Received: 2 April 2025 | Accepted: 29 September 2025

DOI: 10.1002/aps3.70045

RESOURCE ARTICLE

Applications in Plant Sciences

Detecting cryptic ghost lineage introgression in four-taxon genomic datasets

Evan S. Forsythe[1,2] | Blaine S. Pappa[1] | Darren A. Clavette[1] | Devin Y. Mendoza[1]

[1]Biology Program, Oregon State University–Cascades, Bend, Oregon, USA

[2]Department of Integrative Biology, Oregon State University, Corvallis, Oregon, USA

Correspondence

Evan S. Forsythe, Biology Program, Oregon State University–Cascades, Bend, Oregon, USA.
Email: evan.forsythe@osucascades.edu

This article is part of the special section "Branching out: Resolving plant evolution through phylogenetic networks."

Funding information
National Science Foundation, Grant/Award Number: IOS-2114641

Abstract

Premise: Hybridization and introgression are pervasive evolutionary forces that have played fundamental roles in shaping the diversity of wild and domesticated plants. Four-taxon tests for introgression provide a reliable framework for detecting signatures of ancient introgression from genomic data, which have played an important role in revealing the reticulate nature of plant evolution; however, there is emerging evidence that a cryptic process known as ghost lineage introgression has the potential to dramatically skew interpretations of four-taxon introgression statistics, particularly our ability to determine the lineages involved in introgression. This ambiguity limits our ability to resolve the mechanisms and functional implications of introgression because it means we can determine neither the donor nor the recipient of introgressed alleles with confidence.

Methods: Here, we develop ghostbuster, a statistical test designed to detect ghost lineage introgression in genomic data based on patterns of sequence divergence. We employ coalescent simulations to test our method and ascertain the conditions under which it accurately identifies ingroup versus ghost lineage introgression. Finally, to demonstrate the utility of ghostbuster, we apply it to a previously identified introgression event in the plant family Brassicaceae.

Results: Our simulations reveal that ghostbuster accurately distinguishes ghost lineage introgression from ingroup introgression across a range of introgression scenarios, with errors arising only when divergence events are closely spaced or when ancestral population sizes are unbalanced. Our analysis of empirical plant data reveals that the previously identified introgression likely constitutes ghost lineage introgression and, thus, was previously misinterpreted.

Discussion: Our analyses of simulated and empirical data demonstrate that ghostbuster will be a helpful tool in resolving reticulate evolution in plants and other taxa. We demonstrate the biological insights that ghostbuster provides by presenting an updated model of ghost lineage introgression in Brassicaceae, impacting our understanding of the molecular evolution of crop and model species in this important plant lineage. Ghostbuster code is freely available at: https://github.com/EvanForsythe/Ghost_introgression.

KEYWORDS

ghost lineage, hybridization, introgression, phylogenomics

Hybridization is widespread across eukaryotes (Taylor and Larson, 2019) and is especially prominent in plants (Mallet, 2005; Yakimowski and Rieseberg, 2014; Mallet et al., 2016). Hybridization, followed by backcrossing, leads to introgression of alleles between divergent species or populations, a phenomenon that has been well-documented in plants (Rieseberg and Soltis, 1991; Rieseberg, 2006; Suarez-Gonzalez et al., 2016; Forsythe et al., 2020a) and animals

(Dasmahapatra et al., 2012), including humans (Green et al., 2010; Prüfer et al., 2014; Kuhlwilm et al., 2016). Studies of hybridization and introgression have been particularity insightful in disentangling unresolved phylogenetic relationships in plants (Beilstein et al., 2006, 2008; Oyama et al., 2008; Huang et al., 2015; Forsythe et al., 2020a) and uncovering the signatures of adaptive introgression in plant genomes (Hufford et al., 2013; Wang et al., 2023). Advances in genome sequencing technology have led to the development of statistical approaches to detecting patterns of historical introgression in extant species (Green et al., 2010; Durand et al., 2011; Martin et al., 2015; Pease and Hahn, 2015; Hahn and Hibbins, 2019; Hibbins and Hahn, 2022). Many of the most widely used statistics employ a four-taxon sampling approach to detect an abundance of derived alleles shared by non-sister species by detecting imbalanced site pattern or gene tree topology counts (Green et al., 2010; Durand et al., 2011; Martin et al., 2015; Martin and Jiggins, 2017; Zheng and Janke, 2018; Dagilis et al., 2021). Four-taxon statistics have been used to identify foundational cases of adaptive introgression (Dasmahapatra et al., 2012), extensive introgression impacting large portions of the genome (Fontaine et al., 2015; Forsythe et al., 2020a), and introgression events with profound significance to human health and medicine (Prüfer et al., 2014). Given the impact of introgression on the molecular evolution of important taxa, it is paramount to fully resolve a detailed understanding of these introgression events.

One of the most fundamental details regarding an introgression event is to resolve the lineages that exchanged alleles during historical gene flow (Hibbins and Hahn, 2019, 2022; Forsythe et al., 2020b). Four-taxon tests are designed to reveal introgression events between two alternative pairs of non-sister taxa within the ingroup clade of the test species (Durand et al., 2011; Tricou et al., 2022b). These two alternative introgression scenarios are distinguished by the sign (positive versus negative) of the test statistic. Once the two species involved in an introgression event have been identified with a four-taxon test, auxiliary tests (Hibbins and Hahn, 2019, 2022; Forsythe et al., 2020b), which make use of sequence divergence information, can be applied to resolve the directionality of introgression (i.e., the donor and recipient lineage). Furthermore, modified four-taxon tests have also been developed to identify the specific genomic regions (Martin et al., 2015; Martin and Jiggins, 2017) that were introgressed between species, revealing the genomic and functional impacts of introgression.

Combining four-taxon site-pattern tests and divergence-based tests holds potential to resolve introgression events within sampled ingroup species; however, there is a growing body of evidence that suggests that interpretation of these tests is dramatically impacted by cryptic introgression stemming from unsampled or extinct lineages that diverged prior to the divergence of the sampled ingroup species (Durand et al., 2011; Hibbins and Hahn, 2022; Tricou et al., 2022a, 2022b; Tiley et al., 2023; Pang and Zhang, 2024). Introgression from these so-called "ghost lineages" can skew allele patterns and gene tree topologies in a manner that exactly mimics the signature of ingroup introgression produced by four-taxon tests (Tricou et al., 2022b). This caveat is especially problematic because the implication of a misidentified ghost lineage introgression event means that neither the donor lineage nor the recipient lineage is correctly inferred (Tricou et al., 2022a; Pang and Zhang, 2024). Considering that the vast majority of species that have existed on Earth are either extinct or undescribed (Mora et al., 2011) and that fossil and archaic DNA evidence has demonstrated gene flow between now-extinct species (Slon et al., 2018), it is likely that at least some of the previously identified introgression events detected with four-taxon tests have been misinterpreted.

