# A Three-Sample Test for Introgression

Matthew W. Hahn*[,1,2] and Mark S. Hibbins[1]

[1]Department of Biology, Indiana University, Bloomington, IN

[2]Department of Computer Science, Indiana University, Bloomington, IN

***Corresponding author:** E-mail: mwh@indiana.edu.

**Associate Editor**: Claus Wilke

## Abstract

**Many methods exist for detecting introgression between nonsister species, but the most commonly used require either a single sequence from four or more taxa or multiple sequences from each of three taxa. Here, we present a test for introgression that uses only a single sequence from three taxa. This test, denoted $D_3$, uses similar logic as the standard *D*-test for introgression, but by using pairwise distances instead of site patterns it is able to detect the same signal of introgression with fewer species. We use simulations to show that $D_3$ has statistical power almost equal to *D*, demonstrating its use on a data set of wild bananas (*Musa*). The new test is easy to apply and easy to interpret, and should find wide use among currently available data sets.**

***Key words:*** **hybridization, ABBA-BABA, Patterson's *D*, gene flow, admixture.**

Letter

Genome-scale data have revealed extensive evidence for postspeciation introgression across the tree of life (reviewed in Mallet et al. 2016). Many of these analyses have been carried out in a phylogenetic context, using only a single sample from each population or species. Some methods use gene tree topologies themselves as input (e.g., Huson et al. 2005; Meng and Kubatko 2009; Yu et al. 2011; Edelman et al. 2018), whereas others use counts of shared derived alleles that reflect the underlying topologies (e.g., Green et al. 2010; Lohse and Frantz 2014; Pease and Hahn 2015).

All of these methods depend on the expectation under incomplete lineage sorting (ILS) that the two less-frequent topologies in a rooted triplet should be equal in frequency. Asymmetry in gene tree topologies is taken as evidence for introgression, though ancestral population structure can produce similar patterns (Slatkin and Pollack 2008; Durand et al. 2011; Lohse and Frantz 2014). Importantly, the need to distinguish among topologies or between ancestral and derived sites using these methods means that at least four taxa must be sampled, and sometimes more (e.g., Pease and Hahn 2015; Elworth et al. 2018).

Here, we present a test for introgression that only requires a single sample from each of three taxa. With three taxa we cannot infer the frequencies of alternative gene tree topologies. Instead, our test is based on a related prediction of the ILS model: that there is also an expected symmetry in the branch lengths among topologies. While previous papers have used this expectation informally as an argument for gene flow (e.g., Brandvain et al. 2014), we develop an explicit model and test statistic based on pairwise distances to detect the presence of introgression.

## New Approaches

### A Test for Introgression

Assume that lineages *A* and *B* are sister in the species tree, with divergence time $t_1$ (measured in units of $2N$ generations), and that the ancestor of *A* and *B* split from lineage *C* at time $t_2$ (fig. 1*a*). We refer to gene trees having this topology as *AB*, such that the two discordant topologies are *AC* and *BC* (fig. 1*b* and *c*, respectively).

When ILS is the only cause of gene tree incongruence, topology *AB* may be generated in two different ways, with different expected frequencies and branch lengths. Looking backwards in time, we refer to the topology in which lineages *A* and *B* coalesce before $t_2$ as *AB*1 (this is the history shown in fig. 1*a*). Alternatively, the same topology can occur when these lineages coalesce in the ancestral population of all three lineages; we refer to this topology as *AB*2.

The expected frequencies of these four topologies are (Hudson 1983):

$$E[f_{AB2}] = E[f_{AC}] = E[f_{BC}] = (1/3)e^{-(t_2-t_1)} \quad (1)$$

$$E[f_{AB1}] = 1 - e^{-(t_2-t_1)} \quad (2)$$

As mentioned in the first three paragraphs, here we see that the two discordant topologies (*AC* and *BC*) are expected to have the same frequencies. See chapter 9 in Hahn (2018) for more details on the underlying assumptions of this model.

