frontiers in Plant Science

ORIGINAL RESEARCH
published: 14 October 2021
doi: 10.3389/fpls.2021.677009

# There Is No 'Rule of Thumb': Genomic Filter Settings for a Small Plant Population to Obtain Unbiased Gene Flow Estimates

*Alison G. Nazareno*[1,2]* *and L. Lacey Knowles*[1]

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, United States, [2] Department of Genetics, Ecology and Evolution, Federal University of Minas Gerais, Belo Horizonte, Brazil

The application of high-density polymorphic single-nucleotide polymorphisms (SNP) markers derived from high-throughput sequencing methods has heralded plenty of biological questions about the linkages of processes operating at micro- and macroevolutionary scales. However, the effects of SNP filtering practices on population genetic inference have received much less attention. By performing sensitivity analyses, we empirically investigated how decisions about the percentage of missing data (MD) and the minor allele frequency (MAF) set in bioinformatic processing of genomic data affect direct (i.e., parentage analysis) and indirect (i.e., fine-scale spatial genetic structure – SGS) gene flow estimates. We focus specifically on these manifestations in small plant populations, and particularly, in the rare tropical plant species *Dinizia jueirana-facao*, where assumptions implicit to analytical procedures for accurate estimates of gene flow may not hold. Avoiding biases in dispersal estimates are essential given this species is facing extinction risks due to habitat loss, and so we also investigate the effects of forest fragmentation on the accuracy of dispersal estimates under different filtering criteria by testing for recent decrease in the scale of gene flow. Our sensitivity analyses demonstrate that gene flow estimates are robust to different setting of MAF (0.05–0.35) and MD (0–20%). Comparing the direct and indirect estimates of dispersal, we find that contemporary estimates of gene dispersal distance ($\sigma_{rt}$ = 41.8 m) was ~ fourfold smaller than the historical estimates, supporting the hypothesis of a temporal shift in the scale of gene flow in *D. jueirana-facao*, which is consistent with predictions based on recent, dramatic forest fragmentation process. While we identified settings for filtering genomic data to avoid biases in gene flow estimates, we stress that there is no 'rule of thumb' for bioinformatic filtering and that relying on default program settings is not advisable. Instead, we suggest that the approach implemented here be applied independently in each separate empirical study to confirm appropriate settings to obtain unbiased population genetics estimates.

**Keywords: conservation genetics, *Dinizia jueirana-facao*, Fabaceae, spatial genetic structure, parentage assignment**

OPEN ACCESS

***Edited by:***
*Mario Fernández-Mazuecos, Complutense University of Madrid, Spain*

***Reviewed by:***
*Isaac Overcast, INSERM U1024 Institut de Biologie de l'Ecole Normale Supérieure, France*
*José Luis Blanco Pastor, INRA Centre de Recherche de Poitou-Charentes, France*

***Correspondence:**
*Alison G. Nazareno*
*alisongn@ufmg.br*

***Specialty section:***
*This article was submitted to Plant Systematics and Evolution, a section of the journal Frontiers in Plant Science*

***Received:*** *07 March 2021*
***Accepted:*** *16 June 2021*
***Published:*** *14 October 2021*

***Citation:***
*Nazareno AG and Knowles LL (2021) There Is No 'Rule of Thumb': Genomic Filter Settings for a Small Plant Population to Obtain Unbiased Gene Flow Estimates. Front. Plant Sci. 12:677009. doi: 10.3389/fpls.2021.677009*

## INTRODUCTION

High-throughput sequencing technologies that take advantage of restriction endonuclease enzymes to generate reduced representations of genomes (Davey et al., 2011; Andrews et al., 2016) are enabling us to identify, sequence, and genotype thousands of SNPs (i.e., single-nucleotide polymorphisms) in any kind of organism. This use of high-density biallelic SNP markers has heralded a plethora of evolutionary questions at a genome-level in non-model organisms, improving our understanding of the underlying processes at micro- and macroevolutionary scales (Alencar and Quental, 2019; Myers et al., 2019). Furthermore, the increasing number and density of molecular markers across the genome can give more statistical power for accurate population genetic parameters (e.g., Luikart et al., 2003; Nazareno et al., 2017). This becomes invaluable for addressing questions where the processes of interest act locally, and hence at finer spatial and temporal scales, rather than at large spatial or temporal scales. For example, the negative effects of habitat fragmentation often manifest at local scales, especially in organisms with limited dispersal capabilities. As such, the effect of fragmentation may only be detectable with the resolution of genomic data. For example, analysis of hundreds of SNPs in an endangered salamander revealed the effects of fragmentation on genetic diversity and structure (McCartney-Melstad et al., 2018), but such effects went undetected in analyses of few microsatellite markers (Titus et al., 2014).

With time- and cost-efficient techniques and lower genotyping error than other molecular markers (e.g., microsatellites) (e.g., Davey et al., 2011; Seeb et al., 2011; O'Leary et al., 2018), the sharp rise in applications of SNPs in population genetic studies (Andrews et al., 2016) has also been accompanied by many studies on best practices. These include details ranging from library preparation to bioinformatic processing for quality controls to improve the accuracy of SNP data sets (e.g., DePristo et al., 2011; Gautier et al., 2013; Ilut et al., 2014; Mastretta-Yanes et al., 2015; Andrews et al., 2016; Paris et al., 2017; Willis et al., 2017; O'Leary et al., 2018; Díaz-Arce and Rodríguez-Ezpeleta, 2019; Cumer et al., 2021). However, the effects of some SNP filtering practices, especially in relation to parameters regarding the frequency of missing data (MD) and minor allele frequency (MAF), and their effects on population genetic inference have received much less attention (Huang and Knowles, 2014; Andrews et al., 2018; Hall et al., 2020). Some questions, especially those focused on local spatial and temporal scales, are no doubt disproportionately affected by these filtering practices. Ironically, these are also scenarios where the large number of SNPs are required for distinguishing among hypotheses, where such distinction rests on subtle differences in the allele frequency spectrum, and yet, this spectrum is sensitive to filtering practices (see Huang and Knowles, 2014), biasing parameter estimates (Larson et al., 2020).

Here we address the unintended consequences of filtering practices of RADseq on estimates of dispersal, focusing on the MD and MAF settings used in bioinformatic processing. Specifically, we examine the effects of filtering on dispersal estimates for an endangered species threatened with extinction, *Dinizia jueirana-facao* G. P. Lewis and G. S. Siqueira (Fabaceae: Caesalpinioideae), where accurate parameter estimation has downstream consequences for conservation decisions; conservation concerns were the primary motivation for the collection of genetic data in this species. This recently discovered species (Lewis et al., 2017) is facing extinction risks due to habitat fragmentation and degradation. With notable reductions in populations, the species has become increasingly rare, with a few remaining small populations. Consequently, both direct measures of dispersal from parentage analyses, as well as indirect estimates of dispersal from fine-scale spatial genetic structure (SGS; i.e., non-random spatial distribution of genotypes within populations), provide useful information for management activities and policies, including seed collection for *ex situ* conservation, tree breeding, and/or reforestation (e.g., Bittencourt and Sebbenn, 2007; Ramos et al., 2018; de Oliveira et al., 2020).

While the effects of MD and MAF used in bioinformatic processing of SNP data on fine-scale SGS (Attard et al., 2018) employing different relatedness coefficients (e.g., Loiselle et al., 1995; Ritland, 1996, kinship estimators; Queller and Goodnight, 1989) have been investigated more generally (see Hellmann et al., 2016; de Fraga et al., 2017; Escoda et al., 2017; Attard et al., 2018), the results (and therefore recommendations for best practices) from such studies may not be generalizable to small populations for a number of reasons. For example, studies have shown that parentage assignments are accurate with high statistical power when the frequency of both alleles is close to 0.5 (Anderson and Garza, 2006; Baruch and Weller, 2008; Strucken et al., 2016; Andrews et al., 2018; Dussault and Boulding, 2018; but see Andrews et al., 2018) or when there is no missing data in the SNP data set (see Dussault and Boulding, 2018). However, such conditions are unlikely to be met when studying threatened species. Such taxa generally have low genetic variation and their populations are often comprised of closely related individuals.

By informing our study through the analysis of empirical data, we assure that the observed effects of MAF and MD settings when processing genomic data are consistent with the biological realities of being a rare, endangered plant species. We follow our analyses of the endangered plant species *D. jueirana-facao* with a discussion of why different MD and MAF settings among taxa are likely necessary to obtain unbiased estimates of dispersal, as opposed to general guidelines about MD and MAF that do not consider specific applications of RADseq data (e.g., Andrews et al., 2018; Dussault and Boulding, 2018). By comparing the direct and indirect gene flow estimates, we also investigate the effects of forest fragmentation by testing the hypothesis of recent decrease in the scale of gene flow expressed by the dispersal distance. This pattern is expected because only direct estimates should be affected due to the temporal inertia of indirect estimates for few generations (Dutech et al., 2005; Oddou-Muratorio and Klein, 2008). Lastly, evaluation of the sensitivity of both direct and indirect estimates of gene flow to settings of MD and MAF for SNP data sets shows that the assumptions of general bioinformatic guidelines are not likely to be met in species that have recently undergone declines and/or are rare.

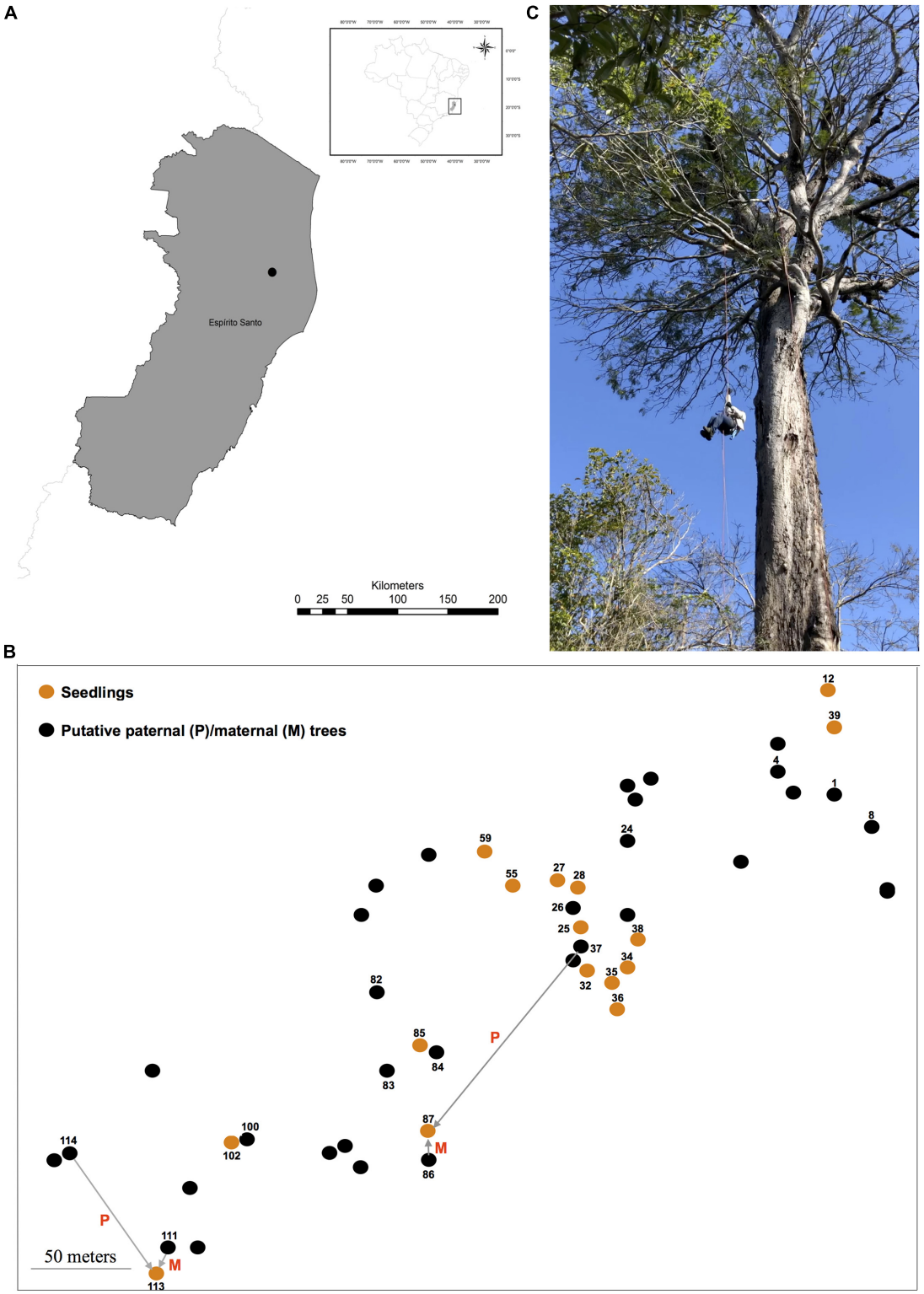


**FIGURE 1 |** The location **(A)** of the studied population (black dot) in the Atlantic Forest (Reserva Natural Vale, Espírito Santo State, Southeast Brazil). The spatial distribution of reproductive trees (i.e., putative maternal and/or paternal parents) and seedlings of *Dinizia jueirana-facao* sampled in the small population is shown in **(B)**. The reproductive trees (black dots) that were assigned as a maternal and/or paternal parent of the seedlings (orange dots) are also shown in **(B)** (displayed numbers; see **Table 3** to identify the maternal and/or paternal parents of the seedlings as shown for seedlings 87 and 113). An image of an adult tree **(C)** is also shown.