Given the potential for misinterpretation of introgression results, tools to help differentiate between ingroup and ghost lineage introgression are of value. Recent work has thoroughly investigated the utility of existing heuristic methods (i.e., methods that use topology counts and genome-wide summary statistics) in accurately identifying ingroup versus ghost lineage introgression, revealing that heuristic methods perform poorly in their ability to correctly identify ghost lineage introgression (Pang and Zhang, 2024). In contrast, this study also demonstrated that a model-based approach, Bayesian Phylogenetics and Phylogeography (BPP) (Yang, 2015; Flouri et al., 2020; Jiao et al., 2021), was able to accurately detect ghost lineage introgression. This study highlighted the utility of full likelihood demographic methods, which make use of both topology and sequence divergence information in sequence data (Hibbins and Hahn, 2022), in accurately identifying modes of introgression. Based on these promising results, BPP is likely to become a valuable addition to introgression analysis workflows. However, the authors also note that such methods are computationally expensive compared to analogous heuristic-based methods, presenting a challenge in analyzing large genomic datasets (Pang and Zhang, 2024). Given this limitation, a heuristic method that explicitly tests for ingroup versus ghost lineage introgression would be a valuable addition to analytical workflows.

Here, we present ghostbuster, a novel statistical test to distinguish ingroup from ghost lineage introgression in four-taxon genomic data. Ghostbuster provides an efficient heuristic approach to using sequence divergence summary statistics to test introgression hypotheses. We define the expected divergence profiles under ingroup versus ghost lineage introgression and perform multispecies coalescent simulations to test the performance and accuracy of ghostbuster under different introgression scenarios. Finally, we apply ghostbuster to a previously described introgression event in the plant family Brassicaceae and present a revised model for introgression in the group.

METHODS

Defining sequence divergence expectations under ingroup and ghost lineage introgression

Widely used four-taxon tests for introgression rely on the topology of gene trees or corresponding site patterns to

indicate the presence of introgression. However, ingroup and ghost lineage introgression can lead to identical topologies and site patterns (Tricou et al., 2022a, 2022b), meaning these pieces of evidence cannot differentiate between alternative introgression scenarios. We reasoned that the divergence profiles of introgressed and non-introgressed loci could help resolve the species involved in introgression events. To outline our expectations for introgressed gene trees, we defined time points corresponding to introgression events (T_{IG}) and speciation events (T_α, T_β, T_G, and T_γ) in a four-taxon (not including a ghost lineage) introgression scenario (Figure 1A). Phylogenetic estimates of the relative timing of divergence events are obtainable by calculating sequence divergence (i.e., node depth) between pairs of taxa on gene trees (Figure 1). Gene trees that were not introgressed are expected to display the "species topology" and will inherently contain a node representing the divergence of P1 and P2; the node depth of this node (K_{12}) should always correspond to T_α (Figure 1A). On the other hand, gene trees that underwent introgression will display the "introgressed topology" and, as such, will inherently contain a node representing the divergence of P2 and P3 with node depth K_{23} (Figure 1B, C). Importantly, unlike K_{12}, K_{23} will differ depending on whether introgression is ingroup ($K_{23} = T_{IG}$) (Figure 1B) versus ghost lineage ($K_{23} = T_\beta$) introgression (Figure 1C).

To detect the sequence divergence profile of introgression events, we propose a simple heuristic, the ΔK statistic:

$$\Delta K = \bar{K}_{12} - \bar{K}_{23} \quad (1)$$

Under ingroup introgression, we expect ΔK to be positive, such that:

$$\Delta K = T_\alpha - T_{IG} \quad (2)$$

Under ghost lineage introgression, we expect ΔK to be negative, such that:

$$\Delta K = T_\alpha - T_\beta \quad (3)$$

Our ΔK statistic provides a measurable readout of the divergence profile of introgression events that are detected with four-taxon statistics.

Implementation of the ghostbuster test

We developed ghostbuster as a Python tool that requires a minimal set of dependencies and computational power. Ghostbuster is designed to analyze the same data used for four-taxon introgression tests (Green et al., 2010; Durand et al., 2011) and is intended for use only after a significant signature of introgression has already been detected. Ghostbuster requires FASTA input files with DNA sequences for single-copy genes/loci/windows spanning the genome. Each file must contain a sequence from at least four species (P1, P2, P3, and an outgroup) (Figure 1) used in upstream four-taxon tests. Sequences from extra species are pruned from files prior to multisequence alignment.

For each input file, the following steps are performed: multiple sequence alignment with MAFFT (Katoh and Standley, 2013), maximum likelihood inference with IQ-TREE (Nguyen et al., 2015), which employs ModelFinder (Kalyaanamoorthy et al., 2017), and finally, topology and branch length analyses using Biopython (Chapman and Chang, 2000) tools (Figure 2). Each tree is rooted by the user-specified outgroup, and branch lengths (measured in substitution per site) are calculated from each gene tree after topology is assessed. A divergence value for each node is obtained by summing the branch lengths of the two external branches subtending the node.

Ghostbuster assumes that the user-specified species tree topology is (((P1,P2)P3)O) and calculates node depth K_{12} from gene trees exhibiting this topology. Gene trees with the topology (((P2,P3)P1)O) are used to calculate K_{23}. Gene trees with the topology (((P1,P3)P2)O) are presumed

