brown bears and polar bears. Under bidirectional gene flow, the mixture of signatures $(- - + \,|\, 0)$ for $4 \rightarrow 1$ and $(- \, 0 + \,|\, 0)$ for $1 \rightarrow 4$ looks like the former; this dominance also applies without the singleton patterns. Thus, even tree A can be biased towards polar bear $\rightarrow$ brown bear gene flow, which contributes to the difference between results using 500 kb versus 5 Mb windows. Another contributing factor is that larger window sizes will favour more recent gene flow events whose traces are less uniformly distributed across the genome. Hence, similar to our own 2022 findings, the bidirectional brown – polar bear admixture appears to have a temporal element where brown bear $\rightarrow$ polar bear is older than polar bear $\rightarrow$ brown bear gene flow. Considering that tree A cannot theoretically detect admixture between EBB and BLK ($3 \leftrightarrow 5$) or the ghost relatives of all brown bears and polar bears ($1234 \rightarrow 3$), we have now recognized that the strong ambiguous signal from Wang et al., 2022 and our own symmetric tree S analyses was indeed largely due to ancient bidirectional gene flow between polar bears and ABC brown bears. Additionally, we detect gene flow or ILS within the brown bear clade (Lan et al., 2022).

We also tried the quasisymmetric tree, $Q = (((1 = \mathrm{ABC}, 2 = \mathrm{BB}), 3 = \mathrm{EBB}), (4 = \mathrm{APB/AK}, 5 = \mathrm{PB}))$ (Figures 7 and 8 bottom rows). Using the ancient sample APB without singleton patterns, the results are very similar to those obtained with the symmetric tree S: a strong but ambiguous $45 \leftrightarrow 1^*$ signal and some evidence for recent polar bear $\rightarrow$ brown bear admixture, particularly in 5 Mb windows (Figure 7) and when modern polar bears occupy both leaves 4 and 5 (Figure 8). Using two modern polar bear samples and incorporating singleton patterns (Figure 8), which would otherwise be inappropriate for ancient samples, the results resemble those obtained with the asymmetric tree A. This provides clear evidence for ancient bidirectional gene flow between polar bears and brown bears, and signs of admixture within brown bears, with the two polar bear samples not separating from each other at all.

This empirical example also highlights the importance of the synchronization assumption (2) with D_{FOIL} and some of the new Δ-statistics. The signature of gene flow from Bruno into North American brown bears detected by Wang et al., 2022 (gray box plots in their Figure 3) was several times more prevalent than the signature of gene flow from modern polar bears into North American brown bears (ochre box plots in their Figure 3). Such results can be explained simply by noting that as an ancient sample, Bruno has less private mutations than the modern polar bear sample, thereby biasing D_{FOIL}, since the synchronization assumption (2) is broken. In contrast, Δ-statistics without singleton patterns on comparable trees (top left panel Figure 7 using APB and Figure S16 in Appendix 6 using Bruno) detect substantially more gene flow (190 windows versus 31, and 118 windows versus 61) from the modern sample than from the ancient one. This is what we would expect under occasional hybridization continuing to this day. We also tested what would have happened if we had used the full Tables 2–4 with the ancient sample APB; the dramatically biased results (Figure S17 and S18) can be found in Appendix 6.