# MATERIALS AND METHODS

## Focal Taxon, Study Area and Sampling

*Dinizia jueirana-facao* is a narrowly restricted tree species endemic to a small area of the Brazilian Atlantic Forest (**Figure 1**). The inflorescences of *D. jueirana-facao* are composed of hermaphrodite yellow flowers, with some apical flowers appearing functionally male due to suppression of gynoecium development (Lewis et al., 2017), and its scimitar-shaped and woody large fruits (40–46 × 8.5–10 cm) contain black and hard seeds (Lewis et al., 2017). Although there is a morphological characterization of the reproductive structures of *D. jueirana-facao* (Lewis et al., 2017), there is no information about how its pollen and seeds are dispersed to date. This canopy-emergent tree (19–40 m) is Critically Endangered due to ongoing decline in the number of adult trees because of habitat deforestation and occurs in only two localities: one within the Reserva Natural Vale (RNV) and the other ca. 12.0 km away from the reserve. Only 12 adult trees at RNV, and another 12 trees, were previously mapped in these restricted areas, which combined cover a little over 100 hectares (ha) (Lewis et al., 2017). Fortunately, after an intense sampling effort in 2019, we were able to expand this sampling for the species in the RNV to include 16 seedlings (H, Height, <61 cm) and 99 trees (DBH, Diameter at Breast Height, >10.0 cm; H > 4.0 m), 34 of which were reproductive (DBH > 87 cm, H > 9.0 m). We also confirmed the number of individuals that reproduced in the observed event plus those that had reproduced in a previous event based on the presence of dried reproductive pods and/or seeds under the plant.

For the genetic study, leaf samples of reproductive trees (34 plants with H varying from 9 to 30 m and with a DBH ranging from 87 to 490 cm) and 16 seedlings (H varying from 20 to 60 cm) closest to a reproductive tree were collected and mapped (**Figure 1**) within the RNV site. The distances between adult trees ranged from 2.36 to 444.80 m (average 128.00 ± 60.67 m), and the distances between seedlings ranged from 2.90 to 375.40 m (average 112.16 ± 84.05 m).

## Library Preparation and Sequencing

We extracted genomic DNA from leaf samples of 50 individuals using the Macherey-Nagel kit (Macherey-Nagel GmbH & Co. KG), following the manufacturer's instructions. We created one genomic library using a double-digest restriction site-associated DNA sequencing (i.e., ddRADseq) protocol (Peterson et al., 2012), with modifications to minimize the risk of high variance in the number of reads per individual (see Nazareno et al., 2017 for more details). Briefly, PCRs were performed on each individual and amplicons were pooled for size selection, instead of pooling samples prior to PCR as recommended by Peterson et al. (2012). Double-stranded DNA concentrations were quantified using the Qubit dsDNA Assay Kit (Invitrogen) and 0.5 μg for each individual was digested with the high-fidelity restriction enzymes *Eco*RI and *Mse*I (New England Biolabs). Digestion reactions were purified with the Agencourt AMPure XP system (Beckman Coulter), following the manufacturer's instructions, with elution in 40 μL water. Adapter ligations were carried out at 23°C for 30 min in a total volume of 30 μL, combining 80 ng of DNA, 0.35 μM of a non-sample specific *Mse*I adaptor (common for all samples), 0.50 μM of a sample specific *Eco*RI double-strand adaptor for each DNA sample, 1U of T4 DNA ligase (New England BioLabs), and 1.5 × T4 ligase buffer. Reactions were heated at 65°C for 10 min and slowly cooled to 23°C. Ligation products were cleaned with the Agencourt AMPure XP system and amplified following the PCR protocol reported by Nazareno et al. (2017). Multiplexed genomic library was prepared with approximately equal amounts of DNA, and DNA fragments at a target range size of 375–475 bp were size-selected using Pippin Prep and a 2% agarose cartridge (Sage Science, Beverly, MA, United States). The library was sequenced (100 bp single-end reads) on a lane of an Illumina HiSeq 2500 flowcell (Illumina Inc., San Diego, CA, United States) at The Centre for Applied Genomics in Toronto, Canada.

## SNPs Identification

Files containing the raw sequence reads were analyzed in Stacks 2.41 (Catchen et al., 2011; Catchen et al., 2013; Rochette et al., 2019) using *de novo* assembly. We used the process_radtags program in Stacks to initially assign reads to individuals and eliminate poor quality reads and reads missing the expected *Eco*RI cut site (options –barcode_dist 2 -q -e ecoRI). All sequences were processed in ustacks to produce consensus sequences of RAD tags, applying a maximum-likelihood framework to estimate the diploid genotype for each individual at each nucleotide position (Hohenlohe et al., 2011). The optimum minimum depth of coverage to create a stack was set at three sequences, the maximum distance allowed between stacks was two nucleotides, and the maximum number of stacks allowed per *de novo* locus was three. The stacks assembly enabled the Deleveraging algorithm (–d), which resolves over-merged tags, and the Removal algorithm (–r), which drops highly repetitive stacks and nearby errors from the algorithm. The alpha value for the SNP model was set at 0.05; as reported by Catchen et al. (2013); low alpha values (i.e., <0.10) avoid underestimating true heterozygous genotypes. Cstacks was used to build a catalog of consensus loci containing all the loci from all the individuals and merging all alleles together. After processing the consensus loci in cstacks, stacks generated were searched against the catalog and SNPs were called using sstacks, tsv2bam, and gstacks (Rochette et al., 2019), with default settings.

## Data Sets With Different Amounts of Missing Data (MD) and Minor Allele Frequency (MAF)

We used POPULATIONS in Stacks (Catchen et al., 2011, 2013; Rochette et al., 2019) to create data sets with five different MD settings (i.e., 0, 5, 10, 15, and 20%) and seven different MAF settings (i.e., 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, and 0.35), for a total of 35 data sets. Note, it was not possible to generate datasets with MD > 20% and MAF > 0.35 because the combination of such parameter settings resulted in very small number of SNPs in our empirical data sets. All data sets include one SNP per

locus, which were identified after filtering loci to confirm Hardy–Weinberg (H-W) equilibrium and linkage disequilibrium (LD) using the adegenet package[1] (Jombart, 2008; Jombart and Ahmed, 2011) implemented in R (R Core Team, 2018) and Arlequin 3.5.2 (Excoffier and Lischer, 2010), respectively. Type I error rates for these tests were corrected for multiple tests using the sequential Bonferroni procedure (Rice, 1989) and SNPs that failed the H-W equilibrium test and/or SNP pairs in LD were excluded.

## Assessing the Effects of MD and MAF on SGS and Indirect Dispersal Distance

Assuming that genotypes of all *D. jueirana-facao* come from a two-dimensional population at drift-dispersal equilibrium, the spatial genetic structure (SGS) was inferred based on pairwise relatedness coefficients between individuals using the SPAGeDi program (Hardy and Vekemans, 2002). Estimators of kinship (co-ancestry) coefficients ($F_{ij}$) and relationship coefficients ($R_{ij}$) were calculated for different groups of individuals, where the groups differ by the geographic distance separating them. Specifically, the kinship coefficient estimators $F_L$ (Loiselle et al., 1995) and $F_r$ (Ritland, 1996) are based on the probability that a random allele from individual *i* is identical by descent to a random allele from individual *j* (Vekemans and Hardy, 2004); $F_r$ is downward biased when very low frequency alleles occur (Ritland, 1996; Vekemans and Hardy, 2004). The $R_{Q\&G}$ estimator (Queller and Goodnight, 1989) is based on the probability that a random allele from individual *i* is identical to one of the alleles from individual *j* (Vekemans and Hardy, 2004). Although the estimators display high discriminate power in allozymes and microsatellites markers (Vekemans and Hardy, 2004), there is limited information on their performance with SNP markers (e.g., Attard et al., 2018), especially for SNPs collected in small populations. In addition to the close relation between *Fij* and *Rij*, the $F_L$, $F_r$, and $R_{Q\&G}$ estimators can be used in a comparative way (for a more detailed account on statistical properties of these estimators see Queller and Goodnight, 1989; Loiselle et al., 1995; Ritland, 1996), given some estimators make no assumption regarding Wright's inbreeding coefficient (i.e., the probability that two randomly chosen alleles of an individual at any homologous locus are identical by descent; Malécot, 1948). Individuals were grouped into six distance classes to maximize the number of pairs of individuals per distance class and the average multilocus relatedness coefficients [$F_{(d)}$ or $R_{(d)}$] per distance class using the SPAGeDi program (Hardy and Vekemans, 2002). The 95% confidence interval (CI) of the standard error of the relatedness coefficients was calculated using a jackknife procedure across all loci.

In order to test for significant SGS, $F_{ij}$ (or $R_{ij}$), all pairs of individuals were plotted against the log pairwise spatial distance and significance of the regression was assessed by 10,000 permutations of multilocus genotypes. To compare the extent that SGS varies among data sets with different MD- and MAF-values, we calculated the *Sp*-statistic, a synthetic measure of SGS intensity that is less sensitive to the sampling scheme, and that expresses the balance between local genetic drift and gene dispersal within population (Vekemans and Hardy, 2004; Hardy et al., 2006). The *Sp*-statistic is defined as: $Sp = -b/(1-F_1)$, where $b$ is the regression slope of $F_{ij}$ on log spatial distance, and $F_1$ is the mean $F_{ij}$ between individuals for the first distance class (Vekemans and Hardy, 2004); the parameters to calculate the *Sp*-statistic were obtained in SPAGeDi, and to compute *Sp* using the $R_{Q\&G}$ estimator, we converted the $R_{ij}$ values in $F_{ij}$ applying the equation $R_{ij} = 2F_{ij}/(1 + F_I)$ as proposed by Hardy and Vekemans (1999).

To investigate the relative sensitivity of the different relatedness estimators to the MD and MAF, we used the mean $F_{ij}$ and the *Sp*-statistic values computed considering all the sampled individuals ($n = 50$). Following this result, we used the estimator with the minimum relative standard deviation (i.e., coefficient of variation) to assess the effects of MD and MAF on SGS estimates. As there is a direct association between co-ancestry and inbreeding (i.e., in generation $T_1$ the inbreeding coefficient is equal the co-ancestry in generation $T_0$; Cockerham, 1966), we also investigated the effects of MD and MAF on Wright's inbreeding coefficient. These analyses were performed separately for the reproductive trees ($n = 34$) and seedlings ($n = 16$) of *D. jueirana-facao*.

Lastly, the root-mean-squared dispersal distance (σ) was calculated using the *Sp*-statistic and the effective population density $D_e$, which is the product of the census density $D$ and $N_e/N$, the ratio of the effective to the census population size, where $\sigma^2 = N_b/4\ \pi\ D_e$ (Vekemans and Hardy, 2004). While $D_e$ in plant populations can be estimated as $D/4$ (Hardy et al., 2006), the $D$-values of reproductive trees (0.79 ind.ha$^{-1}$) and seedlings (0.37 ind.ha$^{-1}$), were multiplied by 0.30 – the $N_e/N$ ratio directly computed for the *D. jueirana-facao* in the RNV site given that the Wright's neighborhood size ($N_b$) equals $1/Sp$ (Vekemans and Hardy, 2004). The $N_e/N$ ratio in the RNV population (i.e., 0.30) was calculated as the number of reproductive trees ($N_e = 34$) divided by the total number of *D. jueirana-facao* plants [$N = 115$; i.e., adult trees ($n = 99$) + seedlings ($n = 16$)].