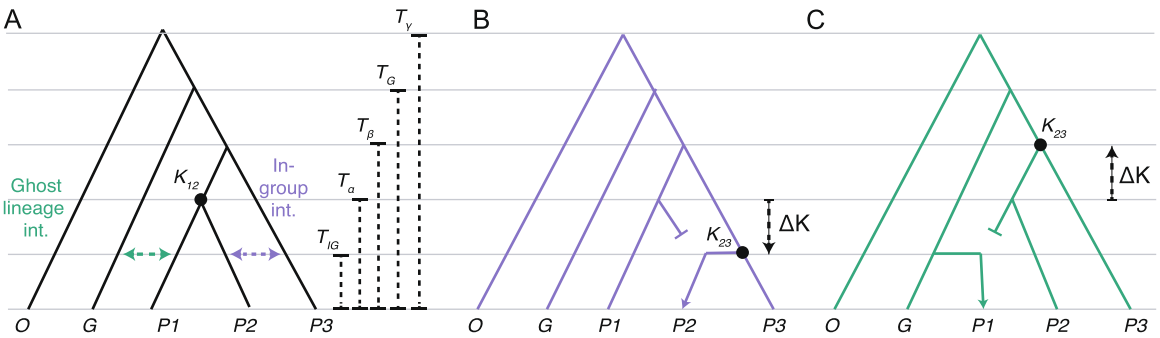

FIGURE 1 Expected sequence divergence under ingroup versus ghost lineage introgression models. (A) Species tree (black branches) with possible introgression events (green and purple arrows) and tree depth times for speciation and introgression events (dotted lines). P1, P2, P3, G, and O refer to ingroup, ghost, and outgroup populations, respectively. The K_{12} node indicates the node of interest on gene trees that exhibit the species topology ((P1,P2),P3). (B, C) Gene trees resulting from ingroup (B) and ghost lineage (C) introgression. The K_{23} node indicates the node of interest on gene trees that exhibit the introgression topology ((P2,P3),P1), displayed in both introgression scenarios. ΔK symbols and arrows indicate the expected difference between K_{12} and K_{23} when comparing introgressed genes to non-introgressed genes. T variables refer to timepoints at which introgression (T_{IG}) and speciation events (T_α, T_β, T_G, and T_γ) occur.

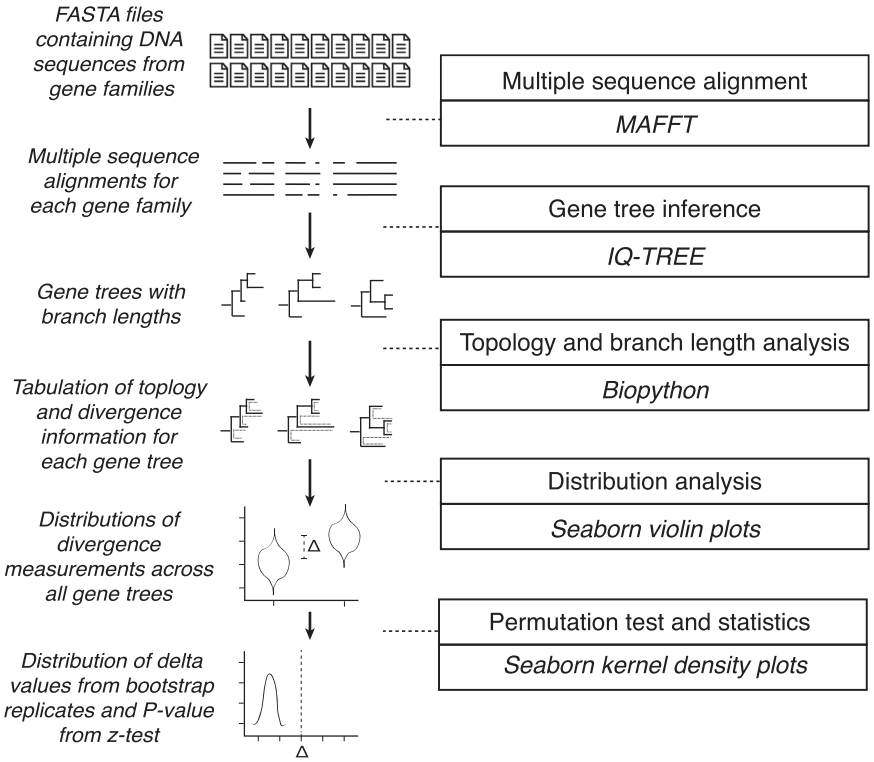

FIGURE 2 Analytical workflow showing the major steps of the ghostbuster test, implemented as a Python tool (available at https://github.com/EvanForsythe/Ghost_introgression).

to originate from incomplete lineage sorting (ILS) (see Discussion) and are not used by ghostbuster. ΔK is calculated by subtracting the average from all K_{23} values from the average from all K_{12} values (see Equation 1). To assess the significance of the ΔK, ghostbuster performs a bootstrap replication permutation test with 100 replicates. For each bootstrap replicate, gene trees are randomly sampled (with replacement) and K_{12} and K_{23} are estimated for the sample to produce a resampling distribution of ΔK. Similar to the approach used by Forsythe et al. (2020b), a two-sided z-test is performed to assess whether the distribution significantly differs from zero.

DNA sequence simulations of ingroup and ghost lineage introgression scenarios

To test ghostbuster performance, we developed simulations using tskit (Kelleher et al., 2016; Wong et al., 2024) and msprime (Baumdicker et al., 2022) multispecies coalescent simulation software packages. We implemented simulations using the script *Data_simulations.py*, available with ghostbuster (https://github.com/EvanForsythe/Ghost_introgression). Speciation events were simulated via the *add_mass_migration()* command with the proportion argument set to 1.0. In a reverse-time (coalescent) perspective, these events are akin to two populations merging; however, when considered from a forward-time direction, they are akin to a speciation event. Introgression events were simulated with the *add_mass_migration()* command with the proportion argument set to 0.2, simulating (in forward-time) an introgression event in which 20% of alleles are introgressed. Introgression was simulated as either "ghost" or "true" by setting the "destination" (i.e., the source in forward-time) of migration. In all cases, we simulate the presence of a ghost lineage, but we never include the resulting sequence from this ghost taxon in downstream analyses, thus effectively mimicking an "unknown" ghost lineage.

Our default time points (in generations) were set to the following: $T_{IG} = 40{,}000$, $T_{\alpha} = 80{,}000$, $T_{\beta} = 120{,}000$, $T_{G} = 160{,}000$, and $T_{\gamma} = 200{,}000$ (Figure 3). For parameter scan analyses, we multiplied all time points or the T_{IG} time point by a branch length scaling factor (SF) (Figure 4). Our default effective population size used in simulations was $N_e = 10{,}000$ (diploid individuals). N_e parameter scans multiplied N_e in P3 by a scaling factor to achieve larger or smaller populations. Additional parameters used during sequence simulations were: sequence length = 10,000,000 bp, mutation rate = 0.0000001 (mutations per base pair per generation), and recombination rate = 0.000000001 (recombination events per base pair per generation). These parameters were held constant for the simulations in this study.