The same model leads naturally to expectations for the times to coalescence between lineages in each of the different topologies. Here, we focus on the expected times to coalescence between *B* and *C* ($t_{B-C}$) and between *A* and *C* ($t_{A-C}$). These times are (Hibbins and Hahn 2019):

$$E[t_{B-C}|AB1] = E[t_{A-C}|AB1] = t_2 + 1 \quad (3)$$

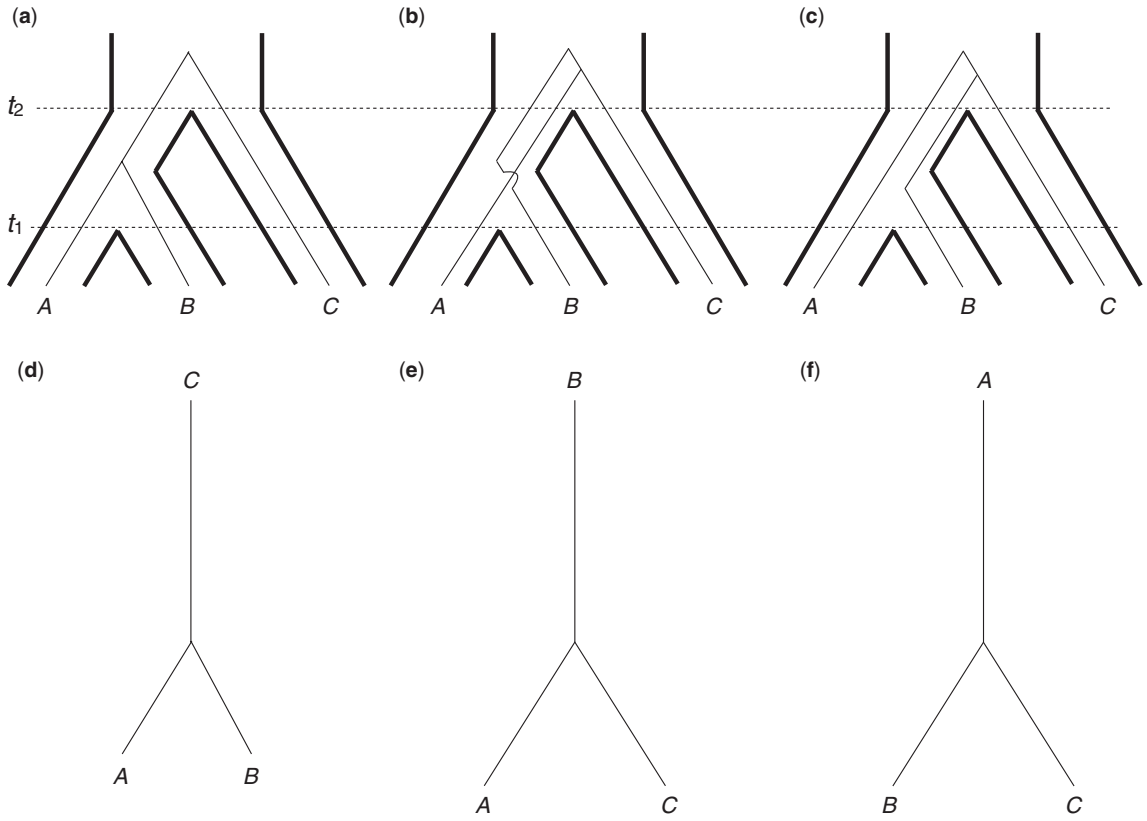


FIG. 1. Topologies produced by incomplete lineage sorting. The top row shows the same species tree (thick lines, with divergence times denoted by $t_1$ and $t_2$) within which three different topologies arise: (*a*) *AB*1, (*b*) *AC*, and (*c*) *BC*. The bottom row (*d-f*) shows the same unrooted topologies as in *a–c*, with approximate branch lengths.