## Assessing the Effects of MD and MAF on Parentage Analysis

We used the CERVUS 3.0.7 program (Marshall et al., 1998; Kalinowski et al., 2007) to investigate how MD and MAF affect cryptic gene flow, $C_{gf}$, which expresses the proportion of genotypes assigned to a candidate parent within the sampled area when the true parent is located outside there. $C_{gf}$ was calculated as $1 - (1 - P_p)^n$, where $n$ is the number of candidate parents within the population, and $P_p$ represents the combined non-exclusion probability (i.e., the probability of not excluding a single randomly chosen unrelated individual from parentage over all loci) of the parent pair, when the parent pair is unknown (Dow and Ashley, 1996). Then, we identified the appropriate data set (i.e., those with $C_{gf} \approx$ zero) to be used on the characterization of direct gene flow in the *D. jueirana-facao* population following the categorical parentage analyses described below.

Parentage analyses were performed according to the maximum likelihood method integrated in the CERVUS

[1]https://CRAN.R-project.org/package=adegenet

(Marshall et al., 1998; Kalinowski et al., 2007), with 50,000 simulated genotypes to estimate the critical value of Delta ($\Delta_{crit}$), considering a genotyping error ratio of 1%. The proportion of candidate parents sampled was set at 90%, which was justifiable given that we had genotyped all known adult trees of *D. jueirana-facao* in the RNV population, which takes into account that 10% of candidate parents may have died in recent years. Parentage assignment was performed comparing the value of $\Delta_{crit}$ with the $\Delta$-score. The $\Delta$-score is used as a criterion for the assignment of parentage and it is defined as the difference in LOD scores between the most likely candidate parent and the second most likely candidate parent. As defined by Marshall et al. (1998), the LOD score is the natural logarithm of the likelihood that the candidate parent is the true parent divided by the likelihood that the candidate parent is not the true parent. Gene flow via pollen and seed dispersal was estimated in seedlings ($n$ = 16) considering all reproductive trees ($n$ = 34) as possible maternal and/or paternal candidates. Putative parents were recognized as those with $\Delta > \Delta_{crit}$ with 95% confidence; seedlings in which the same tree was inferred to be the maternal and paternal parent were considered to represent examples of selfing. Pollen and seed dispersal (Euclidean) distances were calculated considering the distance between seedling and the putative parents. Specifically, seedlings with only one putative parent identified within the population were presumed to represent the distance of seed dispersal, assuming that pollen dispersal distances are more likely to come from geographically more distant reproductive trees that were not sampled (Dow and Ashley, 1996). For seedlings with two putative parents identified among the reproductive trees of the study population, the one nearest to the seedling was presumed to reflect seed dispersal (i.e., the maternal parent), whereas the more distant one was presumed to reflect pollen dispersal (i.e., the paternal parent), again using the assumption that the distance traveled by pollen is likely greater than that of seeds, which has been applied in parentage analyses when the maternal parent is not distinguishable (e.g., Dow and Ashley, 1996; Guidugli et al., 2016; Feres et al., 2021), as is the case of the hermaphroditic tree species such as *D. jueirana-facao*. We also calculated seed ($m_s$) and pollen ($m_p$) immigration rates. Specifically, immigrant seeds and immigrant pollen were represented by seedlings without assigned parents or seedlings that had only one putative parent assigned from the population, respectively, and were compared relative to the total number of seedlings to discern $m_s$ and $m_p$ (Burczyk et al., 1996).

Lastly, for comparison with the root-mean-squared dispersal distance obtained from the SGS analysis, we computed the total direct gene flow ($\sigma^2_{rt}$) using the parentage assignment results. Specifically, the total direct gene flow, $\sigma^2_{rt}$ is equal to 1/2 $\sigma^2_{p-rt}$ + $\sigma^2_{s-rt}$, where $\sigma^2_{p-rt}$ and $\sigma^2_{s-rt}$ are the variances of the pollen and seed dispersal distances, respectively (Crawford, 1984).

# RESULTS

About 145 million single-end raw reads were produced on one sequencing lane of HiSeq 2000 Illumina for the 50 individuals included in the genomic library of *D. jueirana-facao*. The mean number of retained reads that passed the quality filters, including a Phred quality score > 33 with identifiable barcodes, were 2,319,619 ± 152,193 SE. The number of polymorphic SNPs with a minimum 10-fold coverage ranged from 256 (MD = 0%, MAF = 0.35) to 6,898 (MD = 20%, MAF = 0.05). No significant departures from HWE were observed in any data set after a Bonferroni adjustment ($p > 0.00019$). In addition, no LD was observed after a sequential Bonferroni correction for $k$ tests (varying from $k = 3.26 \times 10^4$ with $p < 1.53 \times 10^{-6}$ for 256 SNPs to $k = 2.38 \times 10^7$ with $p < 2.10 \times 10^{-9}$ for 6,898 SNPs).

## Effects of MD and MAF on SGS and Indirect Dispersal Distance

Of the different estimators of relatedness, the Loisele's kinship estimator was comparatively less sensitive to the different settings of MD and MAF (coefficient of variation, $CV$ = 1.37%; **Table 1**), as well as derivatives based on the relatedness estimator – namely, the measure of the extent of SGS captured by the *Sp*-statistic ($CV$ = 1.69%; **Table 1**). Although the standard error of all estimators of relatedness increases when the number of loci decreases (**Supplementary Table 1**), the Loisele's kinship and its derivative *Sp*-statistic were both less sensitive to the number of SNPs analyzed than other relatedness measures, showing a weak, non-significant correlation (**Figure 2**). Therefore, we based tests of fine-scale SGS on the Loisele's kinship (see below).

Both seedlings and adult trees show a similar trend regarding the effects of MD and MAF on the fine-scale SGS estimates. However, estimates of $F_1$ and *Sp*-statistic for data sets with different amounts of MD and MAF were more homogenous in seedlings than adult trees (**Supplementary Table 2**), with no statistical differences were observed among the data sets with the different amounts of MD and MAF (**Figure 3**). Note that no significant correlations between $F_1$ or *Sp*-statistic and the number of SNPs were observed in either seedlings or adult trees (**Supplementary Figure 1**). On the other hand, the inbreeding coefficient estimates appear to be more sensitive to both MD and MAF, becoming inflated and statistical significant with increases of MD (**Supplementary Figure 2** and **Supplementary Table 3**). For instance, with larger amounts of MD (with MAF varying from 0.05 to 0.35), the average inbreeding coefficient in seedlings of *D. jueirana-facao* increases and remains stastistical significant for data sets with 0 and 20% MD (e.g., 0% MD = 0.047, CI = 0.037 to 0.057, and 20% MD = 0.097, CI = 0.090 to 0.104; **Supplementary Table 3**). For adult trees of *D. jueirana-facao*, the average inbreeding coefficient was also statistically significant for 0% and 20% MD and shows similar trend (e.g., 0% MD = –0.065, CI = –0.075 to –0.055, and 20% MD = 0.005, CI = 0.001–0.008, **Supplementary Table 3**).

Based on distance class analysis, a significant linear decrease of the Loiselle's kinship coefficients with the linear spatial distance was detected in both seedlings and adult trees (**Figure 4**). However, the shapes of the kinship curves are distinct for seedlings and adult trees (**Figure 4**), with significant positive values (based on the 95% CI) of up to 60 m for seedlings, and up to 100 m for adult trees. For seedlings, the largest Loiselle's

**TABLE 1 |** Effects of different amounts of missing data (MD) and minor allele frequency (MAF) on the estimates of pairwise relatedness statistics expressed by the mean $F_{ij}$ between individuals for the first distance class ($F_1$) for the empirical data set ($n$ = 50; adults and seedlings of *Dinizia jueirana-facao* combined).

| | | Loiselle's kinship | | | | Ritland's kinship | | | | Queller and Goodnight's relatedness | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MD_MAF** | **SNPs** | $F_1$ (<50 m) | *b* | $R^2$ | *Sp* | $F_1$ (<50 m) | *b* | $R^2$ | *Sp* | $F_1$ (<50m)* | *b* | $R^2$ | *Sp* |
| 1.00_0.35 | 256 | **0.054** | **0.033** | 0.187 | 0.0352 | **0.044** | **0.033** | 0.188 | 0.0350 | **0.055** | **0.074** | 0.192 | 0.0780 |
| 1.00_0.30 | 338 | **0.053** | **0.032** | 0.191 | 0.0340 | **0.043** | **0.032** | 0.190 | 0.0337 | **0.054** | **0.071** | 0.197 | 0.0754 |
| 1.00_0.25 | 456 | **0.052** | **0.032** | 0.205 | 0.0342 | **0.041** | **0.032** | 0.205 | 0.0339 | **0.053** | **0.073** | 0.216 | 0.0766 |
| 1.00_0.20 | 537 | **0.053** | **0.034** | 0.217 | 0.0355 | **0.043** | **0.034** | 0.220 | 0.0355 | **0.052** | **0.073** | 0.218 | 0.0771 |
| 1.00_0.15 | 668 | **0.053** | **0.034** | 0.217 | 0.0356 | **0.044** | **0.034** | 0.215 | 0.0353 | **0.052** | **0.072** | 0.212 | 0.0754 |
| 1.00_0.10 | 778 | **0.054** | **0.034** | 0.218 | 0.0358 | **0.046** | **0.034** | 0.214 | 0.0356 | **0.051** | **0.070** | 0.204 | 0.0742 |
| 1.00_0.05 | 829 | **0.053** | **0.033** | 0.219 | 0.0351 | **0.050** | **0.032** | 0.215 | 0.0341 | **0.050** | **0.072** | 0.217 | 0.0755 |
| 0.95_0.35 | 1,011 | **0.051** | **0.032** | 0.195 | 0.0336 | **0.041** | **0.032** | 0.196 | 0.0334 | **0.049** | **0.069** | 0.200 | 0.0726 |
| 0.95_0.30 | 1,090 | **0.052** | **0.032** | 0.201 | 0.0338 | **0.042** | **0.032** | 0.201 | 0.0335 | **0.049** | **0.068** | 0.203 | 0.0717 |
| 0.95_0.25 | 1,406 | **0.051** | **0.032** | 0.201 | 0.0338 | **0.041** | **0.032** | 0.201 | 0.0336 | **0.049** | **0.069** | 0.209 | 0.0726 |
| 0.95_0.20 | 1,721 | **0.052** | **0.033** | 0.207 | 0.0346 | **0.042** | **0.033** | 0.207 | 0.0344 | **0.049** | **0.069** | 0.208 | 0.0729 |
| 0.95_0.15 | 2,099 | **0.052** | **0.033** | 0.207 | 0.0343 | **0.042** | **0.033** | 0.206 | 0.0340 | **0.049** | **0.069** | 0.209 | 0.0716 |
| 0.95_0.10 | 2,492 | **0.052** | **0.033** | 0.210 | 0.0344 | **0.042** | **0.033** | 0.209 | 0.0340 | **0.050** | **0.070** | 0.215 | 0.0715 |
| 0.95_0.05 | 3,029 | **0.051** | **0.032** | 0.211 | 0.0336 | **0.040** | **0.031** | 0.211 | 0.0324 | **0.051** | **0.071** | 0.224 | 0.0726 |
| 0.90_0.35 | 1,352 | **0.053** | **0.033** | 0.207 | 0.0351 | **0.042** | **0.033** | 0.208 | 0.0348 | **0.049** | **0.070** | 0.210 | 0.0738 |
| 0.90_0.30 | 1,777 | **0.053** | **0.033** | 0.208 | 0.0348 | **0.042** | **0.033** | 0.208 | 0.0345 | **0.049** | **0.069** | 0.208 | 0.0729 |
| 0.90_0.25 | 2,267 | **0.052** | **0.033** | 0.206 | 0.0347 | **0.042** | **0.033** | 0.206 | 0.0343 | **0.048** | **0.069** | 0.210 | 0.0713 |
| 0.90_0.20 | 2,753 | **0.053** | **0.033** | 0.211 | 0.0353 | **0.043** | **0.034** | 0.210 | 0.0350 | **0.048** | **0.068** | 0.207 | 0.0720 |
| 0.90_0.15 | 3,342 | **0.053** | **0.033** | 0.211 | 0.0350 | **0.042** | **0.033** | 0.210 | 0.0346 | **0.048** | **0.068** | 0.207 | 0.0721 |
| 0.90_0.10 | 3,884 | **0.053** | **0.033** | 0.212 | 0.0350 | **0.043** | **0.033** | 0.212 | 0.0346 | **0.050** | **0.069** | 0.215 | 0.0722 |
| 0.90_0.05 | 4,666 | **0.052** | **0.032** | 0.214 | 0.0342 | **0.040** | **0.032** | 0.215 | 0.0329 | **0.050** | **0.070** | 0.223 | 0.0732 |
| 0.85_0.35 | 1,630 | **0.052** | **0.033** | 0.206 | 0.0351 | **0.041** | **0.033** | 0.207 | 0.0348 | **0.047** | **0.069** | 0.208 | 0.0735 |
| 0.85_0.30 | 2,159 | **0.052** | **0.033** | 0.205 | 0.0348 | **0.042** | **0.033** | 0.205 | 0.0345 | **0.047** | **0.068** | 0.205 | 0.0729 |
| 0.85_0.25 | 2,748 | **0.052** | **0.033** | 0.205 | 0.0347 | **0.041** | **0.033** | 0.204 | 0.0344 | **0.047** | **0.068** | 0.208 | 0.0717 |
| 0.85_0.20 | 3,324 | **0.053** | **0.033** | 0.209 | 0.0353 | **0.042** | **0.034** | 0.209 | 0.0350 | **0.047** | **0.068** | 0.208 | 0.0722 |
| 0.85_0.15 | 4,022 | **0.053** | **0.033** | 0.209 | 0.0350 | **0.042** | **0.033** | 0.209 | 0.0346 | **0.047** | **0.068** | 0.206 | 0.0716 |
| 0.85_0.10 | 4,675 | **0.053** | **0.033** | 0.212 | 0.0350 | **0.042** | **0.033** | 0.212 | 0.0346 | **0.049** | **0.069** | 0.214 | 0.0712 |
| 0.85_0.05 | 5,569 | **0.052** | **0.032** | 0.213 | 0.0342 | **0.040** | **0.032** | 0.215 | 0.0329 | **0.049** | **0.067** | 0.223 | 0.0718 |
| 0.80_0.35 | 2,042 | **0.052** | **0.033** | 0.207 | 0.0352 | **0.041** | **0.033** | 0.208 | 0.0348 | **0.047** | **0.068** | 0.208 | 0.0705 |
| 0.80_0.30 | 2,675 | **0.052** | **0.033** | 0.208 | 0.0351 | **0.041** | **0.033** | 0.208 | 0.0347 | **0.047** | **0.068** | 0.207 | 0.0725 |
| 0.80_0.25 | 3,388 | **0.052** | **0.033** | 0.207 | 0.0351 | **0.041** | **0.033** | 0.207 | 0.0348 | **0.047** | **0.068** | 0.211 | 0.0716 |
| 0.80_0.20 | 4,088 | **0.053** | **0.034** | 0.211 | 0.0357 | **0.042** | **0.034** | 0.211 | 0.0354 | **0.048** | **0.069** | 0.212 | 0.0722 |
| 0.80_0.15 | 4,959 | **0.053** | **0.034** | 0.212 | 0.0354 | **0.042** | **0.034** | 0.212 | 0.0350 | **0.048** | **0.068** | 0.211 | 0.0717 |
| 0.80_0.10 | 5,762 | **0.053** | **0.033** | 0.214 | 0.0353 | **0.042** | **0.033** | 0.214 | 0.0348 | **0.049** | **0.069** | 0.217 | 0.0729 |
| 0.80_0.05 | 6,898 | **0.052** | **0.033** | 0.215 | 0.0344 | **0.039** | **0.032** | 0.218 | 0.0330 | **0.050** | **0.070** | 0.225 | 0.0735 |
| **Average** | | 0.052 | 0.033 | 0.208 | 0.0350 | 0.042 | 0.033 | 0.208 | 0.034 | 0.049 | 0.069 | 0.210 | 0.073 |
| ***SD*** | | 0.001 | 0.001 | 0.007 | 0.001 | 0.002 | 0.001 | 0.007 | 0.001 | 0.002 | 0.002 | 0.007 | 0.002 |
| **CV%** | | 1.37 | 1.63 | 3.31 | 1.69 | 4.46 | 2.19 | 3.22 | 2.26 | 3.93 | 2.26 | 3.40 | 2.44 |

