bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Habitat loss does not always entail negative genetic consequences

Carolina S. Carvalho[a§], Éder C. M. Lanes[a§], Amanda R. Silva[a,b], Cecilio F. Caldeira[a], Nelson Carvalho-Filho[a], Markus Gastauer[a], Vera L. Imperatriz-Fonseca[a], Wilson Nascimento Júnior[a], Guilherme Oliveira[a], José O. Siqueira[a], Pedro L. Viana[b], Rodolfo Jaffé[a,c*]

[a] Instituto Tecnológico Vale, Rua Boaventura da Silva 955, 66055-090 Belém-PA, Brazil.

[b] Museu Paraense Emílio Goeldi, Av. Perimetral 1901, Terra Firme, 66077-530, Belém-PA, Brazil.

[c] Departamento de Ecologia, Universidade de São Paulo, Rua do Matão 321, trav 14, 05508-900 São Paulo-SP, Brazil.

[§] Both authors contributed equally

[*] **Corresponding author:** Rodolfo Jaffé - Instituto Tecnológico Vale, Rua Boaventura da Silva 955, 66055-090 Belém-PA, Brazil. Phone number: +55 (91) 3213-5523, Email: rodolfo.jaffe@itv.org

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Although habitat loss has large, consistently negative effects on biodiversity, its genetic consequences are not yet fully understood. In this paper, we assess the genetic consequences of extreme habitat loss driven by mining in two endemic plants from Amazonian Savannas. Our analyses are the first to overcome major methodological limitations like the confounding effect of habitat fragmentation, historical processes underpinning genetic differentiation, time-lags between the onset of disturbances and genetic outcomes, and the need for large numbers of samples, genetic markers and replicated landscapes to ensure sufficient statistical power. We found that both species are remarkably resilient, as genetic diversity and gene flow patterns were unaffected by habitat loss. Our study unambiguously demonstrates that it is not possible to generalize about the genetic consequences of habitat loss, and imply that future conservation efforts need to consider species-specific genetic information.

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In spite of ample evidence showing that habitat loss has large, consistently negative effects on biodiversity [1], very few studies have assessed the consequences of habitat amount on genetic variation [2]. Habitat loss can potentially impact the demographics of natural populations, reducing population size, migration, gene flow and genetic diversity, and thereby increasing inbreeding and extinction risk [3]. Understanding the genetic consequences of habitat loss is therefore essential to safeguard biological diversity and fulfill Aichi Biodiversity Targets and Sustainable Development Goals [4].

Important limitations constrain the quantification of habitat amount effects on genetic variation. Firstly, habitat loss and fragmentation are often confounded, so disentangling the relative contribution of habitat amount requires controlling for fragmentation [1]. Secondly, landscape effects can also be easily confounded with historical processes and the underlying population structure [5]. Thirdly, a coarse resolution of spatial data and time-lags between the onset of disturbances and genetic responses may mask the effects of recent landscape modification [6,7]. Finally, large numbers of samples and genetic markers, and replicated sampling designs that capture enough landscape heterogeneity are needed to detect or rule out possible landscape effects with sufficient statistical power [8,9]. Failure in overcoming any of these limitations may hide important detrimental effects to the maintenance of genetic variability, or reveal spurious patterns unrelated to habitat loss.

Few studies have attempted to quantify the impact of habitat loss on both genetic diversity and gene flow, and neither has yet accounted for all the methodological limitations outlined above [2,7]. Here we fill this important knowledge gap assessing the genetic consequences of extreme habitat loss driven by open-pit mining in two endemic plants from the Eastern Amazon. Firstly, we were able to assess the independent effect of habitat loss, as open-pit mining in our study region rarely involves habitat fragmentation (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S1). We also accounted for the underlying population structure when assessing landscape effects on genetic diversity and gene flow, assessed the sensitivity of our analyses to the resolution of spatial data, and used annual species (which complete a full reproductive cycle and die within one year) to minimize possible time-lag effects. Finally, we sampled hundreds of individuals scattered across two separate regions exposed to mining, and genotyped them at thousands of single nucleotide polymorphisms (SNPs) distributed across their genomes to assure high statistical power.

Our study species were *Brasilianthus carajensis* (Melastomataceae) and *Monogereion carajensis* (Asteraceae), annual herbs endemic to the Carajás Mineral Province in the Eastern Amazon (Fig. 1). Both species seem to be pollinated by insects, their seeds dispersed by the wind, and exclusively occur in the banded iron formations known as Cangas [10,11] which constitute inselbergs of Amazonian Savannas embedded in an evergreen forest matrix. As Cangas harbor one of the world's largest deposits of high-grade iron ore [12], they have attracted substantial attention from mining companies. In fact two of the world's largest iron-ore mines are located in the region (Fig. 1), with operations in Serra Norte dating back to the 1980s, while Serra Sul only began activities in 2014. We predicted that: i) Individuals surrounded by undisturbed habitats would show higher genetic diversity and lower inbreeding than those exposed to habitat loss driven by mining; ii) Gene flow would be best explained by recent landscape modifications, and mining areas would represent barriers to gene flow.

Results

Neutral dataset

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We collected leaf tissue samples of 150 individuals of *B. carajensis* and 207 individuals of *M. carajensis* distributed across the entire occurrence range of both species and surrounding two large iron ore mines (Fig. 1). Samples were frozen and their DNA later extracted and shipped for genotyping-by-sequencing (RAD-sequencing) and bioinformatic processing. We identified a total of 10,016 SNPs in *B. carajensis* and 20,464 SNPs in *M. carajensis,* but after filtering these for quality, depth, linkage disequilibrium, deviations from the Hardy-Wenberg Equilibrium and F_{ST} outlier loci, we obtained sets of neutral and independent markers containing 1,411 and 6,052 loci for each species respectively.

Genetic structure

Two complementary genetic clustering approaches used to assess population structure (Admixture [13] and Discriminant Analysis of Principal Components - DAPC [14]) indicated the presence of three clusters in *B. carajensis* and two in *M. carajensis* (Fig. 2, Supplementary Figure S2-S4). Significant albeit low inbreeding was found in one genetic cluster of each species (Fig. 2). Both species showed spatial autocorrelation in genetic relatedness in Serra Norte but not in Serra Sul, and the strength of spatial autocorrelation was higher in *B. carajensis* (Fig. 2).