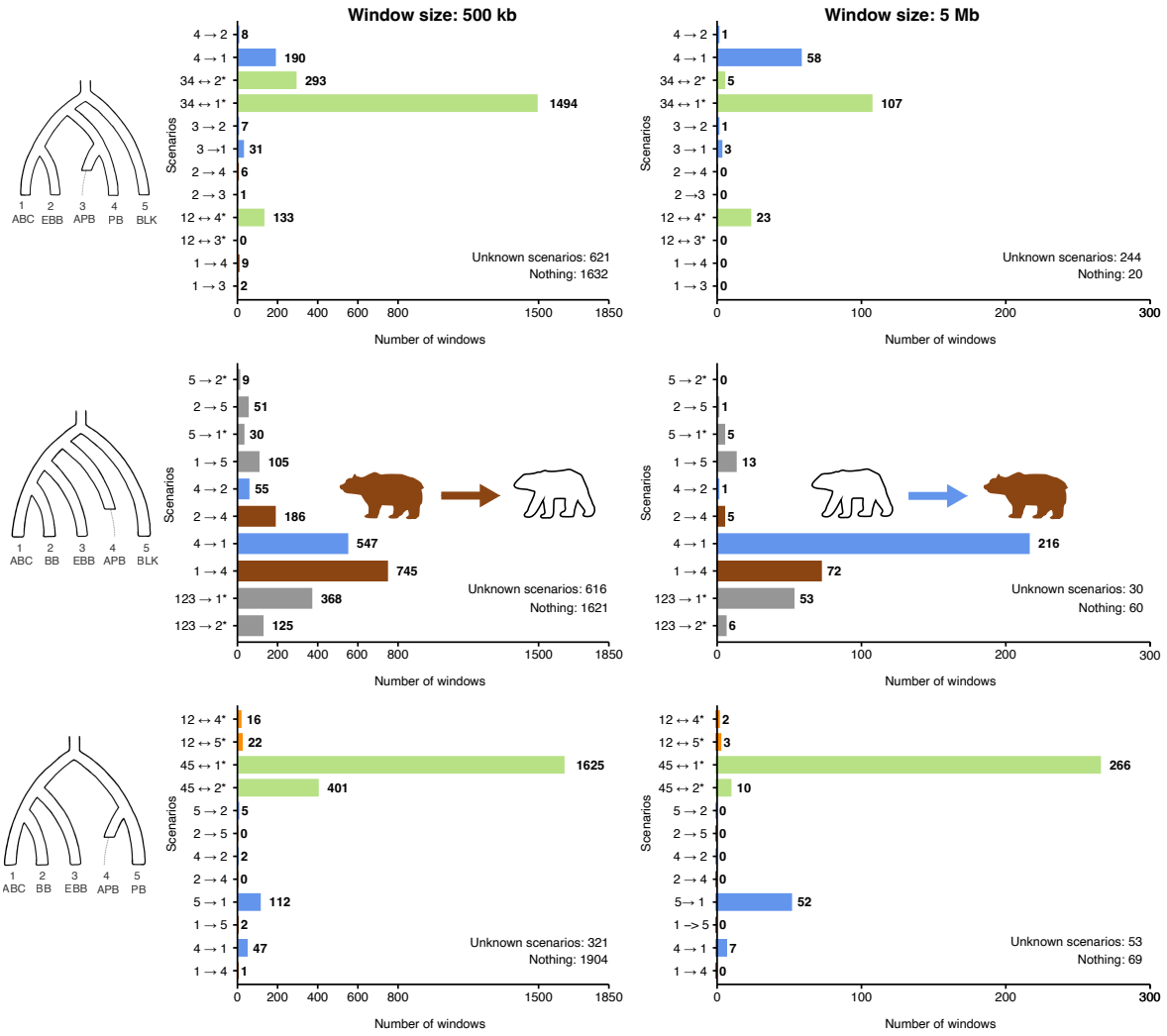

Figure 7: Bar plots of various Δ-statistic analyses of brown bear – polar bear gene flow incorporating a modern and an ancient polar bear (APB). Because APB is an ancient sample, we cannot make the synchronization assumption (2), so we therefore exclude singletons and use the collapsed Tables S1–S3 in Appendix 3. The asterisk means there are additional gene flow events consistent with the signature to the representative written, for example $34 \leftrightarrow 2^*$ in tree S stands for $12345 \rightarrow 1$, $34 \leftrightarrow 2$, $1234 \rightarrow 1$ or $1 \leftrightarrow 5$, according to Table S1 in Appendix 3. The colours of the bars signify the following: Brown — the signature supports only gene flow from brown bears into polar bears; Blue — the signature supports only gene flow from polar bears into brown bears; Orange — the signature supports gene flow in either direction between brown bears and polar bears; Gray — the signature supports some gene flow event that has nothing to do with the brown bear – polar bear question; Green — the signature supports gene flow of either or neither direction between brown bears and polar bears. For example, $34 \leftrightarrow 2^*$ is Green in tree S because it's consistent with gene flow from brown bears into polar bears ($2 \rightarrow 34$), with gene flow from polar bears into brown bears ($34 \rightarrow 2$), and with gene flow that has nothing to do with the brown bear – polar bear question (such as $1234 \rightarrow 1$). The windows with zero signature are counted under "Nothing", and windows with non-zero signature are discordant with any row in Tables S1–S3 in Appendix 3 under "Unknown scenarios". See the main text for description of the populations and interpretation.