*Also shown are the regression slopes of the pairwise values between individuals on the logarithm of the spatial distance (b), the coefficient of determination ($R^2$), and the Sp-statistic–a synthetic measure of spatial genetic structure (SGS) intensity. Bold values denote statistical significance at $p < 0.05$.*

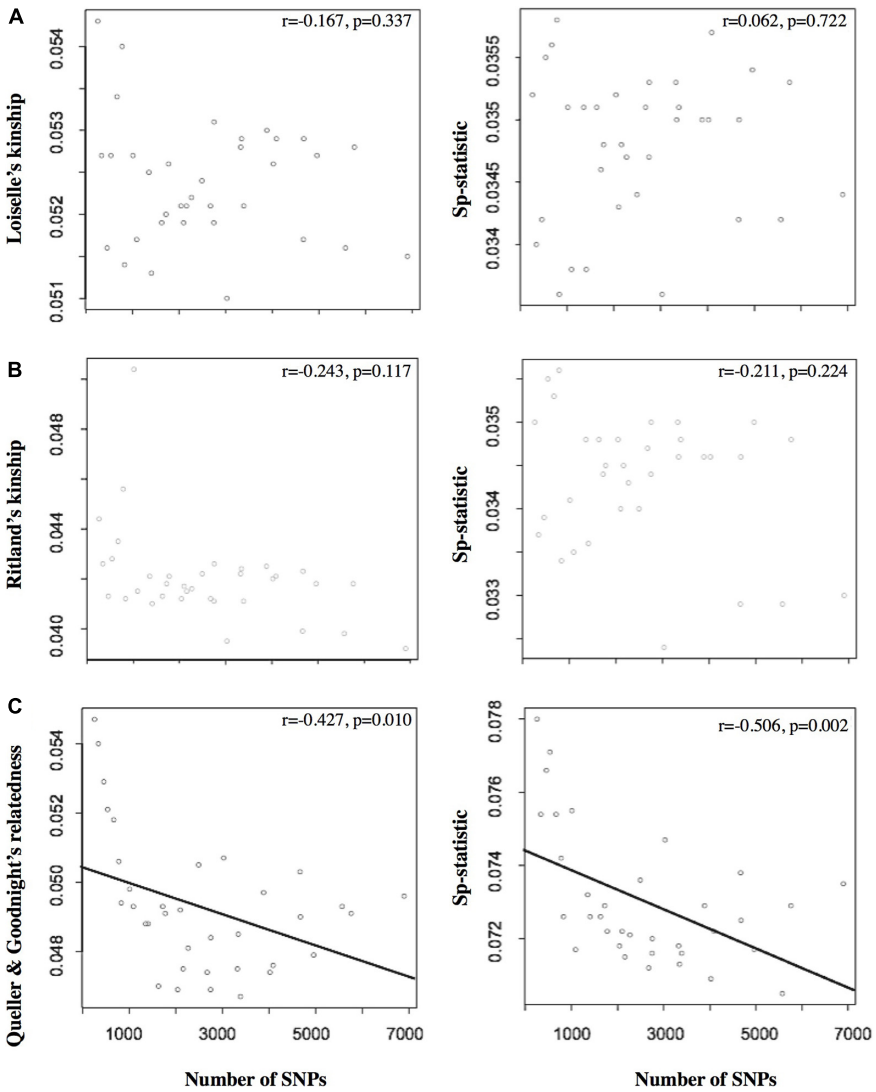


**FIGURE 2 |** The left column presents the relationship between the number of SNPs and the relatedness estimators based on **(A)** Loiselle's kinship (Loiselle et al., 1995), **(B)** Ritland's kinship (Ritland, 1996), and **(C)** Queller and Goodnight's relatedness. The right column shows the relationship between the number of SNPs and the *Sp*-statistic. Note that the number of SNPs differed as a function of the 35 variable MD/MAF data sets (saw as points in the plots) used to filter the SNP data sets during bioinformatic processing. The Pearson's correlation coefficient (*r*) and significance are also shown for each plot.

kinship coefficient ($F_L = 0.155$, $p < 0.05$) was estimated in the first class of distance (0–55 m). This value is between the theoretical expectation for half-siblings ($F_L = 0.125$) and full-sibs ($F_L = 0.25$). For adult trees, the largest kinship coefficient was observed in the first class of distance ($F_L = 0.031$, $p < 0.05$), a value consistent with expectations for second cousins (0.0312). Broadly speaking, stronger SGS was detected in seedlings compared to adult trees, as measured by *Sp*-statistic (*Sp* = 0.0651 in seedlings, *Sp* = 0.0208 in adults; **Table 2**).

The neighborhood size was estimated to be 15 individuals in the seedlings vs. 48 individuals in the adult trees' generation. The average gene dispersal distances were estimated to be 156 m in the seedlings versus 277 m in the adult trees, with small confidence intervals for both (**Table 2**).

## Effects of MD and MAF on Parentage Analysis

The combined non-exclusion probabilities among parent pairs were extremely low, varying from $2.603^{-227}$ to 0.000 (**Supplementary Table 4**), irrespective of the MD and MAF settings. That is, the probability of cryptic gene flow was equal to zero for all but two data sets (**Supplementary Table 4**), indicating that there is no apparent sensitivity of parentage analysis with respect to the MD and MAF values for this empirical dataset.

In order to do comparisons with the indirect gene flow estimates, we also used the data set with 5% of MD and a MAF of 0.05 to quantify the direct gene flow in *D. jueirana-facao*. We assigned a maternal and paternal parent to 11

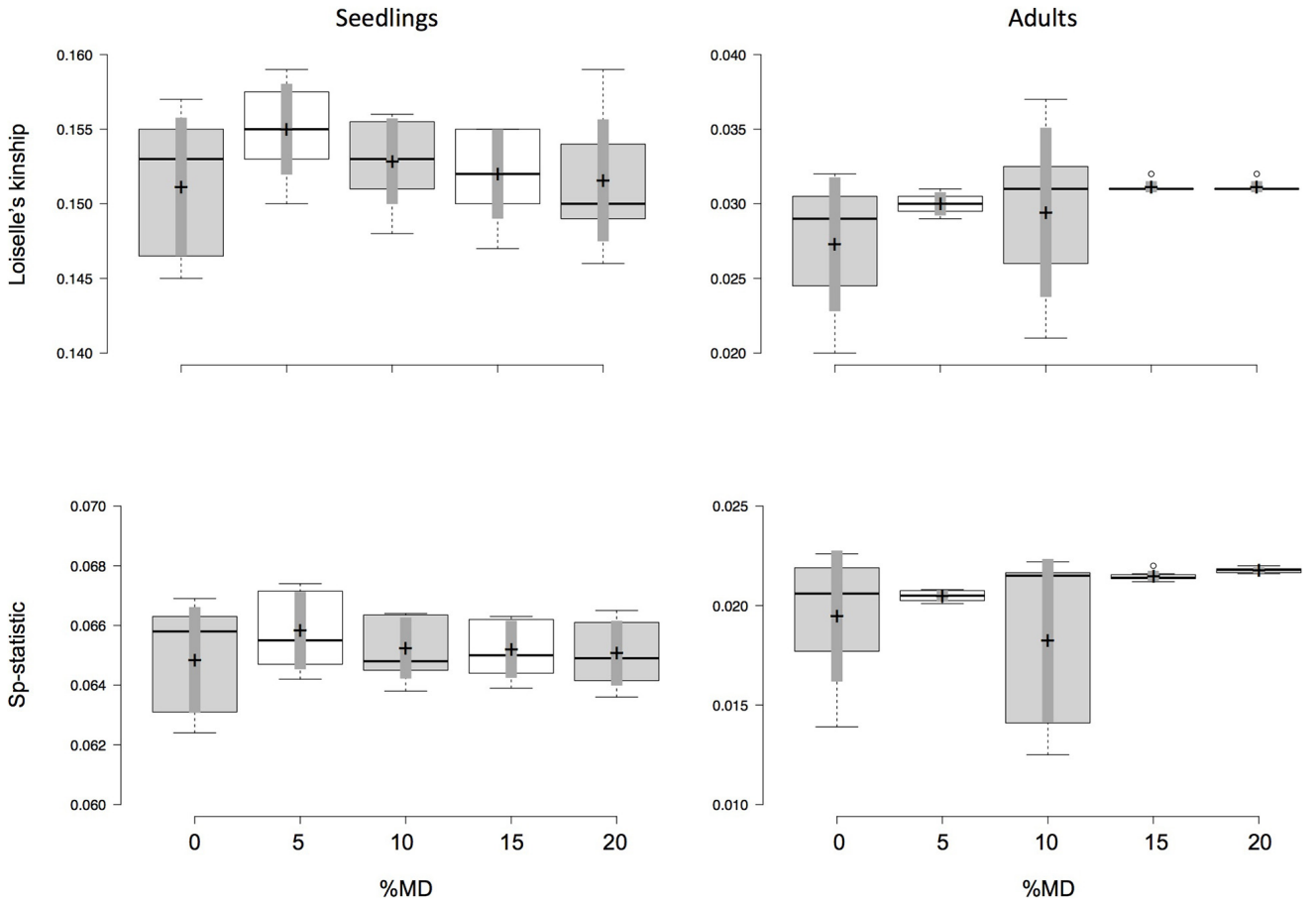


**FIGURE 3 |** Effect of missing data (MD) and minor allele frequency (MAF) on values of Loiselle's kinship measure for the first distance class ($F_1$), which is less than 55 m between seedling and putative parent, and its derivative *Sp*-statistic for seedlings (left column) and adults (right column) of *Dinizia jueirana-facao*. The amount of MAF varied from 0.05 to 0.35 (i.e., seven data sets) for each fixed percentage of MD. Boxplots show the median (center line) and mean (marked by cross, with 95% confidence in gray) of $F_1$ and *Sp*-statistic, and the 25th and 75th percentiles (by the box) with the extent of the whiskers marking 1.5 times the interquartile range from the 25th and 75th percentiles, and outliers (marked by dots), as calculated using R software.