The output of the above simulations is a "full chromosome" alignment comprising numerous introgressed and non-introgressed haplotype blocks. From this full chromosome alignment, we achieved single-gene FASTA files for input into ghostbuster by splitting the full chromosome alignment into 1000 non-overlapping windows, which were

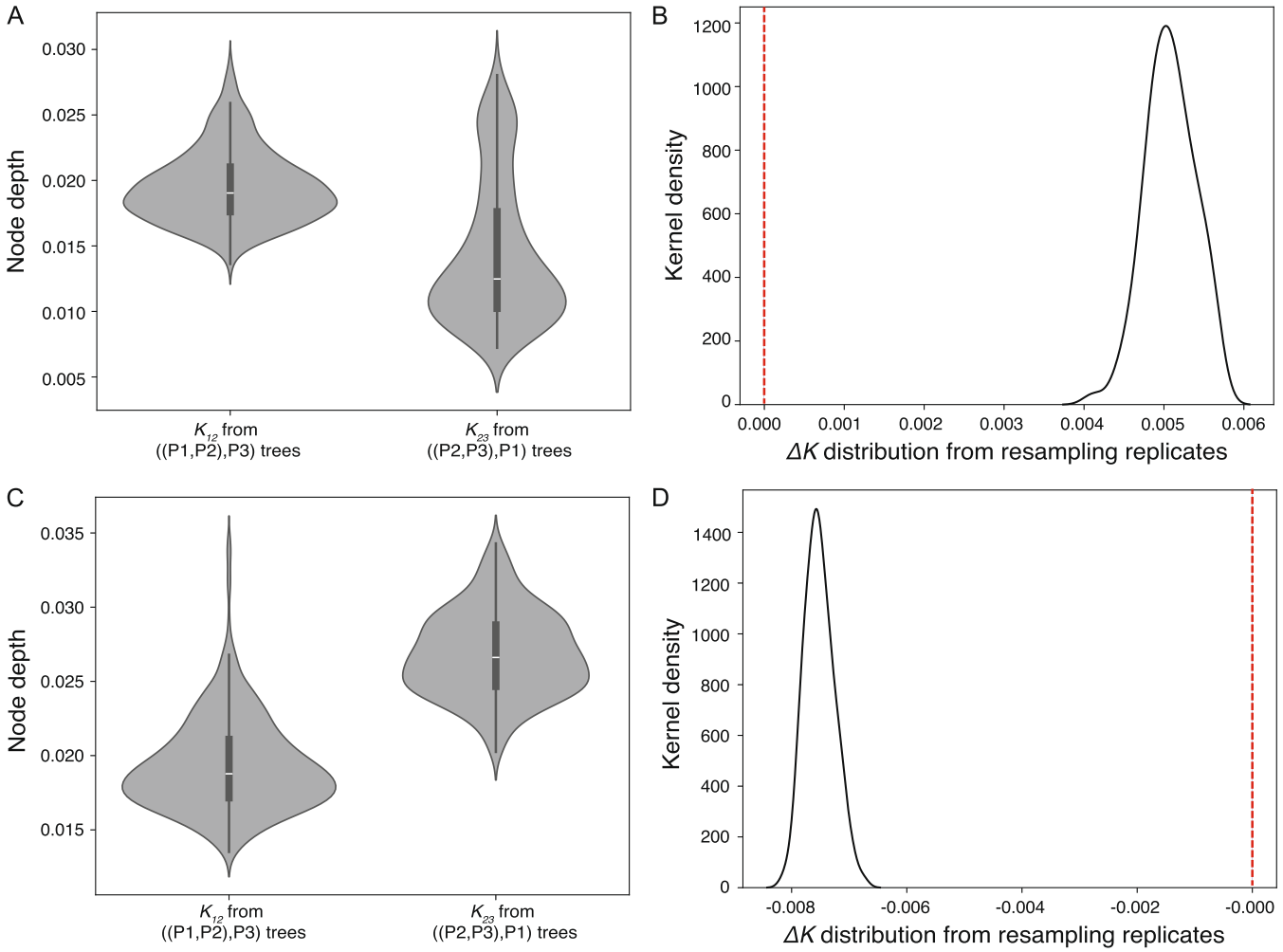

FIGURE 3 Applying the ghostbuster test to simulated data. Ghostbuster output on a simulated "ingroup" introgression dataset, in which introgression was simulated from P3 to P2 (A, B), or a simulated "ghost lineage" introgression, in which introgression was simulated from a ghost lineage into P1 (C, D). (A, C) The distribution of node depths obtained from gene trees with the speciation topology or the introgression topology. (B, D) ΔK distribution from 100 bootstrap resampling replicates. Both distributions significantly ($P \ll 0.01$) differ from zero, indicating significant signatures of ingroup (B) and ghost lineage (D) introgression.

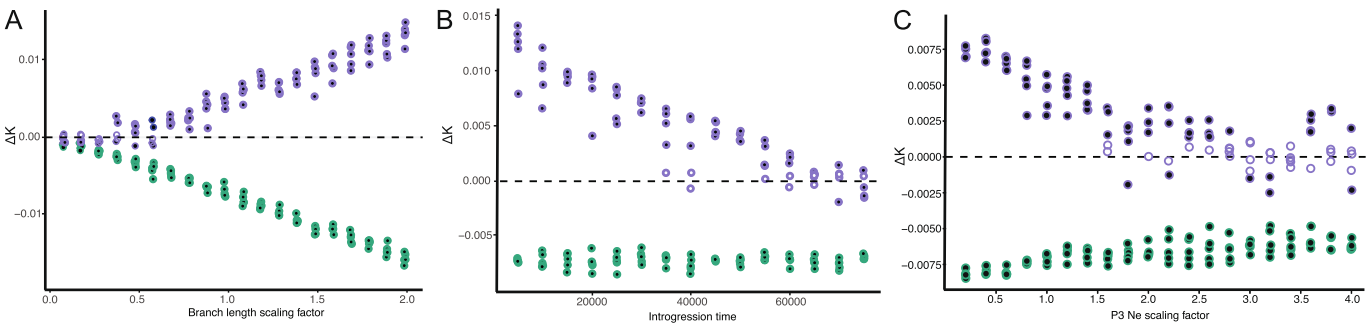

FIGURE 4 Testing the robustness of the ghostbuster method under simulated introgression scenarios. (A–C) ΔK measurements from ghostbuster runs on datasets with simulated ingroup (purple points) and ghost lineage (green points) introgression. Positive and negative ΔK measurements from the ghostbuster test indicate ingroup and ghost lineage introgression, respectively. (A) Parameter scan of varying branch length scaling factor values used to scale the simulated timing of all divergence and reticulation events proportionally. Larger numbers equate to larger tree heights on simulated trees. (B) Parameter scan of introgression scenarios in which introgression time is changed while all other time points are held constant. Larger values indicate introgression (T_{IG}) occurring at deeper time points approaching the timing of the speciation event (T_{α}). (C) Parameter scan of the effective population size (N_e) in the P3 lineage (the donor population in ingroup introgression). P3 N_e was multiplied by P3 N_e scaling factors to achieve unbalanced simulated donor and recipient population sizes. Filled points indicate a significant (z-test, $P < 0.05$) ghostbuster test result. Open points indicate that ΔK distributions did not significantly differ from zero.

used to infer window trees (hereafter referred to as "gene trees") in downstream ghostbuster analyses.