Genetic diversity

To assess the effect of habitat loss on genetic diversity, we regressed individual-level diversity metrics on historical habitat amount (1979) and habitat loss driven by mining in different years (2011, 2014 and 2016). Heterozygosity (H_E) and inbreeding (F) were not influenced by habitat loss, neither in Serra Norte nor in Serra Sul, as the set of best-fitting

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

93 models always included null models or historical (pre-mining) habitat amount (Fig. 3,
94 Supplementary Table S1). Historical habitat amount was found to be associated with
95 inbreeding in both species, although the direction of the effect varied (Fig. 4, Supplementary
96 Table S2).

97 *Gene flow*

98 To assess the effect of habitat loss on gene flow, we first employed a genetic
99 algorithm to optimize gene flow hypotheses [15], then calculated effective resistance between
100 individual samples using Circuitscape V4.0 [16], and finally modeled isolation by resistance
101 (IBR) regressing pairwise genetic relatedness on resistance distances through Maximum
102 Likelihood Population Effects (MLPE) models. Resistance to gene flow due to mining was
103 modeled using land cover maps for different years (2016, 2014, 2011 and 1979). Additional
104 variables found to be important predictors of gene flow in other plants [17,18] were modeled
105 along with land cover, including geographic distance, terrain roughness, elevation, and
106 bioclimatic variables. The optimization of resistance surfaces revealed that Canga was the
107 land cover class representing lowest resistance to gene flow in both species, whereas mining
108 areas and evergreen forests imposed higher resistance (Supplementary Figure S5-S8).
109 However, univariate MLPE regression models revealed that geographic distance usually
110 explained relatedness patterns as well as land cover (Supplementary Table S3), and only pre-
111 mining land cover (1979) was found to explain relatedness patterns better than geographic
112 distance in *M. carajensis* from Serra Norte. Our results thus reveal that mining neither
113 hinders nor facilitates gene flow in these two endemic annual plants. While these results hold
114 across different resolutions (Supplementary Table S3), an independent barrier analysis also
115 failed to identify barriers between individuals separated by mining areas (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S9). Multiple MLPE regression models showed that isolation by geographic distance (IBD) explained genetic relatedness patterns in *B. carajensis*, whereas isolation by resistance (IBR, [16]) was more important in *M. carajensis* (Fig. 3, Supplementary Table S4). In all cases, genetic relatedness decreased with increasing resistance (Fig. 4, Supplementary Table S5).

Germination experiments

Germination experiments revealed that seeds from both species are able to germinate in mining waste substrates. Whereas *M. carajensis* showed similar germination in Canga and mining substrates, germination rates of *B. carajensis* were higher in Canga topsoil (Supplementary Figure S10).

Discussion

Our study is the first to assess the genetic consequences of habitat loss while accounting for all the major limitations constraining the quantification of habitat amount effects on genetic variation. Our results reveal that habitat loss driven by mining did not affect genetic diversity or gene flow in two endemic herbs from Amazonian Savannas. Whereas historical habitat amount was found to influence inbreeding; heterozygosity and inbreeding were not affected by habitat loss in either species, and gene flow was mainly influenced by geographic distance in *B. carajensis* and by pre-mining land cover and local climate in *M. carajensis*.

The genetic structure in *B. carajensis* mirrored that from the co-occurring perennial morning glory *Ipomoea maurandioides* [17], showing two differentiated genetic clusters in Serra Norte, while *M. carajensis* only presented one cluster in Serra Norte and another one

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

137 across the remaining distribution range. This genetic structure was considered when assessing
138 landscape effects on genetic diversity and gene flow, so our results were not biased by
139 historical patterns of genetic differentiation.

140 The maintenance of genetic diversity in spite of extreme habitat loss suggests that our
141 study plants are able to colonize mining environments and maintain gene flow across open-
142 pit mines. Germination experiments revealed that seeds from both species can indeed
143 germinate in mining waste substrates. Additionally, both plant species showed extensive gene
144 flow across mining areas, and mining neither enhanced nor hindered gene flow. Similar
145 results were found for a threatened orchid and the American pika, which showed analogous
146 levels of genetic diversity in mining and natural habitats [19,20], although neither gene flow nor
147 historical effects were assessed. Inbreeding levels in our focus species are comparable to
148 those observed in the widespread *I. maurandioides* [17], and since they were associated with
149 historical habitat amount they seem to reflect density-dependent selfing [21].

150 Both species presented spatial autocorrelation in genetic relatedness in Serra Norte
151 but not in Serra Sul, indicating a more restricted gene flow in the Canga archipelago of Serra
152 Norte than in the large continuous plateau of Serra Sul. Additionally, geographic distance
153 resistance was weakly correlated with recent land cover resistance in Serra Norte but not in
154 Serra Sul, where it was strongly correlated with land cover resistance from all years
155 (Supplementary Figure S11). We thus expected that isolation by resistance (IBR) would be
156 easier to disentangle from isolation by distance (IBD) in Serra Norte than in Serra Sul. In
157 Serra Norte, however, geographic distance and pre-mining land cover (highly correlated with
158 geographic distance) were the best predictors of current gene flow in *B. carajensis* and *M.*
159 *carajensis* populations, respectively. Considering the strong winds characterizing Montane
160 Savanna ecosystems from the Carajás Mineral Province [12], and the fact that wind currents in
161 open landscapes are known to facilitate long-distance dispersal of plant propagules [22,23], we

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

posit that wind-mediated dispersal is driving gene flow across Montane Savannas and open-pit mines. On the other hand, local climate differences also appear to explain gene flow patterns in *M. carajensis* populations from Serra Sul better than IBD, suggesting mismatches in flowering periods [24]. We nevertheless caution that our study design and the little available knowledge on the natural history of these plants do not allow disentangling the relative contribution of pollen and seed dispersal on gene flow.

The absence of an effect of habitat loss on genetic variation can be attributed to time-lags between the onset of disturbances and genetic responses [25]. We overcame this methodological limitation by focusing on species with a short generation time (i.e. completing their life cycle within one year), and by explicitly incorporating time scale into our analyses (evaluating land cover maps from different years). Moreover, our isolation by resistance models primarily reflect recent gene flow, as they explicitly account for the underlying population structure and rely on relatedness estimates calculated from thousands of independent and neutral SNPs. Mining operations began in the 1980s in Serra Norte, allowing enough time (~40 generations) to assess genetic responses to mining. On the other hand, Serra Sul was still pristine by 2013, so only three plant generations were exposed to mining before our samples were collected in 2017. This could explain why geographic distance explained relatedness patterns in Serra Sul better than land cover (Supplementary Table S3). However, land cover did not explain relatedness patterns in either species in Serra Norte, which clearly shows that gene flow has been maintained across mines. In contrast, land cover in existence two decades ago was found to explain gene flow in a perennial narrow endemic morning glory occurring in Serra Norte [17], indicating that our methods should be sufficient to detect an effect of mining should there be one. Additionally, our findings were unaffected by the resolution of spatial data and were supported by an

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

independent barrier analysis, so they clearly reveal that gene flow in our two annual herbs is unaffected by habitat loss driven by mining.