$$E[t_{B-C}|BC] = E[t_{A-C}|AC] = t_2 + 1/3 \quad (4)$$

$$E[t_{B-C}|AB2] = E[t_{A-C}|AB2] = E[t_{B-C}|AC] = E[t_{A-C}|BC] = t_2 + 1/3 + 1 \quad (5)$$

These times can be transformed into genetic distances between tip sequences by assuming an infinite sites mutation model and multiplying by two to account for mutations along both lineages since their common ancestor. Summing the weighted length of branches between any two taxa across all possible topologies leads to the following expected distances:

$$E[d_{B-C}] = 2f_{AB1}(t_2 + 1) + 2f_{AB2}(t_2 + 1/3 + 1) + 2f_{BC}(t_2 + 1/3) + 2f_{AC}(t_2 + 1/3 + 1) \quad (6)$$

$$E[d_{A-C}] = 2f_{AB1}(t_2 + 1) + 2f_{AB2}(t_2 + 1/3 + 1) + 2f_{BC}(t_2 + 1/3 + 1) + 2f_{AC}(t_2 + 1/3) \quad (7)$$

(leaving off the shared mutation parameter, $\mu$, for clarity). Because of the underlying symmetries in topology frequencies and branch lengths under ILS, the expected values of $d_{B-C}$ and $d_{A-C}$ are exactly the same. Notably, these expectations hold for distances calculated without rooted gene trees or polarized substitutions (e.g., fig. 1*d–f*).

Given these results, a natural test of the ILS-only model can be formed using the statistic:

$$D_3 = \frac{d_{B-C} - d_{A-C}}{d_{B-C} + d_{A-C}} \quad (8)$$

Because the two terms in the numerator have the same expected values under ILS alone, the expectation of $D_3$ is 0. The denominator is a normalizing factor that bounds $D_3$ between −1 and +1.

$D_3$ can be significantly different from zero in the presence of gene flow. While the exchange of alleles between lineages *A* and *B* will have no effect on $D_3$, unequal amounts of introgression between either *B* and *C* (fig. 2*a*) or *A* and *C* (fig. 2*b*) can lead to deviations from zero. This occurs because gene flow between a pair of nonsister lineages leads to a breakdown in the symmetry of branch lengths predicted under ILS alone. In particular, introgression between *B* and *C* leads to both more trees with a *BC* topology and a shorter pairwise distance between these two lineages (fig. 2*a*). As a result, $d_{B-C}$ will be smaller than $d_{A-C}$, leading to a negative value of $D_3$. Conversely, gene flow between *A* and *C* leads to positive values of $D_3$. Exact expectations for $D_3$ in the presence of introgression are presented in the Appendix.

## Results and Discussion

### Application of $D_3$

The $D_3$ test is straightforward to carry out, requiring only pairwise distances between three species. Distances can be measured as the percent of sites that differ in an alignment, or