Testing for ghost lineage introgression in empirical plant genomic data

To test ghostbuster on an empirical dataset, we used genomic data from a previous study that identified introgression between model species in the Brassicaceae family (Forsythe et al., 2020a). We obtained coding sequence (CDS) alignments for single-copy genes used in the prior study and applied the ghostbuster test to a representative species from each clade using the species tree topology shown in Figure 5. Analyses were performed on a Linux server equipped with dual AMD EPYC 7713 processors (Advanced Micro Devices, Santa Clara, California, USA) using four parallel threads. Analyses were repeated on a Macintosh laptop (Apple M1 Pro chip; Apple, Cupertino, California, USA), yielding similar runtimes and results.

RESULTS

Testing ghostbuster performance on simulated genomes

To test our ghostbuster approach, we simulated DNA sequence evolution for populations described in Figure 1A. Ingroup introgression was achieved by simulated gene flow from P3 to P2. Ghost lineage introgression was achieved by simulating gene flow from the ghost lineage (G) into P1. This approach mimics ghost lineage introgression because all downstream analyses omit the ghost population, only making use of O, P1, P2, and P3 in divergence analyses (see Methods). To demonstrate the output of ghostbuster, we applied the test to genomes simulated to have undergone ingroup introgression (Figure 3A, B) and ghost lineage introgression (Figure 3C, D) at our default time points (see Methods). Our results show that simulated ingroup introgression leads to a higher K_{12} distribution compared to the K_{23} distribution (Figure 3A). This divergence profile resulted in a ΔK distribution significantly ($P < 0.001$) greater than zero (Figure 3B), consistent with our theoretical expectations for ingroup introgression. When we simulated ghost lineage introgression, we observed the opposite pattern for K_{12} and K_{23} distributions (Figure 3C), resulting in a significantly negative ($P < 0.001$) ΔK distribution (Figure 3D), again consistent with our expectations. These results provide a proof-of-principle for the effectiveness of our ghostbuster test.

To evaluate robustness of the ghostbuster test, we performed parameter scans, in which we altered the timing of simulated speciation and introgression events. First, to explore the effect of overall tree height on ghostbuster results, we employed a branch length SF and multiplied all time points by this number to achieve simulated genomes with overall larger (branch length SF > 1) or smaller (branch length SF < 1) levels of divergence (Figure 4A). We found that increased tree height provided increased resolution for the ghostbuster test. The effect size of ΔK grew increasingly positive (ingroup introgression) or negative (ghost lineage introgression) as branch length SF increased. We simulated five replicate genomes for each branch length SF value, and ghostbuster significantly inferred the correct mode of introgression with 100% accuracy for all branch length SF ≥ 0.7 runs. When applied to simulated ghost lineage introgression, ghostbuster inferred the correct mode for all replicates and SF values, with the exception of a single insignificant result for one replicate at branch length SF = 0.2. In contrast, ghostbuster accuracy suffered when applied to simulated ingroup introgression on branch length SF < 0.7 runs, as evidenced by many cases of erroneous ΔK estimates below zero at low branch length SF values. In some cases, these erroneous results were statistically significant, indicating that ghostbuster is prone to error under rapid

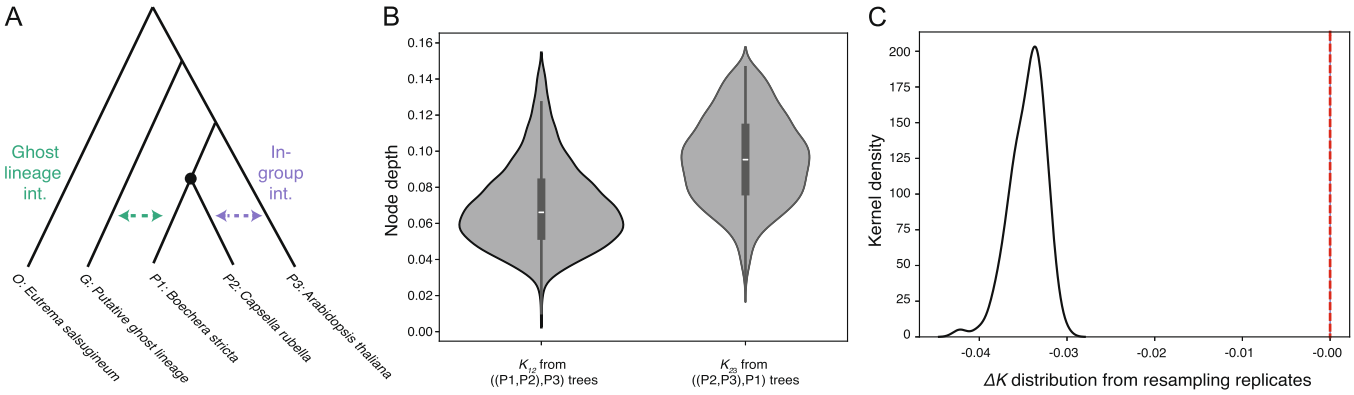

FIGURE 5 Testing for ghost lineage introgression in a previously identified introgression event in Brassicaceae. (A) Species relationships among test species as indicated by the major topology from all gene trees. Green and purple arrows indicate alternative introgression scenarios that would conceivably result in the previously observed ((P2,P3),P1) topology gene trees. P1, P2, P3, G, and O refer to ingroup, ghost, and outgroup populations, respectively. (B) The distribution of node depths obtained from gene trees with the speciation topology (left) or the introgression topology (right). (C) ΔK distribution from 100 bootstrap resampling replicates.

speciation scenarios, when divergence events occur in close succession.

To test the effect of the relative timing of speciation and introgression events, we also performed a parameter scan in which we varied only the timing of introgression (T_{IG}), while leaving all other time points constant. For these simulations, increasing T_{IG} brings it closer to T_{α} (Figure 4B). For ghost lineage introgression simulations, we found that ΔK values are unaffected by T_{IG}. This is consistent with our theoretical expectations because ΔK is affected only by T_{α} and T_{β} in ghost lineage introgression scenarios (Equation 3), which are unchanged in this analysis. In contrast, ingroup ΔK are heavily impacted by T_{IG}, with increasing rates of insignificant and significantly erroneous results at high T_{IG} values. This result again highlights that ghostbuster accuracy suffers when divergence and introgression occur in short order. These divergence time parameter scans show that ghostbuster is a reliable test under most divergence scenarios but lacks resolution under rapid speciation (see Discussion).