Conclusions

Using thousands of genetic markers to study two annual endemic plants in replicated landscapes, we found that extreme habitat loss driven by mining did not result in any detectable genetic consequences. Since our results are not biased by the effect of habitat fragmentation, the underlying genetic structure of plant populations, the resolution of spatial data, nor time-lag effects, they unambiguously reveal that habitat loss does not always entail negative genetic consequences. For instance, our study reveals remarkably resilient species to extreme habitat loss. These findings imply that it is not possible to generalize about the genetic consequences of habitat loss, so future conservation efforts need to consider species individually.

Materials and Methods

Sampling, DNA extraction, and genome size estimation

We collected leaf tissue samples of 150 individuals of *B. carajensis* and 207 individuals of *M. carajensis* between February and June 2017 (SISBIO collection permit N. 48272-4). Samples were collected in the main Canga plateaus of our study area, comprising the entire occurrence range of both species, and care was taken to sample individuals at or around iron ore mines (Fig. 1). To ensure high DNA quality and concentration, we preserved *B. carajensis* samples in silica and *M. carajensis* samples in 10 mL of a NaCl-saturated solution of 2% CTAB [26], and stored them at -80 °C until analysis. Total DNA of *B. carajensis* was extracted using a CTAB 2% protocol [27] followed by a DNA purification protocol [28];

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

whereas the DNeasy Plant Mini Kit (Qiagen, EUA) was used for *M. carajensis*. DNA concentration for both species was quantified using the Qubit High SensitivityAssay kit (Invitrogen), and DNA integrity assessed through 1.2% agarose gel electrophoresis. All DNA samples were adjusted to a final concentration of 5 ng/µL in a final volume of 30 µL. We used flow cytometry to estimate genome size in both species. Nuclei were obtained from fresh leaf tissues chopped along with references in general purpose buffer with 1% Triton X-100 and 1% PVP-30 [29]. The whole sample preparation was conducted on ice until the events acquisition on a PI fluorescence mean under a 575/26 bandpass filter. Triplicates of 1000 PI stained nuclei were analyzed under a 488 nm laser on BD FACS Aria II cytometer. The internal standard used was tomato (*Lycopersicon esculentum*; 2C = 1.98 pg, [30]).

RAD sequencing and SNP discovery

DNA samples were shipped to SNPSaurus (http://snpsaurus.com/) for sequencing and bioinformatic analyses. Briefly, nextRAD genotyping-by-sequencing libraries were prepared [31] using Nextera reagent (Illumina, Inc) and considering the estimated genome size of each species (2C DNA content was 508 Mbp in *B. carajensis* and 6,284 Mbp in *M. carajensis*). The nextRAD libraries were then sequenced on an Illumina HiSeq 4000 (University of Oregon). Reads were trimmed using *BBMap tools* (http://sourceforge.net/projects/bbmap/) and a *de novo* reference was created by collecting 10 million reads evenly from the samples, excluding reads that had counts fewer than 5 or more than 700 for *B. carajensis*; and fewer than 6 or more than 1,000 for *M. carajensis*. The remaining loci were then aligned to each other to identify alleles. All reads were mapped to the reference with an alignment identity threshold of 90% using *BBMap,* generating 150bp contigs. Genotype calling was done using *Samtools* and *bcftools* (https://github.com/samtools/samtools), and the resulting set of

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

231 genotypes were filtered to remove loci with a minor allele frequency of less than 3%.
232 Heterozygous loci or loci that contained more than 2 alleles in a sample (suggesting collapsed
233 paralogs) were removed. The absence of artifacts was checked by counting SNPs at each read
234 nucleotide position and determining that SNP number did not increase with reduced base
235 quality at the end of the read. A total of 43,887 contigs were generated for *B. carajensis*
236 (sequencing depth ranged between 18 and 239), and 36,040 for *M. carajensis* (depth ranging
237 between 14 and 246).

238 *Neutral datasets*

239 The R package *r2vcftools* (https://github.com/nspope/r2vcftools) - a wrapper for
240 VCFtools [32] - was used to perform final quality control on the genotype data. We filtered loci
241 for quality (Phred score > 50 both species), read depth (30 – 240 both species), linkage
242 disequilibrium (LD, $r^2 < 0.6$ and $r^2 < 0.4$ for *B. carajensis* and *M. carajensis*, respectively),
243 and strong deviations from the Hardy Weinberg Equilibrium (HWE, $p < 0.0001$ both species).
244 Additionally, we removed any potential loci under selection detected through genome scans,
245 whereby F_{ST} outlier tests were applied after adjusting false discovery rates ($q = 0.05$)
246 according to the distribution of *p*-values [33]. The resulting sets of neutral and independent loci
247 were then used in all subsequent analyses.

248 *Genetic structure*

249 We used two complementary genetic clustering software to assess population
250 structure: Admixture [13] and DAPC from the *adegenet* package [14,34]. For the former analysis,
251 the number of ancestral populations (*k*) was allowed to vary between 1 and 10, and the best *k*

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was chosen based on cross-validation errors [35]. For the second analyses, the number of clusters was assessed using the function "*find.cluster*", which runs successive *k*-means clustering with an increasing number of clusters, and then determined the best-supported number of genetic clusters using the Bayesian Information Criterion (BIC). Considering the ancestry coefficients assigned by Admixture, we then estimated expected heterozygosity (H_E) and inbreeding coefficients (F) for each genetic cluster. Additionally, we assessed fine-scale spatial genetic structure within each genetic cluster by quantifying spatial autocorrelation in Yang's genetic relatedness between pairs of individuals [36]. To do so we used local polynomial fitting (LOESS) of pairwise relatedness and pairwise geographic distance (https://github.com/rojaff/Lplot).