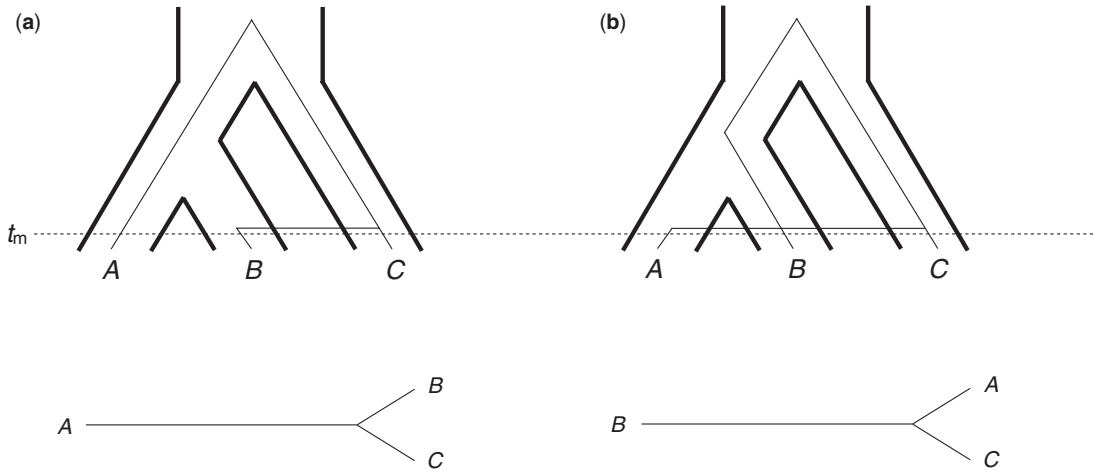


FIG. 2. Topologies produced by introgression. The top row shows the same species tree as in figure 1, but with introgression between (*a*) lineages *B* and *C*, or (*b*) lineages *A* and *C*. Introgression occurs at time $t_m$ in both scenarios. The bottom row again shows the approximate unrooted topologies resulting from introgression. Note how the distance between lineages (*a*) *B* and *C*, or (*b*) *A* and *C* are smaller than in the ILS-only case (fig. 1).

any measure of genetic distance corrected for multiple hits. Ideally, distances should be calculated from regions for which all three lineages have sequences present in the alignment. This will avoid biases that could possibly occur if regions with different ancestral effective population sizes (for example, in regions with different recombination rates; Pease and Hahn 2013) are sampled unequally for the two relevant distances. Otherwise, variation in either $N$ or $\mu$ across sites should not affect the expectation of $D_3$ (see below).

As an example application of this method, we calculated $D_3$ for whole-genome data from three subspecies within *Musa acuminata* (wild bananas; the alignment can be found at https://doi.org/10.6084/m9.figshare.7924727.v1; last accessed August 8, 2019). The tree relating the three subspecies used here is (*M. a. burmannica* (*M. a. malaccensis*, *M. a. banksii*)) (Rouard et al. 2018). As was found using the original *D*-test on these three taxa and an outgroup (Rouard et al. 2018), $D_3$ indicated gene flow between *malaccensis* and *burmannica* ($D_3 = -0.06$; $P < 0.0001$), or species closely related to them (see "Assumptions of $D_3$" below). Distances were calculated as the proportion of sites that differed between sequences and the significance of $D_3$ was determined by a block bootstrap of the *Musa* alignment, as is normally done for the *D*-test (Green et al. 2010).

## Statistical Power of $D_3$ and Comparison with *D*

We tested the power of $D_3$ to detect gene flow with increasing levels of introgression (fig. 3*a*). As the fraction of the genome introgressed approaches 10%, $D_3$ can detect gene flow in 94% of simulated data sets (at $P < 0.05$) with an alignment length of 1 Mb. If we reduce the alignment length to 100 kb we slightly reduce the power to detect gene flow (supplementary fig. 1, Supplementary Material online), while if we increase $\theta$ by 200-fold the power of $D_3$ is increased (supplementary fig. 2, Supplementary Material online). Simulations with variation in $\theta$ across sites reduced power a small amount (supplementary fig. 3, Supplementary Material online). The timing of $t_1$ has no effect on $D_3$ (supplementary fig. 4, Supplementary Material online), while, as expected, the timing of introgression ($t_m$) does have an effect (supplementary fig. 5, Supplementary Material online): when introgression occurs only a short time after speciation, there is little power to detect an asymmetry in branch lengths due to introgression. We also explored an alternative normalization to $D_3$, similar to the *y*-distance introduced in