(69%) of the 16 seedlings with 95% confidence (**Table 3**). For the remaining five seedlings, a putative mother tree could be assigned. Among the 16 seedlings with assigned parentage, eight of the 34 (24%) reproductive trees in the RNV population were assigned as the maternal parent, and five of the 34 (15%) reproductive trees were assigned as the paternal parent; six seedlings were likely a product of selfing ($s$ = 0.375). In total, only nine of 34 (26.5%) reproductive trees of *D. jueirana-facao* are parents of at least one seedling.

The effective pollination distance was greater than the effective seed migration distance, with reproductive trees involved in pollination occurring 31.1–147.4 m (average of 73.80 ± 42.9 m) from the seedlings. The seed dispersal distance varied from 1.0 to 79.5 m, with an average of 19.11 ± 23.72 m. By taking into account the pollen and seed dispersal variances, the total direct gene flow distance of 41.8 m was 3.74-fold smaller than the one based on indirect estimates of gene flow for the seedlings (**Table 2**).

## DISCUSSION

Our approach shows how different measures of gene flow (direct and indirect) are sensitive, or conversely robust, to the settings for the percentage of MD and MAF used in the bioinformatic processing of SNP data in a small plant population for which the very characteristics of being a small population, challenge the assumptions made in other studies investigating the robustness of gene flow measures to different settings used to filter genomic data. That is, although previous studies have investigated the effects of data filtering on SGS (Weir and Goudet, 2017; Attard et al., 2018; Goudet et al., 2018) and parentage analyses (e.g., Anderson and Garza, 2006; Baruch and Weller, 2008; Andrews et al., 2018; Dussault and Boulding, 2018; Hall et al., 2020), our study focuses on the combined effect of MD and MAF, which depending on the measure used to estimate gene flow, can introduce biases when applied to small populations specifically. Our approach is a general one that any researcher might apply to evaluate the potential

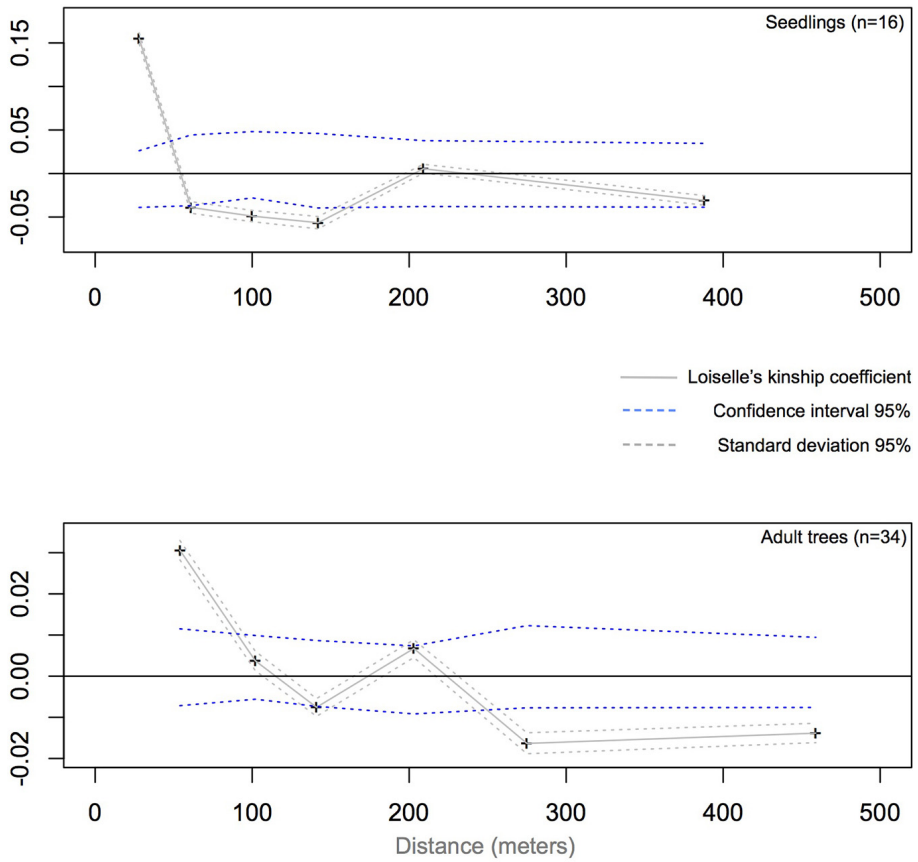


**FIGURE 4 |** Average Loiselle's kinship coefficient ($F_L$) plotted against geographical distance (in gray solid line) between seedlings **(top plot)** and adults **(bottom plot)** of *Dinizia jueirana-facao*. Individuals whose standard deviation of kinship coefficient lie outside the 95% CIs (blue dashed lines) are significantly more similar than would be expected through random sampling. For both reproductive trees and seedlings, a decrease in the kinship values was detected with increasing distances, with significant negative values indicating that distant individuals are not genetically related. The results are based on analyses of a data set generated with 5% MD and a 5% minor allele frequency setting.

**TABLE 2 |** Estimates of fine-scale spatial genetic structure (SGS) and of historical dispersal distance for seedlings and adult tress of *Dinizia jueirana-facao,* as well as the inbreeding coefficient (*F*), the average kinship coefficient ($F_1$) between individuals for the first distance class (i.e., the smallest distance class, which includes distances < 55 m) and its standard error (SE), the *Sp*-statistic, the neighborhood size, *Nb* and its 95% CI, and the root-mean-squared dispersal distance, σ and its 95% CI, are shown considering the effective densities of 0.237 ($D_e$ = 0.79 × 0.30 for adult trees) and 0.111 ($D_e$ = 0.37 × 0.30 for seedlings).

| | SGS parameters | | | | Gene dispersal estimates | |
|---|---|---|---|---|---|---|
| | ***F*** | ***$F_1$*** | ***SE*** | ***Sp*** | ***Nb*** | **σ (m)** |
| Seedlings | **0.071** | **0.1553** | 0.0032 | 0.0651 | 15.37 (15.2–15.4) | 156.40 (155.8–156.9) |
| Adults | **–0.034** | **0.0306** | 0.0012 | 0.0208 | 48.07 (47.9–48.2) | 276.90 (276.3–276.9) |

*Analyses are based on a data set with 5% of missing data and 0.05 of minor allele frequency. Bold values denote statistical significance at $p < 0.05$.*

sensitivity of their data set to MD and MAF settings so that they can accommodate this uncertainty into their analyses and interpretations of the results.

Given that estimates of gene flow have direct consequences for conservation decision on threatened species (e.g., Flanagan et al., 2018; Bowles et al., 2020), we reflect on our findings to emphasize that there is no 'rule of thumb' in the population genomic era (i.e., a universal set of settings for bioinformatic processing of genomic data). Instead, a sensitivity analysis such as the one applied here should be implemented to confirm how robust inferences might be for any particular study/data set with regards to different settings of MD and MAF (e.g., Catchen et al., 2011, 2013; Eaton, 2014). This methodological recommendation is essential to avoid unintentional biases that may result from applying particular filtering criteria to genomic data (e.g., Huang and Knowles, 2014; Hodel et al., 2017). Below, we discuss the implications of our results not only in light of general decisions regarding data filtering, but also on how our findings specifically

**TABLE 3 |** Maximum-likelihood parentage analysis to assign maternal and paternal identities for 16 *Dinizia jueirana-facao* seedlings (i.e., the parents with the highest and second highest LOD score[1]; maternal and paternal individuals were identified by distance from the seedlings; see methods for details).

| ID – Seedlings | ID – 1st parent | 1st parent LOD score | ID – 2nd parent | 2nd parent LOD score | Pair parent LOD score |
|---|---|---|---|---|---|
| pop1_102 | **pop1_100** | 8,56E + 15* | pop1_086 | 4,19E + 15* | 1,79E + 16* |
| pop1_113 | **pop1_111** | 8,15E + 15* | pop1_114 | 8,86E + 15* | 2,16E + 16* |
| pop1_012 | pop1_001 | 3,41E + 15* | **pop1_008** | 2,40E + 15* | 8,37E + 15* |
| pop1_025 | pop1_001 | 3,21E + 15* | **pop1_026** | 5,42E + 14* | 5,90E + 15* |
| pop1_027 | pop1_024 | -1,84E + 16 | **pop1_026** | 6,42E + 15* | 0.00E + 00 |
| pop1_028 | **pop1_026** | 7,49E + 15* | **pop1_026** | 7,49E + 15* | 1,42E + 16* |
| pop1_032 | pop1_026 | 2,50E + 15* | **pop1_037** | 2,16E + 15* | 5,54E + 15* |
| pop1_034 | **pop1_037** | 4,23E + 15* | **pop1_037** | 4,23E + 15* | 8,23E + 15* |
| pop1_035 | **pop1_037** | 4,52E + 15* | **pop1_037** | 4,52E + 15* | 8,30E + 15* |
| pop1_036 | **pop1_037** | 4,60E + 15* | **pop1_037** | 4,60E + 15* | 5,51E + 15* |
| pop1_038 | **pop1_037** | 4,49E + 15* | **pop1_037** | 4,49E + 15* | 8,18E + 15* |
| pop1_039 | **pop1_001** | 7,58E + 15* | **pop1_001** | 7,58E + 15* | 1,46E + 16* |
| pop1_055 | pop1_004 | -1,83E + 16 | **pop1_082** | 2,00E + 16* | 0.00E + 00 |
| pop1_059 | **pop1_037** | 2,20E + 16* | pop1_004 | -2,12E + 16 | 0.00E + 00 |
| pop1_085 | pop1_083 | -8,31E + 15 | **pop1_086** | 1,98E + 16* | 0.00E + 00 |
| pop1_087 | pop1_037 | -5,04E + 15 | **pop1_086** | 2,51E + 15* | 0.00E + 00 |

*The sampling distribution is shown in **Figure 1**. [1]LOD score is the natural logarithm of the likelihood that the candidate parent is the true parent divided by the likelihood that the candidate parent is not the true parent. *Delta criterion based on 95% confidence threshold from 50,000 simulations. Maternal parents are in bold.*

can be useful for conservation strategies in the rare and critically endangered plant species *D. jueirana-facao*.

## Choosing MD and MAF Settings to Assess SGS and Parentage Analysis

The direct and indirect gene flow estimates in the *D. jueirana-facao* population showed little sensitivity to variations in MD and MAF. However, when more loci were used, we observed a decrease in the standard error for direct and indirect gene flow estimates (e.g., relatedness; **Supplementary Table 1**), indicating that the data set with more loci (i.e., more missing data permitted) is better suited to obtain population genetic parameters. Nevertheless, in all analyses we chose to use the data set with 5% MD and 0.05 MAF (3,029 SNPs) instead of 20% MD and 0.05 MAF (6,898 SNPs). This is due to the fact that estimates of inbreeding – a genetic parameter directly associated with estimates of relatedness (Cockerham, 1966) – were unbiased using the 0 or 5% MD data set (**Supplementary Table 3**), and the 5% MD and 0.05 MAF data set offered a greater number of loci than the 0% MD and 0.05 MAF data set.