Land cover maps

To account for time-lag effects when assessing the genetic consequences of habitat loss, we built land cover maps for different years (2016, 2014, 2011 and 1979), comprising pre-mining maps (1979). *Landsat* images (spatial resolution of 30 meters in 7 spectral bands) were used for years 1979 and 2011, while 2014 and 2016 maps were generated using *Sentinel* images (spatial resolution of 10 meters in 4 spectral bands). Images were downloaded from the Earth Explorer Server (https://earthexplorer.usgs.gov/), selecting scenes from the month of July to minimize clouds. All images were converted to ground reflectance in percentage using the ATCOR algorithm of the PCI Geomatica 2016 software. The scenes were joined to create a mosaic of the study area and derive the Normalized Difference Vegetation Index - NDVI [37]. We then employed the eCognition 9 software using a Geographic Object-Based Image Analysis (GEOBIA) to classify land cover types. The Multi-resolution classification algorithm was selected, given that it allows obtaining segments with different sizes due to

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

brightness, shape, smoothness and compactness. Montane Savanna (Canga), Water, Forest, Mine, Pastureland and Urban classes were identified.

Genetic diversity

To assess the effect of habitat loss on genetic diversity, we regressed individual-level diversity metrics (H_E and F) on historical habitat amount and habitat loss driven by mining in different years (2011, 2014 and 2016) using high resolution land cover maps (10 x 10m). By so doing we explicitly evaluated the effect of habitat loss accounting for historical habitat amount. Historical habitat amount was calculated by extracting the proportion of Canga habitat in a buffer surrounding each individual using pre-mining maps (1979). Habitat loss in different years was calculated by subtracting habitat amount for a given year from historical habitat amount. To select an optimal buffer size we first ran uni-variate models using habitat amount extracted from the most recent land cover maps (2016) with buffers varying in size between 100 and 900m, and then compared all models using AIC. As habitat amount calculated with the largest buffers (900m) was always among the best models ($\Delta AIC \leq 2$), we chose this buffer size to encompass a greater portion of lost areas (Supplementary Table S6).

In Serra Norte, which comprises an archipelago of Canga plateaus, we fit linear mixed-effect models, using each plateau as a random effect to account for site-specific characteristics and spatial autocorrelation. In the case of *B. carajensis* from Serra Norte, we also included a random effect specifying the genetic cluster containing each individual (see genetic structure results). In Serra Sul, which comprises a single large plateau, we used generalized least-squares models (GLS) fitted with different correlation structures (linear, exponential, Gaussian, and spherical) to explicitly model spatial autocorrelation. The "*weight*" argument was used in some cases to account for heteroscedasticity. Raw F and

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

298 logit-transformed H_E were used as response variables and models fitted using the *nmle* R
299 package [38]. The set of best models ($\Delta AIC \leq 2$) were compared to reduced models without
300 each predictor variables using likelihood ratio tests (LRT, $\alpha = 0.05$) and all models were
301 validated by plotting residual *vs.* fitted values and by checking for residual autocorrelation.
302 Relative variable importance was calculated dividing the sum of Akaike weights by the
303 number of models containing each predictor variable from those included in set of best-fitting
304 models. Model averaging across the set of best models was used to compute parameters
305 estimates that account for uncertainty in model selection [39].

306 *Gene flow*

307 To assess the effect of habitat loss on gene flow, we employed a genetic algorithm to
308 optimize gene flow hypotheses and then tested them by modeling isolation by resistance
309 (IBR, [16]). Yang's genetic relatedness between pairs of individuals [36] was used as a proxy for
310 gene flow, as it was developed for SNP markers and similar measures of relatedness have
311 proven highly accurate individual-based genetic distance metrics in landscape genetic studies
312 [40]. Resistance to gene flow due to mining was modeled using land cover maps for different
313 years (2016, 2014, 2011 and 1979) containing only the major land cover classes of our study
314 region: Montane Savanna (Canga), Forest (evergreen forest) and Mine. Water bodies,
315 Pastureland and Urban areas were excluded because they occurred outside the extent of our
316 samples (Fig. 1). By so doing we were able to evaluate the permeability to gene flow of each
317 land cover class; and test whether habitat loss driven by mining hindered gene flow across
318 our replicated landscapes. Additional variables found to be important predictors of gene flow
319 in other plants [17,18] were modeled along with land cover, including geographic distance,
320 elevation (DEM retrieved from the USGS Earth Explorer), terrain roughness (generated from

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the DEM using the Terrain Analysis plug-in from QGIS), and bioclimatic variables (retrieved from WorldClim). To select a set of orthogonal variables explaining most climatic variation across our study area, we first ran separate principal component analyses (PCA) for each species using the extracted values from all 19 WorldClim bioclimatic layers plus elevation (scaled). We then selected the three variables showing the strongest correlation with the first, second and third PCA axis (which explained more than 85% of total variance in both *B. carajensis* and *M. carajensis*). These were minimum temperature of coldest month (bio06) and precipitation of wettest (bio16) and coldest quarter (bio19) for *B. carajensis*; and minimum temperature of coldest month (bio06), precipitation of wettest quarter (bio16) and temperature seasonality (bio04) for *M. carajensis*.

A genetic algorithm implemented through the *ResistanceGA* package was used to generate optimized resistance surfaces for each one of these variables [15]. In the case of land cover maps, random initial resistance values were assigned for each class; then pairwise effective distances were measured using random-walk commute times; and finally pairwise genetic distance was regressed on effective distance using maximum likelihood population effect models (MLPE, see below). The whole process was iterated until no significant change was found in the objective function [15]. We then performed the same steps for the remaining continuous predictors, but instead of assigning random initial resistance values, eight types of transformations were applied to the raw values. In this case, two parameters controlling Ricker and Monomolecular functions were iteratively varied during the optimization [15]. Ten independent runs of optimization were conducted for each surface to assess the convergence in parameter estimates [41]. All rasters were set to Universal Transverse Mercator (UTM) projection, and cropped to the extent of sampling locations plus a buffer area of 5 km to minimize border effects [17]. Land cover resistance surfaces and terrain roughness were

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

optimized using 250 x 250 m resolution maps, while 900 x 900 m resolution maps were used for WorldClim layers as this is the highest available. Serra Norte and Serra Sul were analyzed separately aiming to replicate IBR analyses in two separate areas exposed to open-pit mining.