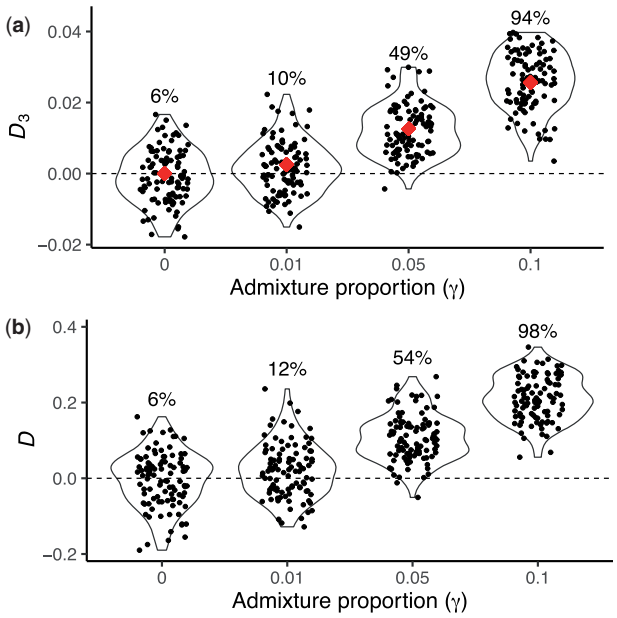


FIG. 3. Statistical power of $D_3$ and *D*. Data were simulated for different proportions of the genome affected by introgression (equivalent to the admixture proportion, $\gamma$), and significance of each data set was determined by block bootstrap (see Materials and Methods). Each black point represents the value of either (*a*) $D_3$ or (*b*) *D* for each simulated data set, with some horizontal jitter added for clarity. Violin plots are used to display the density of values, and in the top panel the red diamonds represent the expected values of $D_3$ for different values of $\gamma$. Percentages reported above each violin plot represent the proportion of simulated data sets that were significantly different from 0 at $P < 0.05$.

Ashander et al. (2018), but found little difference in power (supplementary fig. 6, Supplementary Material online).

In addition to good power to reject the null in the presence of introgression, when there is no gene flow ($\gamma = 0$), the proportion of false positives in $D_3$ is the number we would expect at this significance threshold (fig. 3*a*). We can also see that the expected values of $D_3$ under different levels of introgression (calculated according to the equations given in the Appendix) closely match the mean of simulated data sets (fig. 3*a*). This indicates that we have developed an accurate model for the effect of gene flow on $D_3$.

In order to directly compare these power calculations to the traditional *D*-test, we included an outgroup in the same simulated data sets (the outgroup was simply ignored for $D_3$ calculations). As shown in figure 3*b*, *D* has only slightly more statistical power, despite requiring more data than $D_3$. Our results match similar calculations for *D* carried out previously (e.g., Good et al. 2015; Martin et al. 2015), demonstrating the general power of this class of tests to detect introgression between nonsister lineages.

$D_3$ also has some obvious advantages over similar tests, as it does not require an outgroup (as does the *D*-test) or population samples. Methods such as the $f_3$-test (Reich et al. 2009) require polymorphism data from three taxa, though there are other statistics that only require polymorphism data from two when detecting gene flow between sister species (e.g., Joly et al. 2009; Geneva et al. 2015; Rosenzweig et al. 2016). Even when data from outgroups are available, if there is either ILS or introgression involving these species the *D*-test may not be appropriate. $D_3$ can also detect introgression in both directions (i.e., from *B* into *C* and from *C* into *B*), similar to *D* but unlike $f_3$, which can only detect it in one direction (Peter 2016). It should be noted, however, that the *D*-test will be more robust to sequencing error than $D_3$ because it does not consider mutations on terminal branches (see next section).

The *D*-test has been used with ancient DNA samples, as in the use of Neandertal sequences in the paper introducing this statistic (Green et al. 2010). Although the expectations of branch lengths for $D_3$ given here obviously assume that all sequences are sampled from the present (or are sampled contemporaneously from the past), all of the symmetry expectations hold if the ancient sample is the unpaired lineage (i.e., species *C* in fig. 1). Therefore, there may also be limited cases in which $D_3$ can be applied to ancient samples.

