Combining genotype, phenotype, and environmental data to delineate site-adjusted provenance strategies for ecological restoration

Carolina S. Carvalho[1], Brenna R. Forester[2], Simone K. Mitre[1], Ronnie Alves[1], Vera L. Imperatriz-Fonseca[1], Silvio J. Ramos[1], Luciana C. Resende-Moreira[1], José O. Siqueira[1], Leonardo C. Trevelin[1], Cecilio F. Caldeira[1], Markus Gastauer[1], Rodolfo Jaffé[1,3*]

[1] Instituto Tecnológico Vale, Belém-PA, Brazil.

[2] Colorado State University, Fort Collins-CO, USA

[3] Departamento de Ecologia, Universidade de São Paulo, São Paulo-SP, Brazil.

[*] **Correspondence:**

Rodolfo Jaffé

email: r.jaffe@ib.usp.br

phone: +55 (91) 3213 5523

Summary

Despite the importance of climate-adjusted provenancing to mitigate the effects of climate change, climatic considerations alone are insufficient when restoring highly degraded sites. Here we employ a comprehensive landscape genomic approach to assist the restoration of moderately disturbed and highly degraded sites. Relying on genotyping-by-sequencing we identified thousands of single nucleotide polymorphisms in two plant species of special interest for the restoration of iron-rich Amazon Savannas from the Eastern Amazon. We first used a subset of neutral loci to assess genetic structure and determine the genetic neighborhood size, and then combined uni-variate and multi-variate methods to identify genotype-environment and genotype-phenotype associations, representing local adaptations. We then mapped adaptive genetic variation and predicted adaptive genotypes for both restoration sites. Whereas local provenances where found optimal to restore the moderately disturbed site, a mixture of genotypes was the most promising strategy to recover a highly degraded mining site, to which local provenances were not adapted. We discuss how our results can help define site-adjusted provenancing strategies and argue that our methods can be more broadly applied to assist other restoration and rehabilitation initiatives.

Keywords: Genotype-environment associations (GEA), genotype-phenotype associations (GPA), genome-wide association study (GWAS), landscape genomics, local adaptation, RAD sequencing, restoration genomics, single nucleotide polymorphisms (SNP).

Introduction

In spite of the broadly recognized importance of genetic provenance for restoration initiatives, the use of genomic tools to define provenance strategies is still uncommon. Choosing provenances based on genetic knowledge can help increase genetic diversity and adaptability, thereby contributing to the success of restoration initiatives (Broadhurst *et al.*, 2008; Weeks *et al.*, 2011; Mijangos *et al.*, 2015). Fortunately, the use of genetic provenancing is increasing, as advances in Next-Generation-Sequencing technologies have made possible large-scale assessments of neutral and adaptive genetic variation (Williams *et al.*, 2014; Mijangos *et al.*, 2015; Breed *et al.*, 2019). For instance, neutral loci (i.e., those which are not subject to natural selection) can be used to identify independent demographic units, assess fine-scale spatial genetic structure, and quantify genetic diversity (Allendorf *et al.*, 2013; Balkenhol *et al.*, 2017); whereas adaptive loci (under natural selection) are relevant to detect adaptations to local environmental conditions and delineate adaptive units (Funk *et al.*, 2012; Rellstab *et al.*, 2015).

Restoration genomic studies published so far have assessed the effect of multiple environmental variables on genetic composition in order to identify which individuals or populations are "pre-adapted" to future climates, an approach known as predictive provenancing (Steane *et al.*, 2014; Shryock *et al.*, 2015, 2017; Gugger *et al.*, 2018; Martins *et al.*, 2018; Supple *et al.*, 2018; Rossetto *et al.*, 2019; Lu *et al.*, 2019). Although this emphasis on climate-adjusted provenancing is crucial to proactively combat global climate change (Prober *et al.*, 2015), it has overshadowed the application of genomic methods to restore extremely degraded sites (Lesica & Allendorf, 1999a; Bucharova *et al.*, 2019). Such site-adjusted provenancing may not even incorporate climate change considerations, as highly degraded sites have unique characteristics that make them extremely challenging to restore. First, highly degraded sites usually require immediate restoration or rehabilitation, so adaptations to current environmental conditions are more suitable to guide provenance strategies than those based on future climate (Gastauer *et al.*, 2019). Second, environmental protection agencies usually require the restitution of ecosystems to conditions as close to a pre-disturbance baseline as possible, as well as regular (and costly) monitoring until rehabilitation goals have been achieved (Gastauer *et al.*, 2019). The main efforts thus lay in the quick establishment of viable populations that will restore ecosystem functions and processes, prevent soil erosion, and protect biological diversity, and less in future resilience to climate change. Third, as highly degraded sites have radically different environmental characteristics than natural habitats (Gastauer *et al.*, 2019), site-specific characteristics are of primary importance to define provenance strategies. Such site-specific variables generally need to be measured in situ and at high