Finally, to test whether unbalanced donor and recipient population sizes influence divergence patterns, we performed a parameter scan in which we varied N_e in the P3 population. We applied a P3 N_e SF, which altered the N_e of the P3 population, while retaining the default value (10,000 individuals) for all other simulated populations. We found that increasing the N_e of P3 relative to other populations decreased the ability of the ΔK statistic to correctly infer ingroup introgression (Figure 4C). Increasing the P3 N_e by a factor of 1.6 or greater led to some replicates yielding insignificant or significantly incorrect results for ingroup introgression. Meanwhile, ghostbuster performance on ghost introgression events was unaffected, which is unsurprising given that P3 is neither a recipient nor donor in ghost introgression scenarios. Overall, our parameter scan of N_e demonstrates that ghostbuster is sensitive to differences in population size across lineages (see Discussion).

Applying ghostbuster to resolve introgression in Brassicaceae

With an understanding of ghostbuster performance on simulated data, we applied the ghostbuster test to a previously described introgression event in the plant family Brassicaceae (Forsythe et al., 2020a). We used the most frequent gene tree topology as our species topology (Figure 5A). In the previous study, significant four-taxon test results were interpreted as evidence of ingroup introgression (Figure 5A; purple arrow); however, the authors also discussed the possibility of a cryptic ghost lineage introgression event (Figure 5A; green arrow) but lacked a statistical framework for testing these alternative hypotheses. To resolve this ambiguity, we applied ghostbuster to the same data. We observed a significantly negative ($P < 0.001$) ΔK distribution, indicating that this previously described introgression event is likely a case of ghost lineage introgression (see Discussion). These results highlight the utility of ghostbuster in resolving introgression results, providing a new basis for interpreting four-taxon tests for introgression.

DISCUSSION

A fast heuristic that explicitly tests for ghost lineage introgression

Our ghostbuster analysis of Brassicaceae alignments took 22 minutes to complete using four parallel threads for the phylogenetic inference steps. This runtime using only modest computational resources means that ghostbuster can accomplish real-life analyses using computing resources present in most modern base-model laptops. This efficiency is derived from the heuristic nature of comparing average node depths rather than traversing joint distributions of gene tree topologies and coalescence times. The most computationally intensive step of the ghostbuster workflow is the maximum likelihood gene tree inference step, which we achieve with IQ-TREE. As part of gene tree and branch length inference, we employ IQ-TREE's "-m TEST" options to identify the best model of evolution for phylogenetic inference. Given that there is a model-based phylogenetic inference step, it could be argued that ghostbuster is not strictly a heuristic method. However, the overall efficiency for the test and the incorporation of summary statistics make ghostbuster best aligned with site-pattern and gene tree topology–based methods that have been characterized as heuristic-based methods (Pang and Zhang, 2024). In any case, ghostbuster shows promise as a fast and accurate test that can be routinely applied to the same four-taxon datasets that are widely used to indicate the presence of an introgression event.

Reduced phylogenetic resolution amid rapid speciation and unbalanced population sizes

Our simulation analyses sought to explore coalescence times and population size parameters to understand the limits of ghostbuster performance (Figure 4). Parameter scans of divergence time revealed the ghostbuster test does not perform well under scenarios in which only a small amount of time separates the speciation event (T_{α}) from the introgression event (T_{IG}) (Figure 4A, B). Lack of phylogenetic resolution in these scenarios is not entirely surprising, given that the process of ILS is known to introduce alternative gene tree topologies with skewed coalescence times (Hibbins and Hahn, 2022). Multispecies coalescent simulations inherently model the process of ILS (i.e., deep coalescence), the extent of which is directly related to time between events (Hudson, 2002). Consistent with the idea that ILS introduces noise to ghostbuster, the simulations that produced erroneous or insignificant ghostbuster results

were also the simulations that produced the most gene trees with a topology in which P1 and P3 are sister (data not shown), which likely result from ILS.

Moreover, in addition to introducing a third gene tree topology on four-taxon trees, ILS is also expected to shift coalescence times, resulting in gene trees with deeper node depths. This impact affects gene trees displaying all three possible topologies, meaning that the genome-wide distribution of node depths across gene trees will be a mixed distribution of ILS trees and non-ILS trees. Consistent with this expectation, we observe a bimodal K_{23} distribution in our simulated ingroup introgression example (Figure 3A). Despite ILS adding noise, the introgression signal (lower peak) vastly outweighed the ILS signal (higher peak), leading to a significant and accurate ghostbuster result. While cases in which ILS is more prominent should be interpreted with caution, we do not anticipate it posing a major barrier in the practical application of ghostbuster. This is because the conditions that would render ghostbuster ineffective are also likely to undermine upstream four-taxon introgression tests, making them less likely to produce a significant signature of introgression in the first place. As a result, a ghostbuster analysis would not be warranted under those conditions.

In addition to divergence time, the effective population size used in coalescent simulations is also expected to influence divergence profiles. Because time to coalescence scales with population size (Kingman, 1982), larger populations are expected to exhibit deeper coalescence and elevated divergence times. This phenomenon has the potential to impact ΔK estimates, particularly when the donor population (P3 in our ingroup introgression scenarios) is larger than the recipient population (P2). To explore this effect, we varied P3 N_e and observed a consistent decline in ghostbuster performance when P3 approached three times the size of other populations in our simulated dataset (Figure 4C). Sensitivity to this effect will likely depend on additional parameters, such as T_α, T_{IG}, and mutation rate. Taken together, our results highlight the importance of considering both divergence timing and population size when interpreting ΔK. Additionally, ghostbuster users should consider the possibility of confounding demographic factors such as ancestral population structure (Eriksson and Manica, 2012), continuous migration events, multiple counteracting reticulation events (Than et al., 2008), or gene flow from a deeply diverged lineage (Tricou et al., 2022a). Nevertheless, ghostbuster will serve as a valuable tool for disentangling complex introgression hypotheses, particularly when introgression signals are strong and confounding effects like ILS and unbalanced population sizes are limited or predictable.

An updated model of introgression in Brassicaceae

We previously interpreted four-taxon results in Brassicaceae as an indication of ingroup introgression (Forsythe et al., 2020a). However, our ghostbuster analysis indicates that this introgression event is likely introgression from an unknown ghost lineage (Figure 5), indicating a novel alternative model for introgression in this important group of model and crop plants. We initially hypothesized the presence of this introgression event based on evidence of phylogenetic incongruence between nuclear and chloroplast gene trees (Beilstein et al., 2006, 2008), and subsequent genome-wide phylogenomic analysis revealed incongruence among nuclear genes, with ~88% of nuclear genes displaying a major topology and ~8% of nuclear genes displaying a minor topology (Forsythe et al., 2020a). The most parsimonious interpretation of these highly unbalanced ratios is that the major topology represents the true species branching order and the minor topology represents introgressed genes. However, based on conflicting divergence values in our prior study, we explored an alternative hypothesis in which the minor topology represents the species branching order, while the major topology is the result of extensive introgression, affecting ~88% of nuclear genes. This interpretation was based on a similar pattern and interpretation of introgression in mosquitos (Fontaine et al., 2015). However, a later study reinterpreted the mosquito introgression data considering ghost lineage introgression and found that underlying assumptions about branch lengths do not hold true (Tricou et al., 2022a). Consistent with this, our updated model suggests that the minor topology is indeed the result of ghost lineage introgression. However, it should be noted that the ghostbuster framework developed here requires the user to define the species relationships a priori and does not explicitly test between different species trees, meaning it cannot be used to rule out the possibility of extensive introgression impacting most of the genome. Despite this, the growing understanding of ghost lineage introgression and its impacts on branch lengths favors our updated model as the most plausible scenario.