Using the program Circuitscape V4.0 [16], we then calculated pairwise resistance distances between all samples, employing the optimized resistance surfaces described above plus a surface where all pixels were set to 1 to assess isolation by geographic distance (IBD). To assess isolation by resistance (IBR), defined as the correlation between genetic and resistance distances [16], we fitted mixed-effects regression models using penalized least squares and a correlation structure designed to account for the non-independence of pairwise distances (maximum-likelihood population effects - MLPE: https://github.com/nspope/corMLPE; [42]). Yang's genetic relatedness between individuals was used as the response variable and the different resistance distances (contemporary and historical land cover, elevation, terrain roughness, temperature, precipitation, and geographic distance) as predictors. All MLPE models accounted for the underlying population structure, either by considering only individuals belonging to the same genetic cluster (most cases), or by including an additional random effect specifying if pairwise distances represented individuals from the same or from different genetic clusters (the case of *B. carajensis* from Serra Norte, see genetic structure results). To evaluate the incidence of time-lag effects potentially masking mining effects on gene flow, we first fitted uni-variate models for each species and region using resistance distances from land cover surfaces from all years, plus those from geographic distance surfaces. The best models were selected using the Akaike Information Criterion ($\Delta AIC < 2$), and whenever geographic distance was found among the best models we considered IBD as the most parsimonious gene flow model. To evaluate the sensitivity of our analysis to the resolution (grain size) of spatial data, we also compared uni-

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variate land cover models containing resistance distances computed from surfaces with different grain sizes (100 x 100 m, 300 x 300 m, 600 x 600 m and 900 x 900 m). Results were consistent across the different resolutions (Supplementary Table S3), so we ran all subsequent analysis using a grain size of 900 x 900 m. We then fitted multiple regression models containing resistance distances from the best uni-variate land cover models selected in the previous step and resistance distance from all other optimized surfaces for each species and region. Models containing all possible combinations of non-collinear predictors ($r < 0.6$, Supplementary Figure S11) were compared using the *dredge* function from the package *MuMIn* (https://github.com/rojaff/dredge_mc; [43]), and best models were selected using AIC. Likelihood ratio tests (LRT) were performed to assess the influence of each predictor variable on the best model's log-likelihood [44], and relative variable importance and model-averaging were calculated as described above. Finally, we carried out a barrier analysis to identify genetic discontinuities between individuals by using Monmonier's algorithm and Gabriel's graph implemented in package *adegenet* [34].

Germination experiments

To evaluate if seeds from both study species are able to germinate inside iron ore mines, we ran a set of germination experiments. Seeds from both species were sown over four different substrates (Whatman® paper, Canga topsoil, forest topsoil, and mining waste substrate) placed in plastic boxes (Gerbox – 11 x 11 x 4 cm) and kept in a growth chamber (Fitotron SGC 120, Weiss Technik, UK) under continuous darkness, constant temperature (20°C) and air humidity (60%) for 33 consecutive days, from September 4th to October 7th 2018. Substrates received distilled water until the retention capacity, and water losses by evaporation were replaced daily. All treatments were carried out with five replicates for each

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

392 substrate in each species. Each replicate contained 25 seeds from *B. carajensis* and 12 seeds
393 from *M. carajensis*. The number of germinated seeds was recorded daily, with germination
394 defined as the emission of 2 mm of primary root.

395 *Data availability*

396 Genotype data will be deposited in figshare and url addresses provided upon the acceptance
397 of this manuscript.

398 *Code availability*

399 Custom code has been deposited in GitHub and is cited in the text.

400 **References**

401 1. Fahrig, L. Effects of habitat fragmentation on biodiversity. *Annu. Rev. Ecol. Evol. Syst.*
402 **34**, 487–515 (2003).

403 2. DiLeo, M. F. & Wagner, H. H. A landscape ecologist's agenda for landscape genetics.
404 *Curr. Landsc. Ecol. Reports* **1**, 115–126 (2016).

405 3. Allendorf, F. W., Luikart, G. H. & Atiken, S. N. *Conservation and the genetics of*
406 *populations*. (Wiley-Blackwell, 2013).

407 4. Tittensor, D. P. *et al.* A mid-term analysis of progress toward international biodiversity
408 targets. *Science (80-.).* **346**, 241–244 (2014).

409 5. Llorens, T. M., Ayre, D. J. & Whelan, R. J. Anthropogenic fragmentation may not alter
410 pre-existing patterns of genetic diversity and differentiation in perennial shrubs. *Mol.*
411 *Ecol.* **27**, 1541–1555 (2018).

412 6. Anderson, C. D. *et al.* Considering spatial and temporal scale in landscape-genetic
413 studies of gene flow. *Mol. Ecol.* **19**, 3565–3575 (2010).

414 7. *Landscape Genetics: Concepts, Methods, Applications*. (John Wiley & Sons, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

8. Storfer, A., Murphy, M. A., Spear, S. F., Holderegger, R. & Waits, L. P. Landscape genetics: where are we now? *Mol. Ecol.* **19**, 3496–3514 (2010).

9. McCartney-Melstad, E., Vu, J. K. & Shaffer, H. B. Genomic data recover previously undetectable fragmentation effects in an endangered amphibian. *Mol. Ecol.* **27**, 4430–4443 (2018).

10. Rocha, K. C. de J., Goldenberg, R., Meirelles, J. & Viana, P. L. Flora das cangas da Serra dos Carajás, Pará, Brasil: Melastomataceae. *Rodriguesia* **68**, 997–1034 (2017).

11. Cruz, A. P. O., Viana, P. L. & Santos, J. U. Flora das cangas da Serra dos Carajás, Pará, Brasil: Asteraceae. *Rodriguesia* **67**, 1211–1242 (2016).

12. Skirycz, A. *et al.* Canga biodiversity, a matter of mining. *Front. Plant Sci.* **5**, 653 (2014).

13. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

14. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).

15. Peterman, W. E. ResistanceGA : An R package for the optimization of resistance surfaces using genetic algorithms. *Methods Ecol. Evol.* **9**, 1638–1647 (2018).

16. McRae, B. H. Isolation by Resistance. *Evolution (N. Y).* **60**, 1551 (2006).

17. Lanes, É. C. *et al.* Landscape genomic conservation assessment of a narrow-endemic and a widespread morning glory from Amazonian Savannas. *Front. Plant Sci.* **9**, 532 (2018).

18. Dyer, R. J. Landscapes and Plant Population Genetics. in *Landscape Genetics: Concepts, Methods, Applications* (eds. Balkenhol, N., Cushman, S., Storfer, A. & Waits, L.) 181–198 (Wiley Online Library, 2016).

19. Esfeld, K. *et al.* Molecular data indicate multiple independent colonizations of former lignite mining areas in Eastern Germany by Epipactis palustris (Orchidaceae). *Biodivers. Conserv.* **17**, 2441–2453 (2008).