### Assumptions of $D_3$

Several points about the test introduced here merit further discussion and explanation. Although the expectations underlying $D_3$ require few assumptions, there are a few things to be cautious about.

First, we have assumed that the pairwise distances used as input to $D_3$ accurately reflect coalescence times. This will only strictly be true for sequences evolving under an infinite sites model with the same shared mutation rate across lineages. Such conditions likely hold only for relatively closely related species, limiting the use of $D_3$ to recent divergences. To test the robustness of this assumption, we conducted further simulations in which lineage *B* was made to have a faster rate of substitution after the split with *A*. While $D_3$ appears robust to small changes in substitution rates, above a ~0.01% difference in rates there is an extremely high rate of false positives (supplementary fig. 7, Supplementary Material online). These simulations are also analogous to cases in which there are unequal rates of sequencing errors among lineages, or other causes of asymmetry such as mapping bias to a reference genome. To detect such situations, we recommend quantifying the proportion of genomic windows with either a positive or negative value of $D_3$. When there is no introgression the expectation is that 50% of windows will have positive (or negative) values, and under introgression there is only a slight excess of such windows (supplementary fig. 8, Supplementary Material online). In contrast, lineage-specific differences in rates of mutation and/or sequence quality will cause most genomic windows to have either a positive or negative value of $D_3$. Using these expectations should help to distinguish the causes of significant results.

Second, while values of $D_3$ significantly different from zero can be interpreted as rejecting an ILS-only model (given the above assumptions), such results do not strictly mean that introgression is the cause of rejection, or that introgression occurred between the sampled lineages. As with the *D*-test, population structure in the ancestor of all three lineages can produce deviations from the ILS-only expectations (Slatkin and Pollack 2008; Durand et al. 2011). In these cases additional analyses may be needed to distinguish among alternative causes of significant $D_3$ values (e.g., Lohse and Frantz 2014). Likewise, significant results among the three lineages tested do not necessarily mean that gene flow occurred between these species or their direct ancestors. Either unsampled extant lineages or extinct "ghost" taxa may have been the ones directly involved in the introgression event, while the sampled lineages show the effects of reduced divergence. In either of the cases discussed here, one must simply be cautious as to how significant results are interpreted.

Finally, we have assumed that the rooted species tree is known, even though the test does not require an outgroup. Of course it is often the case that the species tree can be inferred from either smaller amounts of sequence data or morphological characters, and so a rooted species tree may be known despite the lack of genome-scale data from an outgroup taxon. However, if the species relationships are not known, a conservative approach would be to test all three combinations of pairwise distances (i.e., $d_{B-C} - d_{A-C}$, $d_{B-C} - d_{A-B}$, and $d_{A-C} - d_{A-B}$). If all three are significantly different from zero, then it is likely that introgression has acted in the system.

## Materials and Methods

In order to determine the statistical power of the tests discussed here, we simulated multi-locus data sets. For each of four different values of the admixture proportion ($\gamma$), our main results are based on 100 simulated data sets consisting of 1,000 nonrecombining loci each using the coalescent simulator *ms* (Hudson 2002). Except where stated explicitly, the

species tree used had $t_1 = 0.3$ and $t_2 = 0.6$, and simulations with introgression had $t_m = 0.05$ (in units of $4N$ generations). All simulations also included an outgroup taxon that diverged at $t_o = 4$, though data from the outgroup was only used for calculations involving $D$. Gene trees from *ms* were passed to Seq-Gen (Rambaut and Grassly 1997) to simulate 1-kb alignments under the Jukes–Cantor model with $\theta = 0.01$. All simulation commands are provided at https://github.com/mhibbins/D3_introgression.