Our previous study of introgression in Brassicaceae found evidence of cytonuclear selection acting during introgression (Forsythe et al., 2020a). Based on previous interpretations (under the assumption of ingroup introgression), the recipient of this introgression was presumed to be Clade C, which includes the emerging biofuel crop *Camelina sativa* (L.) Crantz. Our updated model of ghost lineage introgression suggests that, instead, the recipient of introgression is Clade B, represented by *Boechera stricta* Al-Shehbaz, a genetic model used to study the genetics of asexual reproduction in plants (Schranz et al., 2006). It should be noted that our new model does not invalidate the prior findings related to cytonuclear selection, but it does change our understanding of the lineage in which this selection acted, which is an important component of resolving an introgression event. This updated interpretation of introgression highlights the biological insights gained from testing for ghost lineage introgression. Moreover, this new model, combined with the relative abundance of genomic resources available in the Brassicaceae family, creates an exciting opportunity to search for close

relatives/descendants of the putative donor ghost lineage among extant Brassicaceae species.

Finally, given that we observed inconclusive and erroneous results in some scenarios in our simulated data (Figure 4), it is important to ask whether Brassicaceae introgression falls in the range of divergence times and population size for which ghostbuster performance is robust. Our default simulated T_{β}, T_{α}, and T_{IG} are 120, 80, and 40 thousand generations, respectively, which would correspond to introgression and speciation events in the Pleistocene epoch for annual plants. In contrast, divergence events in our Brassicaceae dataset date to deeper time points corresponding to the Miocene epoch 13 to 9 million years ago (Beilstein et al., 2010; Huang et al., 2015). Consistent with this, comparisons of the node depth values in our simulated (Figure 3) versus empirical (Figure 5) data show that the node depths in the Brassicaceae analysis (Figure 5B, C) fall on the high end of the spectrum of simulation scenarios we tested, where ghostbuster performance was more robust. Moreover, the shapes of the node depth distributions do not appear to be bimodal. Both observations suggest that our empirical analysis was not meaningfully impacted by ILS, which is an encouraging sign that our ghostbuster results for the Brassicaceae data are reliable.

Our simulations of unbalanced population size showed that ingroup introgression can mimic ghost lineage introgression when P3 is larger than other populations (Figure 4C); in our Brassicaceae dataset, this would be the case if the *Arabidopsis* lineage showed signs of an exceptionally large population. This does not appear to be the case, as there is evidence that the ancestral *Arabidopsis* population is instead smaller than the ancestral *Capsella* lineage (Foxe et al., 2009; Durvasula et al., 2017). Therefore, based on these estimates, we do not see evidence that population size imbalance has skewed our empirical analyses. Taken together, our work adds important theoretical and methodological advancement to a growing body of work that seeks to resolve the fine details of introgression events identified via tractable four-taxon tests.

AUTHOR CONTRIBUTIONS

E.S.F. conceived the project. E.S.F., D.A.C., and D.Y.M. wrote computer code for the analyses. B.S.P. tested the code and ran analyses. E.S.F. and B.S.P. drafted and edited the manuscript, and all authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors thank Daniel Sloan and members of the Sloan laboratory at Colorado State University for helpful discussion. This work was supported by funding from the National Science Foundation (IOS-2114641) awarded to E.S.F.

DATA AVAILABILITY STATEMENT

Ghostbuster code is freely available at: https://github.com/EvanForsythe/Ghost_introgression. Python scripts used to generate all simulated data are also available with the ghostbuster distribution. Data used for empirical analyses in this work are available with the original publication (Forsythe et al., 2020a).

ORCID

Evan S. Forsythe https://orcid.org/0000-0002-3865-2245

REFERENCES

Baumdicker, F., G. Bisschop, D. Goldstein, G. Gower, A. P. Ragsdale, G. Tsambos, S. Zhu, et al. 2022. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* 220: iyab229.

Beilstein, M. A., I. A. Al-Shehbaz, and E. A. Kellogg. 2006. Brassicaceae phylogeny and trichome evolution. *American Journal of Botany* 93: 607–619.

Beilstein, M. A., I. A. Al-Shehbaz, S. Mathews, and E. A. Kellogg. 2008. Brassicaceae phylogeny inferred from phytochrome A and *ndhF* sequence data: Tribes and trichomes revisited. *American Journal of Botany* 95: 1307–1327.

Beilstein, M. A., N. S. Nagalingum, M. D. Clements, S. R. Manchester, and S. Mathews. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* 107: 18724–18728.

Chapman, B. A., and J. T. Chang. 2000. Biopython: Python tools for computational biology. *ACM SIGBIO Newsletter* 20: 15–19.

Dagilis, A. J., D. Peede, J. M. Coughlan, G. I. Jofre, E. R. R. D'Agostino, H. Mavengere, A. D. Tate, and D. R. Matute. 2021. 15 years of introgression studies: Quantifying gene flow across Eukaryotes. bioRxiv 2021.06.15.448399 [preprint]. Available at: https://doi.org/10.1101/2021.06.15.448399 [posted 16 June 2021; accessed 29 January 2026].

Dasmahapatra, K. K., J. R. Walters, A. D. Briscoe, J. W. Davey, A. Whibley, N. J. Nadeau, A. V. Zimin, et al. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487: 94–98.

Durand, E. Y., N. Patterson, D. Reich, and M. Slatkin. 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution* 28: 2239–2252.

Durvasula, A., A. Fulgione, R. M. Gutaker, S. I. Alacakaptan, P. J. Flood, C. Neto, T. Tsuchimatsu, et al. 2017. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences, USA* 114: 5213–5218.

Eriksson, A., and A. Manica. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proceedings of the National Academy of Sciences, USA* 109: 13956–13960.

Flouri, T., X. Jiao, B. Rannala, and Z. Yang. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Molecular Biology and Evolution* 37: 1211–1223.

Fontaine, M. C., J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, et al. 2015. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* 347: e1258524.