20. Waterhouse, M. D., Blair, C., Larsen, K. W. & Russello, M. A. Genetic variation and fine-scale population structure in American pikas across a human-modified landscape. *Conserv. Genet.* **18**, 825–835 (2017).

21. Leimu, R., Mutikainen, P., Koricheva, J. & Fischer, M. How general are positive relationships between plant population size, fitness and genetic variation? *J. Ecol.* **94**, 942–952 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

22. Soons, M. B., Heil, G. W., Nathan, R. & Katul, G. G. Determinants of long-distance seed dispersal by wind in grasslands. *Ecology* **85**, 3056–3068 (2004).

23. Heydel, F., Cunze, S., Bernhardt-Römermann, M. & Tackenberg, O. Long-distance seed dispersal by wind: disentangling the effects of species traits, vegetation types, vertical turbulence and wind speed. *Ecol. Res.* **29**, 641–651 (2014).

24. Dick, C. W., Hardy, O. J., Jones, F. A. & Petit, R. J. Spatial scales of pollen and seed-mediated gene flow in tropical rain forest trees. *Trop. Plant Biol.* **1**, 20–33 (2008).

25. Schlaepfer, D. R., Braschler, B., Rusterholz, H.-P. & Baur, B. Genetic effects of anthropogenic habitat fragmentation on remnant animal and plant populations: a meta-analysis. *Ecosphere* **9**, e02488 (2018).

26. Rogstad, S. H. Saturated NaCl-CTAB solution as a means of field preservation of leaves for DNA analyses. *Taxon* **41**, 701 (1992).

27. Doyle, J. & Doyle, J. L. Genomic plant DNA preparation from fresh tissue-CTAB method. *Phytochem Bull* **19**, 11–15 (1987).

28. Michaels, S. D., John, M. C. & Amasino, R. M. Removal of polysaccharides from plant DNA by ethanol precipitation. *Biotechniques* **17**, 274–276 (1994).

29. Loureiro, J., Rodriguez, E., Dolezel, J. & Santos, C. Two new nuclear isolation buffers for plant DNA flow cytometry: A test with 37 species. *Ann. Bot.* **100**, 875–888 (2007).

30. Dolezel, J., Sgorbati, S. & Lucretti, S. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol. Plant.* **85**, 625–631 (1992).

31. Russello, M. A., Waterhouse, M. D., Etter, P. D. & Johnson, E. A. From promise to practice: pairing non-invasive sampling with genomics in conservation. *PeerJ* **3**, e1106 (2015).

32. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

33. François, O., Martins, H., Caye, K. & Schoville, S. D. Controlling false discoveries in genome scans for selection. *Mol. Ecol.* **25**, 454–469 (2016).

34. Jombart, T. & Ahmed, I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).

35. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G. & François, O. Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196**, 973–983 (2014).

36. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

37. Tarpley, J. D., Schneider, S. R. & Money, R. L. Global vegetation indices from the NOAA-7 meteorological satellite. *J. Clim. Appl. Meteorol.* **23**, 491–494 (1984).

38. Pinheiro, J., Bates, D., DebRoy, S. & Sarkar, D. Linear and nonlinear mixed effects models. *R Packag. version* (2009).

39. Burnham, K. K. P. & Anderson, D. R. D. *Model selection and multimodel inference: a practical information-theoretic approach (2nd ed)*. *Springer* **172**, (2002).

40. Shirk, A. J., Landguth, E. L. & Cushman, S. A. A comparison of individual-based genetic distance metrics for landscape genetics. *Mol. Ecol. Resour.* **17**, 1308–1317 (2017).

41. Khimoun, A. *et al.* Landscape genetic analyses reveal fine-scale effects of forest fragmentation in an insular tropical bird. *Mol. Ecol.* **26**, 4906–4919 (2017).

42. Clarke, R. T., Rothery, P. & Raybould, A. F. Confidence limits for regression relationships between distance matrices: Estimating gene flow with distance. *J. Agric. Biol. Environ. Stat.* **7**, 361–372 (2002).

43. Barton, K. & Barton, M. K. Package 'MuMIn'Version 1. (2015).

44. Jaffé, R. *et al.* Beekeeping practices and geographic distance, not land use, drive gene flow across tropical bees. *Mol. Ecol.* **25**, 5345–5358 (2016).

Acknowledgments

Funding was provided by Instituto Tecnológico Vale, National Council for Scientific and Technological Development (CNPq) grants 301616/2017-5 (RJ), 307479/2016-1, and 402756/2018-5 (GO) and 300714/2017-3 (EL), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) grant 88887.156652/2017-00 (CSC). We thank Alexandre Castilho, Cesar Neto and Waléria Monteiro for assistance in the field, Gleiciane Salvador and Manoel Lopes for help in the laboratory, and Prof. Ing. Jaroslay Dolezel for provinding the standards used in flow cytometry analyses.