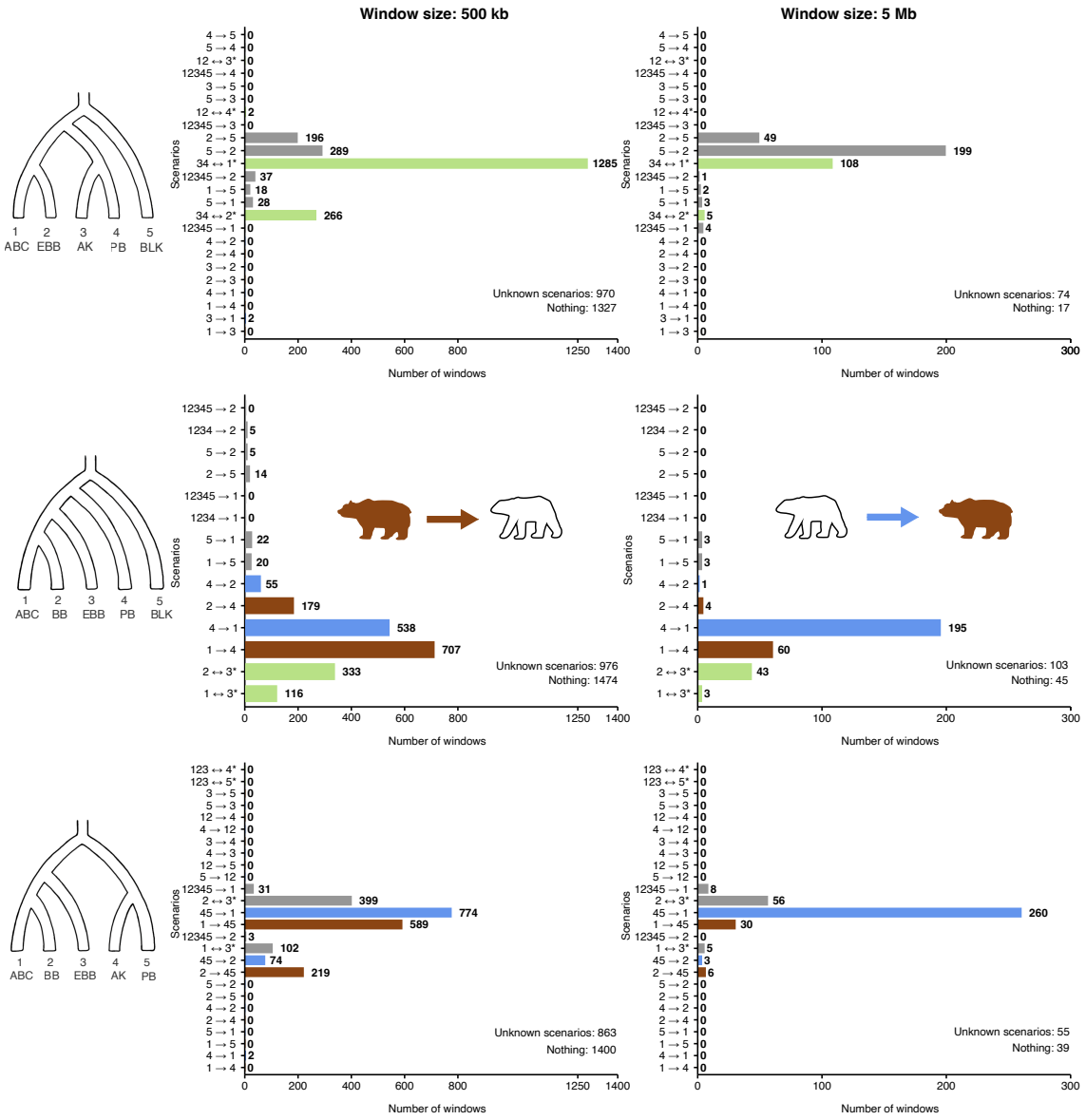

Figure 8: Bar plots of various Δ-statistic analyses of brown bear – polar bear gene flow incorporating two modern polar bears. Because all the samples are modern, we commit to the synchronization assumption (2) (although the generation times between populations likely vary a little), and therefore we use the complete Tables 2–4 in the main text. The asterisk means there are additional gene flow events consistent with the signature to the representative written, for example $34 \leftrightarrow 2^*$ in tree S stands for $12345 \rightarrow 1$, $34 \leftrightarrow 2$, $1234 \rightarrow 1$ or $1 \leftrightarrow 5$, according to Table S1 in Appendix 3. The colours of the bars are as in Figure 7. The windows with zero signature are counted under "Nothing", and windows with non-zero signature are discordant with any row in Tables S1–S3 in Appendix 3 under "Unknown scenarios". See the main text for description of the populations and interpretation.

We also tried the analysis without windows for all trees, instead blocks of various sizes (100 kb, 500 kb, 1 Mb, 5 Mb, 10 Mb) for jacknifing (Busing et al., 1999), reported in Tables S11–S13. The predominant gene flow event detected was from modern polar bears into ABC bears, although when using the (incompatible) combination of singleton patterns and the ancient sample, the method would often produce a signature that doesn't match any unidirectional gene flow event. The result highlights the single event assumption (7): as

discussed above, the combination of $4 \to 1$ and $34 \leftrightarrow 1$ in tree S looks like $4 \to 1$, the combination of $4 \to 1$ and $1 \to 4$ in tree A looks like $4 \to 1$, and the combination of $5 \to 1$ and $45 \leftrightarrow 1$ in tree Q looks like $5 \to 1$ — a phenomenon already visible in Figures 7 and 8. Without using the window based approach, a researcher might unknowingly break the single event assumption (7) and miss out on the true complexity.

Finally, to confirm that our dataset shows admixture signals comparable to Wang et al., 2022, and to compare our results from Δ-statistics with D_{FOIL}, we ran D_{FOIL} using our expanded D_{FOIL} prediction table (Appendix 5). We tested the symmetric tree $S = (((1 = \text{ABC}, 2 = \text{EBB}), (3 = \text{X}, 4 = \text{PB})), 5 = \text{BLK})$, where three individuals each