## Guidelines, but Not a Rule of Thumb

Anonymous sequencing methods, such as the reduced representation of genomes protocols [e.g., Genotyping-By-Sequencing (GBS), restriction site-associated DNA sequencing (RAD-seq), double-digest RAD-seq (ddRADseq), and Complexity Reduction of Polymorphic Sequences (CRoPS), ezRAD, and CUTseq; see Andrews et al., 2016; Zhang et al., 2019 for more details], are commonly used to identify biallelic SNPs for a broad range of questions in a diversity of taxa. With this increased application of SNPs, there has been increasing attention being paid to the downstream effects of bioinformatic data processing (e.g., Huang and Knowles, 2014; Paris et al., 2017; Cumer et al., 2021). For example, artifacts generated during bioinformatic processing (or even during genomic library construction) can result in misleading biological conclusions (e.g., O'Leary et al., 2018; Larson et al., 2020), such as incorrect inferences about the geographic structure of genetic variation because of the settings used to filter genomic data (see Larson et al., 2020).

Both the molecular technology used to produce genetic data (e.g., RADseq, where the number of individuals used and size selection of fragments used to create the library, in conjunction with genome size of the species, influence the coverage and missing data), and the bioinformatic settings used to process the genomic data determines the properties of a data set (e.g., the amount of, and which data, are retained for analysis). For example, to estimate intra- and interpopulation genetic parameters, the MAF filter set as low as 1–5% (Rochette and Catchen, 2017) will help ensure that alleles will be found within each population. However, the MAF depends on the sample size, which means that even with a low setting for the MAF filter, inferences may not be robust (e.g., with five individuals sampled per locality, the MAF in the population is 10%). The evolutionary history of the species itself can also impact how many and which loci are retained when filtering the data. For example, the level of divergence among the sampled individuals is a reflection of the evolutionary context (e.g., persistent stable populations versus expanding ones). Likewise, the frequencies of SNPs vary as a function of the demographic history of the species (or populations; e.g., a population expansion versus subdivided population structure).

The desirable properties of a retained data set depends upon the application of the genomic data and the assumptions of the particular methods that will be used to analyze the data (Flanagan et al., 2018). For example, some applications of

genomic data require especially high confidence in the calls of SNPs, whereas others do not (e.g., genome-wide association studies versus phylogenetic inference, respectively) because of differences in the relative sensitivity of an inference to errors in SNP-calling. Likewise, some analytical methods require no missing data. Other methods can accommodate missing data, but these methods vary in how more or less robust they are to missing data. Consequently, the inherent properties of data sets will be unique to each study, as will the level of uncertainty that can be accommodated for accurate inference. All of this means that there are no rules of thumb (i.e., a general suite of settings) for bioinformatic processing of genomic data. Instead, the filtering of genomic data will require data set specific settings, and depending on the analytical method being applied, different data sets may need to be generated from the same genomic data for any one particular study (e.g., Resende-Moreira et al., 2019; Massatti and Knowles, 2020; Marske et al., 2020).

All these considerations for exploring the sensitivity of inferences to the properties of data sets that are either intrinsic to the species evolutionary history itself, or are shaped by the bioinformatic processing of the genomic data, apply to the empirical data of *D. jueirana-facao* for inferences about its dispersal. For example, the sensitivity analyses indicated that the inbreeding coefficient was more sensitive to the percentage of MD than either the direct or indirect estimates of gene flow. As such, applying the arbitrary cutoff of MD (e.g., 20%) that is often applied (e.g., Catchen et al., 2013; Kang et al., 2017; Wyngaarden et al., 2017; Flanagan et al., 2018; Ríos et al., 2020; Soghigian et al., 2020), or even advised (see Catchen et al., 2013), would generate spurious results. Moreover, these different estimates would also change the interpretation and conclusions we might make about *D. jueirana-facao*. For instance, the average inbreeding coefficient in adult trees was statistically significant for a range of percent MD and ranged from –0.065 (for 0% MD) to 0.005 (for 20% MD). Because the interpretation of these statistic shifts when the inbreeding coefficient is negative or positive, an arbitrarily chosen 20% MD would have indicated a low, but significant, level of inbreeding in the adult trees of *D. jueirana-facao*. Does this mean that 20% should not be used in other studies? Absolutely not. Although the use of the full data set (i.e., no missing data) instead of 20% MD provided an unbiased estimate of the level of inbreeding in the trees of *D. jueirana-facao*, the use of 20% MD for other data sets may be a good setting, maximizing the number of loci without biasing the results (e.g., Hovmöller et al., 2013).

We observed that the sensitivity to MD varies among the summary statistics, as it does among studies. For example, Hodel et al. (2017) found that some, but not all, genetic estimates (e.g., $F_{IS}$, $F_{ST}$, and $H_e$) were sensitive to the amount of MD (i.e., varied depending upon the amount of MD) in mangroves. Likewise, comparing simulated with empirical ddRAD data set, Attard et al. (2018) also reported that relatedness estimates, specifically that proposed by Ritland (1996), is robust to MD between 0 and 40%. We also found that there were no detectable effects of MD on the relatedness estimates (**Table 1**); however, our study differs in that Loiselle's kinship, not Ritland's kinship, is less sensitive to the MD, as well as the MAF setting used to filter the genomic data of *D. jueirana-facao*. This difference may relate to the intrinsic properties arising from the demographic history of *D. jueirana-facao* given that Ritland's kinship estimator tends to give downward biased estimates when rare alleles occur (Ritland, 1996; Vekemans and Hardy, 2004).

## Does Different Filtering Setting per Study Confound Comparisons Across Studies?

The comparison across species in their genetic structuring is essential for exploring the generality of evolutionary hypotheses (e.g., the identification of shared effects of climate induced distributional shifts (e.g., Knowles et al., 2016; Myers et al., 2019). However, the different criteria for bioinformatic processing of genomic data across studies does not compromise comparisons across studies. As discussed above, standardizing these setting would introduce biases of varying degrees across studies; if the inferences for individuals studies are biased, there is no reasonable argument that standardization could improve the accuracy of conclusions drawn from comparing those studies. In addition, the bias introduced by applying a single standard across species would obscure any effort to quantify the uncertainty in estimating parameters or testing hypotheses of interspecific similarities or dissimilarities (e.g., Andrews et al., 2018; Díaz-Arce and Rodríguez-Ezpeleta, 2019).

Likewise, the view that since more stringent settings will reduce SNP-calling errors, these settings are desirable as being "more conservative" is inaccurate. For example, because the amount of data impacts both the accuracy and error associated with parameter estimates (e.g., Arnold et al., 2013; Marandel et al., 2020), the loss of information when applying strict filtering criteria that severely reduce the number of SNPs will not be outweighed by reducing potential SNP-calling errors (e.g., Mastretta-Yanes et al., 2015; Paris et al., 2017; Díaz-Arce and Rodríguez-Ezpeleta, 2019; Cumer et al., 2021).

## A Small, but Not Isolated Population and Implication for Conservation Management

The impact of the specific history of *D. jueirana-facao* – that is, a small plant population that resulted from habitat loss – appears to affect how generalizable previous suggestions about MD and MAF might be (e.g., Anderson and Garza, 2006; Baruch and Weller, 2008; Strucken et al., 2016; Andrews et al., 2018). Other studies have similarly documented different degrees of sensitivity, and they were not all based on small populations (e.g., Chattopadhyay et al., 2014; Hodel et al., 2017; Attard et al., 2018; Dussault and Boulding, 2018; O'Connell and Smith, 2018; Crotti et al., 2019; Díaz-Arce and Rodríguez-Ezpeleta, 2019; Larson et al., 2020; Cumer et al., 2021). Together, this reinforces that any general guideline for bioinformatic settings still need to be examined with sensitivity analyses on a case-by-case basis. Such study-specific and/or data set specific settings

therefore also become an important part of investigating the crises many species face due to habitat loss and shrinking population sizes.

Our findings demonstrate the robustness of gene flow estimates, as well as the sensitivity of some summary statistics, and provide essential information about the uncertainty arising from the settings of MD and MAF used in the bioinformatic processing of the genomic data for *D. jueirana-facao*. This information for this small, fragmented population is vital to avoiding biased inferences about gene flow that inform conservation and management policies of *D. jueirana-facao*. In particular, with contemporary estimates of gene dispersal distance ($\sigma_{rt}$ = 41.8 m) ∼ fourfold lower than the historical estimates, the genetic consequences of the recent restriction in the scale of gene flow identifies the magnitude of the threat posed by forest fragmentation and loss of habitat in *D. jueirana-facao*.

The response of different plant species with specific pollination or seed dispersal syndromes may vary when faced with anthropogenic disturbances (e.g., Bacles and Jump, 2011; Hardy et al., 2019). However, with respect to tree species, they exhibit a trend of reduced contemporary gene flow, even if they differ in their respective dispersal syndromes (e.g., Oddou-Muratorio and Klein, 2008; Guidugli et al., 2016; but see Bacles et al., 2005). However, the magnitude of the decrease of contemporary gene flow varied among tree species. For instance, a contemporary estimate of gene dispersal distance was twofold smaller than the historical estimates for the insect-pollinated and animal-dispersed tree species *Sorbus torminalis* (Oddou-Muratorio and Klein, 2008), whereas in the wind-dispersed tree species *Entandrophragma cylindricum* the reduction in contemporary gene flow was almost threefold compared with historical estimates of gene flow (Monthe et al., 2017).

With respect to the dispersal distance, contemporary versus historical measures differ in *D. jueirana-facao*. However, we note that realized gene dispersal distances are markedly higher for pollen than seeds, which is consistent with other studies (e.g., Oddou-Muratorio and Klein, 2008; Berens et al., 2013; Guidugli et al., 2016; Hardy et al., 2019). The effectiveness of pollen transport may be a key contributor to the resilience of *D. jueirana-facao* to losses related to anthropogenic threats, given that selfing rates estimated from parentage analysis were not exceedingly high. Predominantly outcrossing tree species such as *D. jueirana-facao*, like other species, are expected to show pollen dispersal over long distances (e.g., Bacles et al., 2005; Nazareno and Carvalho, 2009; Guidugli et al., 2016; Hardy et al., 2019). Indeed, even in the highly fragmented landscape that *D. jueirana-facao* inhabits, we observed a moderate frequency (31.25%) of gene immigration, indicating that pollen movement beyond the edges of the small fragment may reach distances of 12 km (i.e., there is long-distance pollen dispersal between the forest fragment and the nearest pollen source). Gene flow by pollen dispersal beyond the edges of seemingly isolated forest fragments has been documented for distinct tree species, including species that are animal-pollinated (e.g., Nason and Hamrick, 1997; Sato et al., 2006; Nazareno and Carvalho, 2009; Ottewell et al., 2012; Côrtes et al., 2013; Saro et al., 2014; Guidugli et al., 2016; Garcia et al., 2019; Skogen et al., 2019; Lompo et al., 2020). Our result suggests that the open landscape due to deforestation, where the small population of *D. jueirana-facao* are located, facilitates pollen flow and may ameliorate the expected detrimental genetic effects of forest fragmentation. In fact, the inbreeding rate in the seedlings of *D. jueirana-facao* was close to zero. However, the sustainability of the small number of individuals of the species in the long-term is unclear, given that the maintenance of gene flow depends on the preservation of very small populations of *D. jueirana-facao* that reside in forest remnants that are highly fragmented across the landscape.

Our estimates of indirect dispersal distance also provide direct practical guidance for the conservation of *D. jueirana-facao*. For example, our genomic study suggests that efforts toward managed reseeding programs should focus on collecting seeds for breeding, conservation, and restoration programs from reproductive trees separated by at least 100 m. This finding, along with the maximum estimate of direct gene dispersal observed within the population (∼275.0 m), should be taken into account in management strategies to promote more favorable conditions for the establishment and retention of new generations of seedlings. Furthermore, we noted that the population of *D. jueirana-facao* received a moderate percentage of long-distance immigrant pollen, indicating that this small population is not genetically isolated. This direct dispersal pattern is relevant for the *in situ* conservation of remaining local populations since gene flow over long distances can enhance and/or increase connectivity between the two remaining fragments of *D. jueirana-facao*. As such, any human activities that would jeopardize the connectivity between fragments, and in essence exacerbating the negative genetic effects of small, isolated populations (e.g., Spielman et al., 2004), would place the viability of forest remnants in immediate peril. Several strategies have proven to increase connectivity between fragments, notably the establishment of corridors along forest patches (e.g., Rosot et al., 2018), and increasing the porosity of the matrix (e.g., Rodrigues et al., 2009; Rubio and Saura, 2012). Such measures must be informed by empirical data in which the gene flow capacity of species is measured directly on the particular landscape associated with the species of interest to avoid applying generalities to manage fragmented plant populations, when these populations exhibit species-specific responses, as well as species-specific means for alleviating the negative genetic consequences of population loss and fragmentation.