Author contributions

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

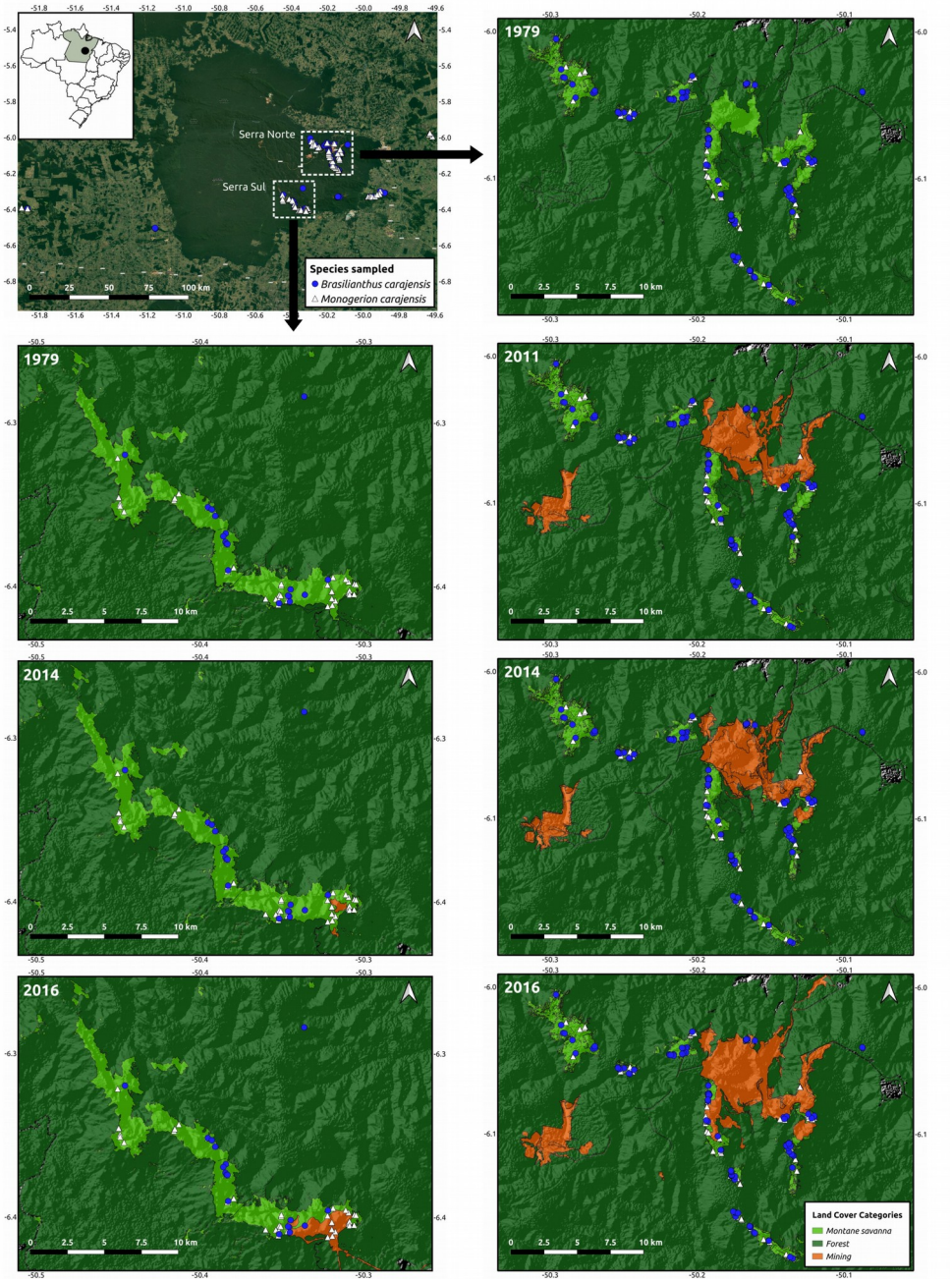
RJ conceived, designed and coordinated the project. RJ and PLV coordinated the field work and sampling. ECML, ARS, CFC and MG performed laboratory work. RJ, ECML, CSC, ARS and CFC performed the data analysis. The first draft of the paper was written by CSC and ECML with input from RJ. All authors contributed to discussing the results and editing the paper.

Competing interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

516 **Figure legends**



517 **Fig. 1.** Map of the study region depicting the location of the collected samples from

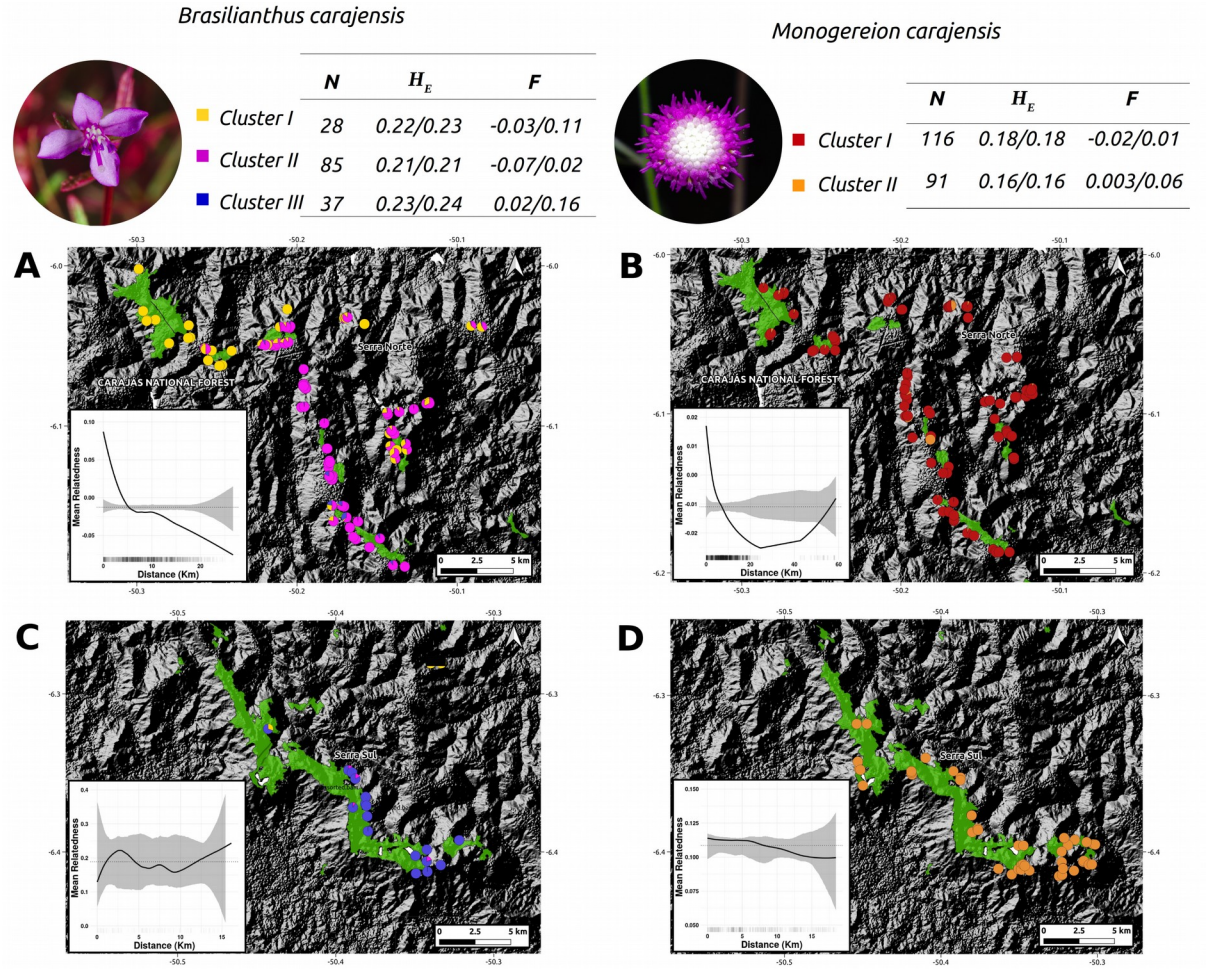
518 *Brasilianthus carajensis* (blue circles) and *Monogereion carajensis* (white triangles) in Serra