for populations ABC, EBB, and PB were incorporated in independent runs of all combinations, and X could represent either APB, Bruno, a modern polar bear AK34, or three different modern polar bears from the AK group. The same window sizes were used as in our Δ-statistic analyses (100 kb, 500 kb, 1 Mb, 5 Mb, 10 Mb), and we calculated the absolute and log_{10} numbers of windows in each scenario to compare our results with Wang et al., 2022.

Our D_{FOIL} analyses yielded results consistent with Wang et al., 2022 regardless of whether X represented a single polar bear (AK34) or three polar bears, except a wider range was observed in the box plots when more than one modern polar bear individual was in the third position (Appendix 6, Figure S19 and S20). Across all combinations, bidirectional admixture scenarios between polar and brown bears received the strongest support. Notably, there was a discrepancy between modern and ancient polar bears in terms of the number of windows supporting most scenarios, which is consistent with results from (Wang et al., 2022).

4 Discussion

We have developed a new method for gene flow detection — Δ-statistics — that can be seen as a natural generalization of the classic D-statistic to five-leaf trees. Additionally, Δ-statistics provide opportunities to discern admixture directions. Like the D-statistic, we worked with minimal assumptions and avoided parametrizing any unnecessary features of the species tree. Instead, we leveraged symmetry that resides in all trees, even the "asymmetric" tree A, to construct difference statistics whose expected signs are informative about admixture direction. We point out that while the good behaviour of the statistics under the null hypotheses relies on *symmetry*, their good behaviour under various gene flow events to some degree relies on *asymmetry* as well; for example, the direction of gene flow $12 \leftrightarrow 4$ can be worked out in tree Q but not in tree S, ultimately because the position of the root in Q is less symmetric. The insight that we are drawing on the interplay between symmetry and asymmetry can be compared to the qpAdm method (part of the ADMIXTOOLS package (Patterson et al., 2012), see (Harney, Patterson, Reich, & Wakeley, 2021)), where the *reference* populations must be asymmetrically aligned with respect to the *source* and *target* populations.