The resulting data sets of 1,000 loci were concatenated together to calculate either $D$ or $D_3$. Calculations of $D_3$ used the proportion of sites that differed between simulated alignments as the genetic distance. Significance of each simulated data set was determined by block bootrapping 1,000 times (with block size equal to 10-kb). The resulting values of $D$ or $D_3$ were used to generate a $z$ distribution, and a nominal value of $P < 0.05$ was used as a threshold for significance. All simulated data sets are provided at https://figshare.com/projects/D3_introgression_test/64862 (last accessed August 8, 2019).

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

The authors are grateful for Mathieu Rouard's assistance with the *Musa* data, and for discussions with Jeff Good and Eric Stone that helped to improve the manuscript. Yaniv Brandvain, Peter Ralph, an anonymous reviewer, and the associate editor all also made helpful suggestions. This work was supported by the Precision Health Initiative of Indiana University and National Science Foundation grant DBI-1564611.

## References

Ashander J, Ralph P, McCartney-Melstad E, Shaffer HB. 2018. Demographic inference in a spatially-explicit ecological model from genomic data: a proof of concept for the Mojave Desert Tortoise. *bioRxiv*:354530.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. 2014. Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genet*. 10(6):e1004410.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol*. 28(8):2239–2252.

Edelman NB, Frandsen P, Miyagi M, et al. 2018. Genomic architecture and introgression shape a butterfly radiation. *bioRxiv*:466292.

Elworth RAL, Allen C, Benedict T, Dulworth P, Nakhleh L. 2018. $D_{GEN}$: a test statistic for detection of general introgression scenarios. *bioRxiv*:348649.

Geneva AJ, Muirhead CA, Kingan SB, Garrigan D. 2015. A new method to scan genomes for introgression in a secondary contact model. *PLoS One* 10(4):e0118621.

Good JM, Vanderpool D, Keeble S, Bi K. 2015. Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. *Evolution* 69(8):1961–1972.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328(5979):710–722.

Hahn MW. 2018. Molecular population genetics. Sunderland (MA): Sinauer Associates/Oxford University Press.

Hibbins MS, Hahn MW. 2019. The timing and direction of introgression under the multispecies network coalescent. *Genetics* 211(3):1059–1073.

Hudson RR. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* 37(1):203–217.

Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337–338.

Huson DH, Klopper T, Lockhart PJ, Steel MA. 2005. Reconstruction of reticulate networks from gene trees. *Res Comput Mol Biol Proc*. 3500:233–249.

Joly S, McLenachan PA, Lockhart PJ. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am Nat*. 174(2):E54–E70.

Lohse K, Frantz LA. 2014. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics* 196(4):1241–1251.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *Bioessays* 38(2):140–149.

Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol Biol Evol*. 32(1):244–257.

Meng C, Kubatko LS. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theor Popul Biol*. 75(1):35–45.

Pease JB, Hahn MW. 2013. More accurate phylogenies inferred from low-recombination regions in the presence of incomplete lineage sorting. *Evolution* 67(8):2376–2384.

Pease JB, Hahn MW. 2015. Detection and polarization of introgression in a five-taxon phylogeny. *Syst Biol*. 64(4):651–662.

Peter BM. 2016. Admixture, population structure, and *F*-statistics. *Genetics* 202(4):1485–1501.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci*. 13:235–238.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. *Nature* 461(7263):489–494.

Rosenzweig BK, Pease JB, Besansky NJ, Hahn MW. 2016. Powerful methods for detecting introgressed regions from population genomic data. *Mol Ecol*. 25(11):2387–2397.

Rouard M, Droc G, Martin G, Sardos J, Hueber Y, Guignon V, Cenci A, Geigle B, Hibbins MS, Yahiaoui N, et al. 2018. Three new genome assemblies support a rapid radiation in *Musa acuminata* (wild banana). *Genome Biol Evol*. 10:3129–3140.

Slatkin M, Pollack JL. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. *Mol Biol Evol*. 25(10):2241–2246.

Yu Y, Than C, Degnan JH, Nakhleh L. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst Biol*. 60(2):138–149.