519 Norte (right panels) and Serra Sul (left panels). Hill shade maps are shown overlaid with land

520 cover color maps for the different years analyzed. The location of the Carajás Mineral

521 Province within Brazil is shown on the upper left corner.

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



522 **Fig. 2.** Map showing the ancestry coefficients from *Brasilianthus carajensis* (A and C) and
523 *Monogereion carajensis* (B and D) in Serra Norte (upper panels) and Serra Sul (lower panels)
524 determined using the Admixture software. Montane Savanna areas are shown in green against
525 hill shade layers. Smaller lower-left corner plots show spatial autocorrelation in genetic
526 relatedness, where black solid lines are the LOESS fit to the observed relatedness, and gray
527 shaded regions are 95% confidence bounds around the null expectation (black dotted lines).
528 Genetic diversity measures for each genetic cluster are shown in the upper tables. Sample
529 sizes (N) are followed by mean expected heterozygosity (H_E) and mean inbreeding coefficient
530 (F), and values represent 95% confidence intervals.

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 3. Relative variable importance in the set of best-fitting models ($\Delta AIC \leq 2$) for *Brasilianthus carajensis* and *Monogereion carajensis* in Serra Norte and Serra Sul (see methods and Supplementary Tables S1 and S4 for details). Individual-level genetic diversity metrics (H_E and F) were response variables and habitat amount in 1979 and habitat loss in 2011, 2014 and 2016 were predictors in genetic diversity models. Pairwise inter-individual genetic relatedness was the response variable and resistance distances computed from optimized surfaces were predictors in IBR models. Likelihood Ratio Test (LRT) were performed to assess if each predictor variable significantly improved the model's log-likelihood (significance levels are highlighted with: *p < 0.05; **p < 0.01; and ***p < 0.001).

bioRxiv preprint doi: https://doi.org/10.1101/528430; this version posted January 23, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

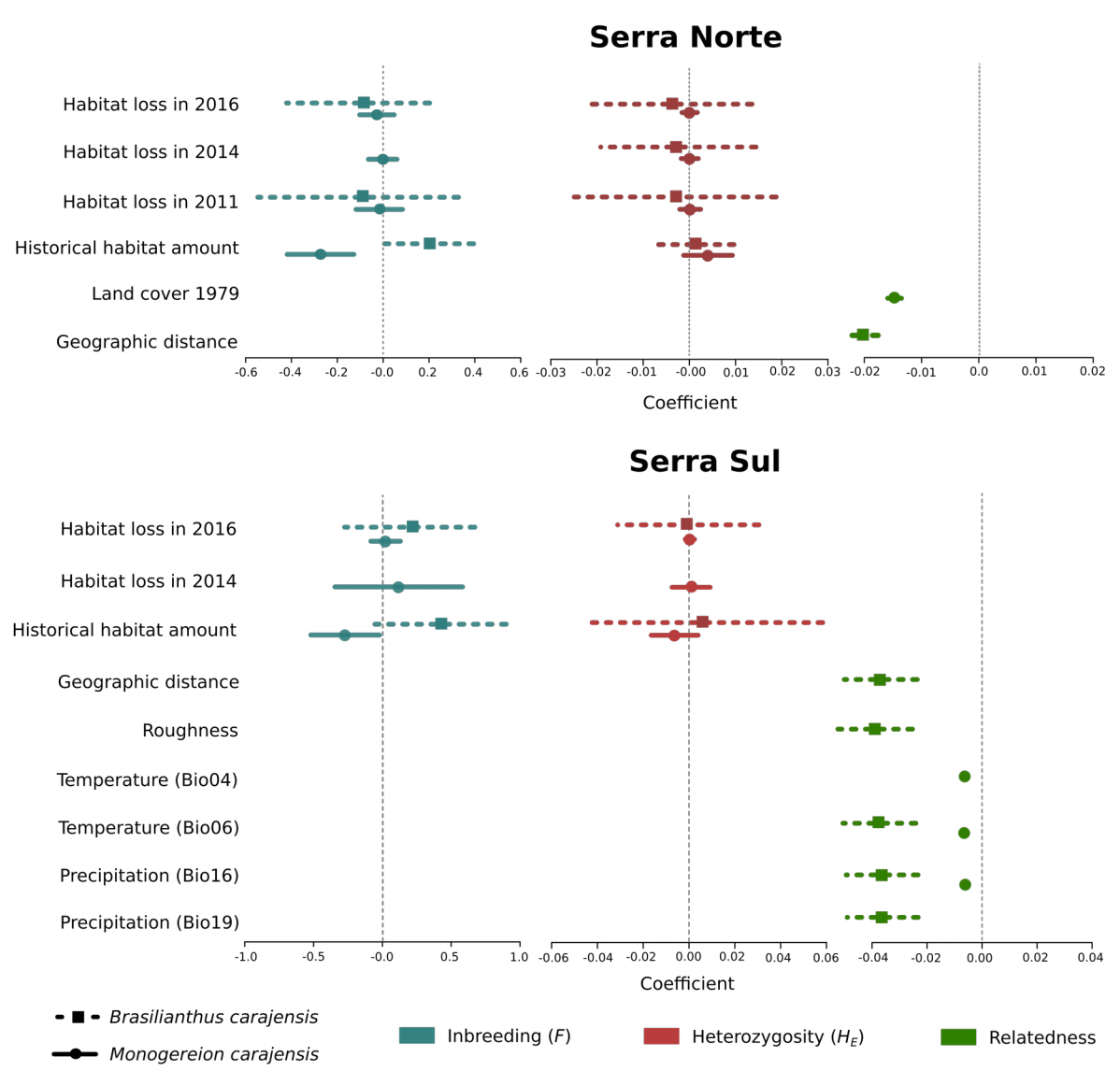


Fig. 4. Coefficient plots for the set of best-fitting models (ΔAIC $\leq$ 2) for *Brasilianthus carajensis* and *Monogereion carajensis* in Serra Norte and Serra Sul (see methods and Supplementary Tables S2 and S5 for details). Points represent model-averaged regression coefficients and lines the 95% confidence intervals.