To compare our method to our inspiration, D_{FOIL} (Pease & Hahn, 2015), we must stress that D_{FOIL} makes the synchronization assumption (2), whereas the classic D does not, and we can either choose to make it or not with Δ-statistics. Without the synchronization assumption (2), our method is applicable to ancient samples and populations with unequal generation times, and Δ-statistics have more statistical power than D_{FOIL} does. For the ability to infer gene flow events we should compare Table S1 in Appendix 3 to Table S4 in Appendix 5, because the original D_{FOIL} paper considered several gene flow events as being outside the realm of possibilities. The only difference in the tables is that D_{FOIL} is blind to gene flow from other terminal branches into 5, while Δ-statistics can detect it but not distinguish it from a number of other gene flow events — and which is better is context dependent. On the other hand, if we make the synchronization assumption (2), our ability to distinguish different gene flow events is superior (full Table 2 in the main text), while statistical power remains comparable to D_{FOIL}.

In addition to highlighting the limitations of D_{FOIL}, we here discuss inaccuracies in another attempt to discern $3 \to 2$ and $2 \to 3$ in tree S — the Partitioned D method (Eaton & Ree, 2013). The proposed statistics are $(D_{12} = {}_S\Delta^*_5, D_1 = -{}_S\Delta^*_3, D_2 = -{}_S\Delta^*_4)$, of which D_{12} is supposedly informative about gene flow direction. Simulations presented in their Table 3 result in signatures $(\pm\pm 0)$ for $3 \to 2$ (tests 4.1–4.6) and (0 ± 0) for $2 \to 3$ (tests 6.1–6.6), where $\pm$ means a significantly non-zero statistic. However, as we have seen in Appendix 2, the actual expected signatures are $(++?)$ for $3 \to 2$ and $(++-)$ for $2 \to 3$, where the question mark stands for an inconsistent statistic whose sign depends on graph parameters. Hence, D_{12} is not informative about gene flow direction, and the ability of the Partitioned D method to infer direction relies on benevolent graph parameters and insufficient power to detect the correct signature. The Eaton

& Ree paper juxtaposes the method against an alternative idea of comparing several classic D-statistics calculated on a four-population subtree, which in fact is a valid technique. Gene flow into the unsampled population cannot be detected, whereas gene flow from the unsampled population is effectively the same as gene flow from its sister population. This is why tests 2.4–2.7 in their Table 2 are consistent with the null $D = 0$, but tests 3.4–3.7 are not. In fact, by the law of total probability, ignoring one of the populations 1–4 in tree S and computing the classic D among the rest is equivalent to using the Δ-statistics ${}_S\Delta^*_{1-6}$, ${}_S\Delta^*_{2-6}$, ${}_S\Delta^*_{3-5}$ and ${}_S\Delta^*_{4-5}$.

One more point of comparison to Δ-statistics is the D_{GEN} method (Elworth, Allen, Benedict, Dulworth, & Nakhleh, 2018). The D_{GEN}-statistic is defined using equal probability sets, as the Δ-statistics are. However, the user is expected to know the actual probabilities of each allelic pattern in order to construct the statistic as a sum of differences expected to be zero under the null hypothesis, and in turn positive under the single alternative hypothesis. This makes the D_{GEN} method sensitive to graph parameters such as branch lengths and coalescence rates within branches of the species tree (recall that it is in general ambiguous which of the patterns BAABA and ABABA become more common under gene flow $3 \to 2$ in tree S), so D_{GEN} cannot be considered a true generalization of the classic D-statistic.

Finally, using an empirical example of brown and polar bear admixture that has been thoroughly explored over the last 10 years, we find support for our own principal conclusions (Lan et al., 2022) that admixture between the two species has been most evidently bidirectional following their initial split into distinct species. Furthermore, using Δ-statistics with window sizes varying by a factor of 10, we uncovered what appear to be at least two episodes of gene flow that differed in their directional biases, only the most recent bias favoring the frequently cited hypothesis of polar bear into ABC brown bear introgression.