spatial resolution, since they may not be available as spatial layers in open-access repositories (and if they are, their spatial resolution may be too coarse to reflect the reality of environmental conditions on the ground). Finally, site-adjusted provenancing strategies need to consider local adaptations to climate, soil, terrain, and even biological interactions, whereas climate-adjusted provenancing focuses exclusively on climate.

Both the degree and size of disturbances are known to play an important role in determining site-adjusted provenancing strategies (Lesica & Allendorf, 1999b). Whereas local genotypes are best suited to restore sites where the degree of disturbance is low, adaptations found in distant populations may facilitate establishment in highly degraded sites to which local genotypes may not be adapted. Local provenancing is also desirable when the size of the disturbance is large, as this strategy minimizes changes in the genetic composition of local natural populations due to introgression. Although the degree of genetic introgression will depend on the plant's breeding system and the strength of selection, introduced alleles will have little chance of spreading widely across natural populations if the size of the restored site is small. Mixtures of genotypes from different populations have thus been suggested as the best strategy to recover small, highly degraded sites, given that enhanced genetic variation is more likely to rapidly generate local adaptations to novel ecological challenges (Lesica & Allendorf, 1999b). In either case, determining the most appropriate site-adjusted provenancing strategy will require assessment of the size of local provenances and the spatial distribution of local adaptations (Breed *et al.*, 2019).

Although the use of neutral genetic markers to identify independent demographic units is now common practice (Coates *et al.*, 2018), few restoration studies have delineated seed sourcing strategies based on population genetic structure (Durka *et al.*, 2017) or the genetic neighborhood size (the distance at which genetic composition stops being spatially autocorrelated) (Krauss & Koch, 2004; Krauss *et al.*, 2013; Rossetto *et al.*, 2019). On the other hand, only three restoration genomic studies so far have identified putative adaptive loci and then mapped adaptive genetic variation (Steane *et al.*, 2014; Shryock *et al.*, 2015; Martins *et al.*, 2018). While the assessment of phenotype through common garden and reciprocal transplant experiments to identify local adaptations has a long history (Aitken & Bemmels, 2016), no study has yet combined genotype-environment associations (GEA) with genotype-phenotype associations (GPA) to delineate seed sourcing areas. This approach could improve the inference of potential candidate genes, provide important insights into genes underlying fitness-related traits, and avoid the logistical challenges associated with the implementation of long-term common garden experiments (Talbot *et al.*, 2016; Vangestel *et al.*, 2018; Mahony *et al.*, 2019).

Here we employ a comprehensive landscape genomic approach to assist the restoration of moderately disturbed and highly degraded sites. Relying on genotyping-by-sequencing we identified thousands of single nucleotide polymorphisms in two plant species of special interest for the restoration of exhausted mining sites from the Carajás Mineral Province, located in the Eastern Amazon (Skirycz *et al.*, 2014; Souza-Filho *et al.*, 2019). We first used a subset of neutral loci to assess broad and fine-scale genetic structure and determine the genetic neighborhood size. Subsequently, we combined univariate and multivariate methods to identify GEA and GPA and employed spatial principal components analyses (sPCA) to map adaptive genetic variation while accounting for spatial autocorrelation in genetic composition. Finally, we predicted the adaptive genotypes best suited to environmental conditions at each restoration site (Fig. 1).