## DATA AVAILABILITY STATEMENT

SNP data sets for *D. jueirana-facao* are available for download from the Dryad Digital Repository (https://doi.org/10.5061/dryad.0vt4b8h01). The raw data generated for *D. jueirana-facao* are available to download from the ENA (European Nucleotide Archive) under accession number ERP129560.

## AUTHOR CONTRIBUTIONS

AGN and LK designed the study. AGN collected the samples, conducted the molecular work, performed the analyses, and led the writing of the manuscript with input from LK, who also provided analytical support. Both authors contributed to the article and approved the submitted version.

## FUNDING

We thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for funding a one-year visit of AGN to the University of Michigan, Ann Arbor (88887.369570/2019-00), and funding from the University of Michigan to LK. Additional funds were provided by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) through a Pq-2 grant to AGN (306182/2020-3). This research was also sponsored by the Mohamed bin Zayed Species Conservation Fund (project number: 202525142). This study was also financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

## ACKNOWLEDGMENTS

We thank Thaís Martins and Domingos A. Folli for assistance during fieldwork. We also thank Reserva Natural Vale for their support of this research.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.677009/full#supplementary-material

## REFERENCES

Alencar, L. R. V., and Quental, T. B. (2019). Exploring the drivers of population structure across desert snakes can help to link micro and macroevolution. *Mol. Ecol.* 28, 4529–4532. doi: 10.1111/mec.15247

Anderson, E. C., and Garza, J. C. (2006). The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics* 172, 2567–2582. doi: 10.1534/genetics.105.048074

Andrews, K. R., Adams, J. R., Cassirer, E. F., Plowright, R. K., Gardner, C., Dwire, M., et al., (2018). A bioinformatic pipeline for identifying informative SNP panels for parentage assignment from RADseq data. *Mol. Ecol. Resour.* 18, 1263–1281. doi: 10.1111/1755-0998.12910

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., and Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet.* 17, 81–92. doi: 10.1038/nrg.2015.28

Arnold, B., Corbett-Detig, R. B., Hartl, D., and Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Mol. Ecol.* 22, 3179–3190. doi: 10.1111/mec.12276

Attard, C. R. M., Beheregaray, L. B., and Moller, L. M. (2018). Genotyping-by-sequencing for estimating relatedness in nonmodel organisms: avoiding the trap of precise bias. *Mol. Ecol. Resour.* 18, 381–390. doi: 10.1111/1755-0998.12739

Bacles, C. F. E., and Jump, A. S. (2011). Taking a tree's perspective on forest fragmentation genetics. *Trends Plant Sci.* 16, 13–18. doi: 10.1016/j.tplants.2010.10.002

Bacles, C. F. E., Burczyk, J., Lowe, A. J., and Ennos, R. A. (2005). Historical and contemporary mating patterns in remnant populations of the forest tree *Fraxinus* excelsior. *Evolution* 59, 979–990. doi: 10.1554/04-653

Baruch, E., and Weller, J. I. (2008). Estimation of the number of SNP genetic markers required for parentage verification. *Anim. Genet.* 39, 474–479. doi: 10.1111/j.1365-2052.2008.01754.x

Berens, D. G., Griebeler, E. M., Braun, C., Chituyi, B. B., Nathan, R., and Böhning−Gaese, K. (2013). Changes of effective gene dispersal distances by pollen and seeds across successive life stages in a tropical tree. *Oikos* 122, 1616–1625. doi: 10.1111/j.1600-0706.2013.00515.x

Bittencourt, J. V. M., and Sebbenn, A. M. (2007). Patterns of pollen and seed dispersal in a small, fragmented population of the wind-pollinated tree *Araucaria angustifolia* in southern Brazil. *Heredity* 99, 580–591. doi: 10.1038/sj.hdy.6801019

Bowles, E., Marin, K., Mogensen, S., MacLeod, P., and Fraser, D. J. (2020). Size reductions and genomic changes within two generations in wild walleye populations: associated with harvest? *Evol. Appl.* 13, 1128–1144. doi: 10.1111/eva.12987

Burczyk, J., Adams, W. T., and Shimizu, J. Y. (1996). Mating patterns and pollen dispersal in a natural knobcone pine (*Pinus* attenuata Lemmon) stand. *Heredity* 77, 251–260. doi: 10.1038/sj.hdy.6880410

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3 (Bethesda)* 1, 171–182. doi: 10.1534/g3.111.000240

Catchen, J. M., Hohenlohe, P., Bassham, S., Amores, A., and Cresko, W. (2013). Stacks: an analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi: 10.1111/mec.12354

Chattopadhyay, B., Garg, K. M., and Ramakrishnan, U. (2014). Effect of diversity and missing data on genetic assignment with RAD-Seq markers. *BMC Res Notes* 7:841. doi: 10.1186/1756-0500-7-841

Cockerham, C. C. (1966). Group inbreeding and coancestry. *Genetics* 56, 89–104. doi: 10.1093/genetics/56.1.89

Côrtes, M. C., Uriarte, M., Lemes, M. R., Gribel, R., Kress, W. J., Smouse, P. E., et al., (2013). Low plant density enhances gene dispersal in the Amazonian understory herb *Heliconia acuminata*. *Mol. Ecol.* 22, 5716–5729. doi: 10.1111/mec.12495

Crawford, T. J. (1984). "What is a population?," in *Evolutionary Ecology*, ed. B. Shorrocks (Oxford: Blackwell Scientific Publications), 135–173.

Crotti, M., Barratt, C. D., Loader, S. P., Gower, D. J., and Streicher, J. W. (2019). Causes and analytical impacts on missing data in RADseq phylogenetics: insights from an African frog (*Afrixalus*). *Zool. Scripta* 48, 157–167. doi: 10.1111/zsc.12335

Cumer, T., Pouchon, C., Boyer, F., Yannic, G., Rioux, D., Bonin, A., et al., (2021). Double-digest RAD-sequencing: do pre- and post-sequencing protocol parameters impact biological results? *Mol. Genet. Genom.* 296, 457–471. doi: 10.1007/s00438-020-01756-9

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012

de Fraga, R., Lima, A. P., Magnusson, W. E., Ferrão, M., and Stow, A. J. (2017). Contrasting patterns of gene flow for Amazonian snakes that actively forage and those that wait in ambush. *J. Heredity* 108, 524–534. doi: 10.1093/jhered/esx051

de Oliveira, S. S., Campos, T., Sebbenn, A. M., and d'Oliveira, M. V. N. (2020). Using spatial genetic structure of a population of *Swietenia macrophylla* king to integrate genetic diversity into management strategies in Southwestern Amazon. *Forest Ecol. Manag.* 464:118040. doi: 10.1016/j.foreco.2020.118040

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., et al., (2011). A framework for variation discovery and genotyping using next−generation DNA sequencing data. *Nat. Genet.* 43, 491–501.

Díaz-Arce, N., and Rodríguez-Ezpeleta, N. (2019). Selecting RAD-Seq data analysis parameters for population genetics: the more the better? *Front. Genet.* 10:533. doi: 10.3389/fgene.2019.00533

Dow, B. D., and Ashley, M. V. (1996). Microsatellite analysis of seed dispersal and parentage of sampling in bur oak, *Quercus macrocarpa*. *Mol. Ecol.* 5, 615–627. doi: 10.1111/j.1365-294x.1996.tb00357.x

Dutech, C., Sork, V. L., Irwin, A. J., Smouse, P. E., amd Davis, F. W. (2005). Gene flow and fine-scale genetic structure in a wind-pollinated tree species, *Quercus lobate* (Fagaceae). *Am. J. Bot.* 92, 252–261. doi: 10.3732/ajb.92.2.252

Dussault, F. M., and Boulding, E. G. (2018). Effects of minor allele frequency on the number of single nucleotide polymorphisms needed for accurate parentage assignment: a methodology illustrated using Atlantic salmon. *Aquac. Res.* 49, 1368–1372. doi: 10.1111/are.13566

Eaton, D. A. R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30, 1844–1849. doi: 10.1093/bioinformatics/btu121

Escoda, L., González−Esteban, J., Gómez, A., and Castresana, J. (2017). Using relatedness networks to infer contemporary dispersal: application to the endangered mammal *Galemys pyrenaicus*. *Mol. Ecol.* 26, 3343–3357. doi: 10.1111/mec.14133

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Feres, J. M., Nazareno, A. G., Borges, L. M., Guiducli, M. C., Bonifacio-Anacleto, F., and Alzate-Marin, A. L. (2021). Depicting the mating system and patterns of contemporary pollen flow in trees of the genus *Anadenanthera* (Fabaceae). *PeerJ* 9:e10579. doi: 10.7717/peerj.10579

Flanagan, S. P., Forester, B. R., Latch, E. K., Aitken, S. N., and Hoban, S. (2018). Guidelines for planning genomic assessment and monitoring locally adaptive variation to inform species conservation. *Evol. Appl.* 11, 1035–1052. doi: 10.1111/eva.12569

Garcia, A. S., Bressan, E. A., Ballester, M. V. R., Figueira, A., and Sebbenn, A. M. (2019). High rates of pollen and seed flow in *Hymenaeae* stignocarpa on a highly fragmented savanna landscape in Brazil. *New For.* 50, 991–1006. doi: 10.1007/s11056-019-09710-3

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., et al., (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Mol. Ecol.* 22, 3165–3178. doi: 10.1111/mec.12089

Goudet, J., Kay, T., and Weir, B. S. (2018). How to estimate kinship. *Mol. Ecol.* 27, 4121–4135. doi: 10.1111/mec.14833

Guidugli, M. C., Nazareno, A. G., Feres, J. M., Contel, E. P. B., Mestriner, M. A., and Alzate-Marin, A. L. (2016). Small but not isolated: a population genetic survey of the tropical tree *Cariniana estrellensis* (Lecythidaceae) in a highly fragmented habitat. *Heredity* 116, 339–347. doi: 10.1038/hdy.2015.108

Hall, D., Zhao, W., Wennstrom, U., Gull, B. A., and Wang, X.-R. (2020). Parentage and relatedness reconstruction in *Pinus sylvestris* using genotyping-by-sequencing. *Heredity* 124, 633–646. doi: 10.1038/s41437-020-0302-3

Hardy, O. J., and Vekemans, X. (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* 83, 145–154. doi: 10.1046/j.1365-2540.1999.00558.x

Hardy, O. J., and Vekemans, X. (2002). SPAGeDi: a versatile computer pro- gram to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* 2, 618–620. doi: 10.1046/j.1471-8286.2002.00305.x

Hardy, O. J., Delaide, B., Hainaut, H., Gillet, J.-F., Gillet, P., Kaymak, E., et al., (2019). Seed and pollen dispersal distances in two African legume timber trees and their reproductive potential under selective logging. *Mol. Ecol.* 28, 3119–3134. doi: 10.1111/mec.15138

Hardy, O. J., Maggia, L., Bandou, E., Caron, H., Chevallier, M. H., Doligez, A., et al., (2006). Fine-scale genetic structure and gene dispersal inferences in 10 Neotropical tree species. *Mol. Ecol.* 15, 559–571. doi: 10.1111/j.1365-294x.2005.02785.x

Hellmann, J. K., Sovic, M. G., Gibbs, H. L., Reddon, A. R., O'Connor, C. M., Ligocki, I. Y., et al., (2016). Within−group relatedness is correlated with colony−level social structure and reproductive sharing in a social fish. *Mol. Ecol.* 25, 4001–4013. doi: 10.1111/mec.13728

Hodel, R. G. J., Chen, S., Payton, A. C., McDaniel, S. F., Soltis, P., and Soltis, D. E. (2017). Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: comparing microsatellites and RAD-Seq and investigating loci filtering. *Sci. Rep.* 7:17598.

Hohenlohe, P., Amish, S., Catchen, J., Allendorf, F., and Luikart, G. (2011). RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow trout and westslope cutthroat trout. *Mol. Ecol. Resour.* 11, 117–122. doi: 10.1111/j.1755-0998.2010.02967.x

Hovmöller, R., Kubatko, L. S., and Knowles, L. L. (2013). Effects of missing data on species tree estimation under the coalescent. *Mol. Phylogenet. Evol.* 69, 1057–1062. doi: 10.1016/j.ympev.2013.06.004

Huang, H., and Knowles, L. L. (2014). Unforeseen consequences of excluding missing data from Next-Generation Sequences: simulation study of RAD sequences. *Syst. Biol.* 65, 357–365. doi: 10.1093/sysbio/syu046

Ilut, D. C., Nydam, M. L., and Hare, M. P. (2014). Defining loci in restriction-based reduced representation genomic data from nonmodel species: sources of bias and diagnostics for optimal clustering. *BioMed Res. Int.* 2014:675158.