Forsythe, E. S., A. D. L. Nelson, and M. A. Beilstein. 2020a. Biased gene retention in the face of introgression obscures species relationships. *Genome Biology and Evolution* 12: 1646–1663.

Forsythe, E. S., D. B. Sloan, and M. A. Beilstein. 2020b. Divergence-based introgression polarization. *Genome Biology and Evolution* 12: 463–478.

Foxe, J. P., T. Slotte, E. A. Stahl, B. Neuffer, H. Hurka, and S. I. Wright. 2009. Recent speciation associated with the evolution of selfing in *Capsella*. *Proceedings of the National Academy of Sciences, USA* 106(13): 5241–5245.

Green, R. E., J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, et al. 2010. A draft sequence of the Neandertal genome. *Science* 328: 710–722.

Hahn, M. W., and M. S. Hibbins. 2019. A three-sample test for introgression. *Molecular Biology and Evolution* 36: 2878–2882.

Hibbins, M. S., and M. W. Hahn. 2019. The timing and direction of introgression under the multispecies network coalescent. *Genetics* 211: 1059–1073.

Hibbins, M. S., and M. W. Hahn. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics* 220: iyab173.

Huang, C.-H., R. Sun, Y. Hu, L. Zeng, N. Zhang, L. Cai, Q. Zhang, et al. 2015. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. *Molecular Biology and Evolution* 33: 394–412.

Hudson, R. R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337–338.

Hufford, M. B., P. Lubinksy, T. Pyhäjärvi, M. T. Devengenzo, N. C. Ellstrand, and J. Ross-Ibarra. 2013. The genomic signature of crop-wild introgression in maize. *PLoS Genetics* 9: e1003477.

Jiao, X., T. Flouri, and Z. Yang. 2021. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *National Science Review* 8: nwab127.

Kalyaanamoorthy, S., B. Q. Minh, T. K. F. Wong, A. Von Haeseler, and L. S. Jermiin. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods* 14: 587–589.

Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Kelleher, J., A. M. Etheridge, and G. McVean. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology* 12(5): e1004842.

Kingman, J. F. C. 1982. The coalescent. *Stochastic Processes and their Applications* 13: 235–248.

Kuhlwilm, M., I. Gronau, M. J. Hubisz, C. De Filippo, J. Prado-Martinez, M. Kircher, Q. Fu, et al. 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature* 530: 429–433.

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology and Evolution* 20: 229–237.

Mallet, J., N. Besansky, and M. W. Hahn. 2016. How reticulated are species? *BioEssays* 38: 140–149.

Martin, S. H., J. W. Davey, and C. D. Jiggins. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Molecular Biology and Evolution* 32: 244–257.

Martin, S. H., and C. D. Jiggins. 2017. Interpreting the genomic landscape of introgression. *Current Opinion in Genetics and Development* 47: 69–74.

Mora, C., D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm. 2011. How many species are there on earth and in the ocean? *PLoS Biology* 9: e1001127.

Nguyen, L. T., H. A. Schmidt, A. Von Haeseler, and B. Q. Minh. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32: 268–274.

Oyama, R. K., M. J. Clauss, N. Formanová, J. Kroymann, K. J. Schmid, H. Vogel, K. Weniger, et al. 2008. The shrunken genome of *Arabidopsis thaliana*. *Plant Systematics and Evolution* 273: 257–271.

Pang, X. X., and D. Y. Zhang. 2024. Detection of ghost introgression requires exploiting topological and branch length information. *Systematic Biology* 73: 207–222.

Pease, J. B., and M. W. Hahn. 2015. Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology* 64: 651–662.

Prüfer, K., F. Racimo, N. Patterson, F. Jay, S. Sankararaman, S. Sawyer, A. Heinze, et al. 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 505: 43–49.

Rieseberg, L. H. 2006. Hybrid speciation in wild sunflowers. *Annals of the Missouri Botanical Garden* 93: 34–48.

Rieseberg, L. H., and D. E. Soltis. 1991. Phylogenetic consequences of cytoplasmic gene flow in plants. *Evolutionary Trends in Plants* 5: 65–84.

Schranz, M. E., L. Kantama, H. De Jong, and T. Mitchell-Olds. 2006. Asexual reproduction in a close relative of *Arabidopsis*: A genetic investigation of apomixis in *Boechera* (Brassicaceae). *New Phytologist* 171: 425–438.

Slon, V., F. Mafessoni, B. Vernot, C. de Filippo, S. Grote, B. Viola, M. Hajdinjak, et al. 2018. The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* 561: 113–116.

Suarez-Gonzalez, A., C. A. Hefer, C. Christe, O. Corea, C. Lexer, Q. C. B. Cronk, and C. J. Douglas. 2016. Genomic and functional approaches reveal a case of adaptive introgression from *Populus balsamifera* (balsam poplar) in *P. trichocarpa* (black cottonwood). *Molecular Ecology* 25: 2427–2442.

Taylor, S. A., and E. L. Larson. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology and Evolution* 3: 170–177.

Than, C., D. Ruths, and L. Nakhleh. 2008. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9: e322.

Tiley, G. P., T. Flouri, X. Jiao, J. W. Poelstra, B. Xu, T. Zhu, B. Rannala, et al. 2023. Estimation of species divergence times in presence of cross-species gene flow. *Systematic Biology* 72: 820–836.

Tricou, T., E. Tannier, and D. M. de Vienne. 2022a. Ghost lineages can invalidate or even reverse findings regarding gene flow. *PLoS Biology* 20: e3001776.

Tricou, T., E. Tannier, and D. M. De Vienne. 2022b. Ghost lineages highly influence the interpretation of introgression tests. *Systematic Biology* 71: 1147–1158.

Wang, T., A. D. J. Van Dijk, J. Bucher, J. Liang, J. Wu, G. Bonnema, and X. Wang. 2023. Interploidy introgression shaped adaptation during the origin and domestication history of *Brassica napus*. *Molecular Biology and Evolution* 40: msad199.

Wong, Y., A. Ignatieva, J. Koskela, G. Gorjanc, A. W. Wohns, and J. Kelleher. 2024. A general and efficient representation of ancestral recombination graphs. *Genetics* 228(1): iyae100.

Yakimowski, S. B., and L. H. Rieseberg. 2014. The role of homoploid hybridization in evolution: A century of studies synthesizing genetics and ecology. *American Journal of Botany* 101: 1247–1258.

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

Zheng, Y., and A. Janke. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinformatics* 19: e10.

How to cite this article: Forsythe, E. S., B. S. Pappa, D. A. Clavette, and D. Y. Mendoza. 2026. Detecting cryptic ghost lineage introgression in four-taxon genomic datasets. *Applications in Plant Sciences* 14(3): e70045. https://doi.org/10.1002/aps3.70045