Acknowledgements

We are grateful for useful feedback from Dr. Jeffery Ross-Ibarra. The authors acknowledge financial support from U.S. National Science Foundation awards 2221988 and 2139311 to C.L. and 2030871 to V.A.A.
K.L. was supported by Academy of Finland Center of Excellence in Tree Biology (TreeBio AoF CoE 346139).

References

Barlow, A., Cahill, J. A., Hartmann, S., Theunert, C., Xenikoudakis, G., Fortes, G. G., … others (2018). Partial genomic survival of cave bears in living brown bears. *Nature ecology & evolution*, *2*(10), 1563–1570.

Busing, F. M., Meijer, E., & Leeden, R. V. D. (1999). Delete-m jackknife for unequal m. *Statistics and Computing*, *9*(1), 3–8.

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

Cahill, J. A., Heintzman, P. D., Harris, K., Teasdale, M. D., Kapp, J., Soares, A. E., … others (2018). Genomic evidence of widespread admixture from polar bears into brown bears during the last ice age. *Molecular Biology and Evolution*, *35*(5), 1120–1129.

Cahill, J. A., Stirling, I., Kistler, L., Salamzade, R., Ersmark, E., Fulton, T. L., … Shapiro, B. (2015). Genomic evidence of geographically widespread effect of gene flow from polar bears into brown bears. *Molecular ecology*, *24*(6), 1205–1217.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*(1), s13742–015.

Eaton, D. A., & Ree, R. H. (2013). Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: *Orobanchaceae*). *Systematic biology*, *62*(5), 689–706.

Elworth, R. A. L., Allen, C., Benedict, T., Dulworth, P., & Nakhleh, L. (2018). D_{GEN}: A test statistic for detection of general introgression scenarios. *BioRxiv*, 348649.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … others (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722.

Harney, É., Patterson, N., Reich, D., & Wakeley, J. (2021). Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics*, *217*(4), iyaa045.

Harris, A. M., & DeGiorgio, M. (2017). Admixture and ancestry inference from ancient and modern samples through measures of population genetic drift. *Human Biology*, *89*(1), 21–46.

Huson, D. H., Klöpper, T., Lockhart, P. J., & Steel, M. A. (2005). Reconstruction of reticulate networks from gene trees. In *Research in computational molecular biology: 9th annual international conference, recomb 2005, cambridge, ma, usa, may 14-18, 2005. proceedings 9* (pp. 233–249).

Kulathinal, R. J., Stevison, L. S., & Noor, M. A. (2009). The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS genetics*, *5*(7), 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. *Scientific Reports*, *7*(1), 46487.

Lan, T., Leppälä, K., Tomlin, C., Talbot, S. L., Sage, G. K., Farley, S. D., … others (2022). Insights into bear evolution from a pleistocene polar bear genome. *Proceedings of the National Academy of Sciences*, *119*(24), e2200016119.

Leppälä, K., Nielsen, S. V., & Mailund, T. (2017). admixturegraph: an R package for admixture graph manipulation and fitting. *Bioinformatics*, *33*(11), 1738–1740.

Lindqvist, C., Schuster, S. C., Sun, Y., Talbot, S. L., Qi, J., Ratan, A., … others (2010). Complete mitochondrial genome of a pleistocene jawbone unveils the origin of polar bear. *Proceedings of the National Academy of Sciences*, *107*(11), 5053–5057.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., … others (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science*, *338*(6104), 222–226.

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

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., … Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093.

Pease, J. B., & Hahn, M. W. (2015). Detection and polarization of introgression in a five-taxon phylogeny. *Systematic biology*, *64*(4), 651–662.

Talbot, S. L., & Shields, G. F. (1996). Phylogeography of brown bears (*Ursus arctos*) of Alaska and paraphyly within the *Ursidae*. *Molecular phylogenetics and evolution*, *5*(3), 477–494.

Tricou, T., Tannier, E., & de Vienne, D. M. (2022). Ghost lineages highly influence the interpretation of introgression tests. *Systematic Biology*.

Wang, M.-S., Murray, G. G., Mann, D., Groves, P., Vershinina, A. O., Supple, M. A., … others (2022). A polar bear paleogenome reveals extensive ancient gene flow from polar bears into brown bears. *Nature Ecology & Evolution*, *6*(7), 936–944.