Jombart, T. (2008). Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi: 10.1093/bioinformatics/btn129

Jombart, T., and Ahmed, I. (2011). Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi: 10.1093/bioinformatics/btr521

Kalinowski, S. T., Taper, M. L., and Marshall, T. C. (2007). Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. *Mol. Ecol.* 16, 1099–1106. doi: 10.1111/j.1365-294x.2007.03089.x

Kang, J., Ma, X., and He, S. (2017). Population genetics analysis of the Nujiang catfish Creteuchiloglanis macropterus through a genome-wide single nucleotide polymorphisms resource generated by RAD-Seq. *Sci. Rep.* 7:2813.

Knowles, L. L., Massatti, R., He, Q., Olson, L. E., and Lanier, H. C. (2016). Quantifying the similarity between genes and geography across Alaska's alpine small mammals. *J. Biogeogr.* 43, 1464–1476. doi: 10.1111/jbi.12728

Larson, W. A., Isermann, D. A., and Feiner, Z. S. (2020). Incomplete bioinformatic filtering and inadequate age and growth analysis lead to an incorrect inference of harvested-induced changes. *Evol. Appl.* 14, 278–289. doi: 10.1111/eva.13122

Lewis, G. P., Siqueira, G. S., Banks, H., and Bruneau, A. (2017). The majestic canopy-emergent genus *Dinizia* (Leguminosae: Caesalpinioideae), including a new species endemic to the Brazilian state of Espírito Santo. *Kew Bull.* 72:48.

Loiselle, B. A., Sork, V. L., Nason, J., and Graham, C. (1995). Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am. J. Bot.* 82, 1420–1425. doi: 10.1002/j.1537-2197.1995.tb12679.x

Lompo, D., Vicenti, B., Konrad, H., Duminil, J., and Geburek, T. (2020). Fine-scale spatial genetic structure, mating, and gene dispersal patterns in *Parkia biglobosa* populations with different levels of habitat fragmentation. *Am. J. Bot.* 107, 1041–1053. doi: 10.1002/ajb2.1504

Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4, 981–999. doi: 10.1038/nrg1226

Malécot, G. (1948). *Les mathématiques de l'hérédité mendelienne généralisée. [Chap. 3 in Malécot (1966)*.

Marandel, F., Charrier, G., Lamy, J.-B., Le Cam, S., Lorance, P., and Trenkel, V. M. (2020). Estimating effective population size using RADseq: effects of SNP selection and sample size. *Ecol. Evol.* 10, 1929–1937. doi: 10.1002/ece3.6016

Marshall, T. C., Slate, J., Kruuk, L. E. B., and Pemberton, J. M. (1998). Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* 7, 639–655. doi: 10.1046/j.1365-294x.1998.00374.x

Marske, K. A., Thomaz, A. T., and Knowles, L. L. (2020). Dispersal barriers and opportunities drive multiple levels of phylogeographic concordance in the Southern Alps of New Zealand. *Mol. Ecol.* 29, 4665–4679. doi: 10.1111/mec.15655

Massatti, R., and Knowles, L. L. (2020). The historical context of contemporary climatic adaptation: a case study in the climatically dynamic and environmentally complex southwestern United States. *Ecography* 43, 735–746. doi: 10.1111/ecog.04840

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., and Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Mol. Ecol. Resour.* 15, 28–41. doi: 10.1111/1755-0998.12291

McCartney-Melstad, E., Vu, J. K., and Shaffer, H. B. (2018). Genomic data recover previously undetectable fragmentation effects in an endangered amphibian. *Mol. Ecol.* 27, 4430–4443. doi: 10.1111/mec.14892

Monthe, F. K., Hardy, O. J., Doucet, J.-L., Loo, J., and Duminil, J. (2017). Extensive seed and pollen dispersal and assortative mating in the rain forest tree *Entandrophragma cylindricum* (Meliaceae) inferred from indirect and direct analyses. *Mol. Ecol.* 26, 5279–5291. doi: 10.1111/mec.14241

Myers, E. A., Xue, A. T., Gehara, M., Christian, L. C., Rabosky, A. R. D., Lemos-Espinal, J., et al., (2019). Environmental heterogeneity and not vicariant biogeographic barriers generate community-wide population structure in desert-adapted snakes. *Mol. Ecol.* 28, 4535–4548. doi: 10.1111/mec.15182

Nason, J. D., and Hamrick, J. L. (1997). Reproductive and genetic consequences of forest fragmentation: two case studies of Neotropical canopy trees. *J. Heredity* 88, 264–276. doi: 10.1093/oxfordjournals.jhered.a023104

Nazareno, A. G., and Carvalho, D. (2009). What the reasons for no inbreeding and high genetic diversity of the Neotropical fig tree *Ficus arpazusa*? *Conserv. Genet.* 10, 1789–1793. doi: 10.1007/s10592-008-9776-x

Nazareno, A. G., Bemmels, J. B., Dick, C., and Lohmann, L. G. (2017). Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Mol. Ecol.* 17, 1136–1147. doi: 10.1111/1755-0998.12654

O'Connell, K. A., and Smith, E. N. (2018). The effect of missing data on coalescent species delimitation and a taxonomic revision of whipsnakes (Colubridae: Masticophis). *Mol. Phylogenet. Evol.* 127, 356–366. doi: 10.1016/j.ympev.2018.03.018

O'Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., and Portnoy, D. S. (2018). These aren't the loci you'e looking for: principles of effective SNP filtering for molecular ecologists. *Mol. Ecol.* 27, 3193–3206. doi: 10.1111/mec.14792

Oddou-Muratorio, S., and Klein, E. K. (2008). Comparing direct vs. indirect estimates of gene flow within a population of a scattered tree species. *Mol. Ecol.* 17, 2743–2754. doi: 10.1111/j.1365-294x.2008.03783.x

Ottewell, K., Grey, E., Castillo, F., and Karubian, J. (2012). The pollen dispersal kernel and mating system of an insect-pollinated tropical palm, *Oenocarpus bataua*. *Heredity* 109, 332–339. doi: 10.1038/hdy.2012.40

Paris, J. R., Stevens, J. R., and Catchen, J. M. (2017). Lost in parameter space: a road map for STACKS. *Methods Ecol. Evol.* 8, 1360–1373. doi: 10.1111/2041-210x.12775

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* 7:e37135. doi: 10.1371/journal.pone.0037135

Queller, D. C., and Goodnight, K. F. (1989). Estimating relatedness using genetic markers. *Evolution* 43, 258–275. doi: 10.2307/2409206

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramos, S. L. F., Dequigiovanni, G., Sebbenn, A. M., Lopes, M. T. G., and Vasconcelos, J. L. (2018). Paternity analysis, pollen flow, and spatial genetic structure of a natural population of Euterpe precatoria in the Brazilian Amazon. *Ecol. Evol.* 8, 11143–11157. doi: 10.1002/ece3.4582

Resende-Moreira, L. C., Knowles, L. L., Thomas, A. T., Prado, J. R., Souto, A. P., Lemos-Filho, J. P., et al., (2019). Evolving in isolation: genetic tests reject recent connections of Amazonian savannas with the central Cerrado. *J. Biogeogr.* 46, 196–211. doi: 10.1111/jbi.13468

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution* 43, 223–225. doi: 10.1111/j.1558-5646.1989.tb04220.x

Ríos, N., Casanova, A., Hermida, M., Pardo, B. G., Martínez, P., Bouza, C., et al., (2020). Population genomics in *Rhamdia* quelen (Heptapteridae, Siluriformes) reveals deep divergence and adaptation in the Neotropical region. *Genes* 11:109. doi: 10.3390/genes11010109

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet. Res.* 67, 175–185. doi: 10.1017/s0016672300033620

Rochette, N. C., and Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using Stacks. *Nat. Protoc.* 12, 2640–2659. doi: 10.1038/nprot.2017.123

Rochette, N. C., Rivera-Colón, A., and Catchen, J. (2019). Stacks 2: analytical methods for paired-end sequencing improve RADseq-based population genomics. *Mol. Ecol.* 28, 4737–4754. doi: 10.1111/mec.15253

Rodrigues, R. R., Lima, R. A. F., Gandolfi, S., and Nave, A. G. (2009). On the restoration of high diversity forests: 30 years of experience in the Brazilian Atlantic Forest. *Biol. Conserv.* 142, 1242–1251. doi: 10.1016/j.biocon.2008.12.008

Rosot, M. A. D., Maran, J. C., Luz, N. B., Garrastazú, M. C., Oliveira, Y. M. M., Franciscon, L., et al., (2018). Riparian forest corridors: a prioritization analysis to the landscape sample units of the Brazilian National Forest Inventory. *Ecol. Indic.* 93, 501–511. doi: 10.1016/j.ecolind.2018.03.071

Rubio, L., and Saura, S. (2012). Assessing the importance of individual habitat fragments as irreplaceable connecting elements: an analysis of simulated and real landscape data. *Ecol. Complex.* 11, 28–37. doi: 10.1016/j.ecocom.2012.01.003

Saro, I., Robledo-Arnuncio, J. J., Gonzáles-Pérez, M. A., and Sosa, P. A. (2014). Patterns of pollen dispersal in a small population of the Canarian endemic palm (*Phoenix canariensis*). *Heredity* 113, 215–223. doi: 10.1038/hdy.2014.16

Sato, T., Isagi, Y., Sakio, H., Osumi, K., and Goto, S. (2006). Effect of gene flow on spatial genetic structure in the riparian canopy tree *Cercidiphyllum japonicum* revealed by microsatellite analysis. *Heredity* 96, 79–84. doi: 10.1038/sj.hdy.6800748

Seeb, J. E., Carvalho, G., Hauser, L., Naish, K., Roberts, S., and Seeb, L. W. (2011). Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol. Ecol. Resour.* 11, 1–8. doi: 10.1111/j.1755-0998.2010.02979.x

Skogen, K. A., Overson, R. P., Hilpman, E. T., and Fant, J. B. (2019). Hawkmoth pollination facilitates long-distance pollen dispersal and reduces isolation across a gradient of land-use change. *Ann. Mo. Bot. Garden* 104, 495–511. doi: 10.3417/2019475

Soghigian, J., Gloria-Soria, A., Robert, V., Le Goff, G., Failloux, A.-B., and Powell, J. R. (2020). Genetic evidence for the origin of Aedes aegypti, the yellow fever mosquito, in the southwestern Indian Ocean. *Mol. Ecol.* 29, 3593–3606. doi: 10.1111/mec.15590

Spielman, D., Brook, B. W., and Frankham, R. (2004). Most species are not driven to extinction before genetic factors impact them. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15261–15264. doi: 10.1073/pnas.0403809101

Strucken, E. M., Lee, S. H., Lee, H. K., Song, K. D., Gibson, J. P., and Gondro, C. (2016). How many markers are enough? Factors influencing parentage testing in different livestock populations. *J. Anim. Breed. Genet.* 133, 13–23. doi: 10.1111/jbg.12179

Titus, V. R., Bell, R. C., Becker, C. G., and Zamudio, K. R. (2014). Connectivity and gene flow among Eastern Tiger Salamander (*Ambystoma tigrinum*) populations in highly modified anthropogenic landscapes. *Conserv. Genet.* 15, 1447–1462. doi: 10.1007/s10592-014-0629-5

Vekemans, X., and Hardy, O. J. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Mol. Ecol.* 13, 921–935. doi: 10.1046/j.1365-294x.2004.02076.x

Weir, B. S., and Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics* 206, 2085–2103. doi: 10.1534/genetics.116.198424

Willis, S. C., Hollenbeck, C. M., Puritz, J. B., Gold, J. R., and Portnoy, D. S. (2017). Haplotyping RAD loci: an efficient method to filter paralogs and account for physical linkage. *Mol. Ecol. Resour.* 17, 955–965. doi: 10.1111/1755-0998.12647

Wyngaarden, M. V., Snelgrove, P. V. R., DiBacco, C., Hamilton, L. C., Rodríguez-Ezpeleta, N., Jeffery, N. W., et al., (2017). Identifying patterns of dispersal, connectivity and selection in the sea scallop, Placopecten magellanicus, using RADseq-derived SNPs. *Evol. Appl.* 10, 102–117. doi: 10.1111/eva.12432

Zhang, X., Garnerone, S., Simonetti, M., Harbers, L., Nicos, M., Mirzazadeh, R., et al., (2019). CUTseq is a versatile method for preparing multiplexed DNA sequencing libraries from low-input samples. *Nat. Commun.* 10:4732.

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.