We focus on the Carajás Mineral Province, which harbors one of the world's largest deposits of iron ore and huge iron ore mining projects, with operations dating back to the 1980s (Poveromo, 1999; Souza-Filho *et al.*, 2019). The banded ironstone formations known as *Cangas* (where iron-ore deposits are concentrated) are characterized by shallow, acidic, nutrient-depleted and metal-rich soils and marked by high solar radiation, hot temperatures, and a severe drought period (Skirycz *et al.*, 2014), which pose severe challenges to plant growth. Environmental legislation in Brazil requires the rehabilitation of disabled iron ore mining sites (Gastauer *et al.*, 2019), which constitute extremely degraded and difficult to restore environments. These mining sites are characterized by extensive vegetation and soil removal, compacted, nutrient-poor soils and steep slopes (Whiting *et al.*, 2004; Boyer & Wratten, 2010; Garris *et al.*, 2016). The successful restoration and rehabilitation of these highly degraded sites thus requires an appropriate selection of plant species (Giannini *et al.*, 2017) and seed sourcing zones to ensure that the introduced plants can effectively colonize and establish viable populations. However, mine rehabilitation programs in the region employ seed mixtures of exotic plants, due to the scarcity of native seeds and their lower germination and growth rates (Silva *et al.*, 2018). Additionally, some natural Canga environments from the region have been repeatedly disturbed by fires, illegal cattle ranching and the introduction of alien plant species, such as grasses and ferns. This is the case in Serra da Bocaina, which became a National Park in 2017 and has since been protected (Mota *et al.*, 2018). No restoration initiatives have yet been implemented to recover the original ecosystems found in Serra da Bocaina, but provenance strategies are expected to differ substantially from those required to rehabilitate disabled mines.

Being native species and dominant in Canga environments, *Mimosa acutistipula var. ferrea* Barneby and *Dioclea apurensis* Kunth (both legumes) are among the most promising plants for use in Canga restoration and mineland rehabilitation programs (Giannini *et al.*, 2017). Considered

metallophyte species, both exhibit biological mechanisms to tolerate and thrive in metalliferous soils (Whiting *et al.*, 2004; Preite *et al.*, 2019). Moreover, they are extremely abundant in pristine Cangas ecosystems, and interact symbiotically with nitrogen-fixing bacteria, contributing to soil enrichment and thus acting as pioneer species in restoration sites (Nunes *et al.*, 2015; Silva *et al.*, 2018; Ramos *et al.*, 2019a). *Mimosa acutistipula* is drought tolerant and well adapted to the low nutrient content of Canga soils (Silva *et al.*, 2018). On the other hand, *D. apurensis* requires low nutrient inputs and shows high nutrient use efficiency (Ramos *et al.*, 2019b). Moreover, this species is a fast-growing liana with a ground-covering growth form, enabling the revegetation and stabilization mine pits and waste piles. Both species have high germination rates (Ramos *et al.*, 2019a), can be observed growing on minelands, and seem to be central in plant-pollinator networks (unpublished data). Considering the heterogeneous and hostile environment where both species occur (Mitre *et al.*, 2018), and their life-history similarities, we expected to find similar patterns of neutral and adaptive genetic structure in both species across our study area. We also expected that local populations would not be adapted to the environmental conditions found in exhausted mining sites, whereas populations from Serra da Bocaina would show adaptations to local environmental conditions. Based on our results, we propose two different site-adjusted provenance strategies for the restoration of a degraded mine site and a disturbed but unmined Canga environment. We discuss the merits of our approach and argue that it can be more broadly applied to define site-adjusted provenancing strategies.

Material and Methods

Sampling

We collected samples of 180 individuals of *M. acutistipula* var. *ferrea* and 167 individuals of *D. apurensis* between February and May of 2018 (SISBIO collection permit N. 48272-6), across the three major Canga highlands of the Carajás Mineral Province (Fig. 2). For each individual plant, we collected a sample of root-proximal soil (0-5 cm) for chemical characterization and leaflet samples for phenotype and genotype analyses. These Canga ecosystems are composed of several physiognomies, comprising grasslands, scrublands, wetlands and forest formations (Mota *et al.*, 2015), which differ in terms of the plant communities they support as well as in their soil chemistry (Mitre *et al.*, 2018). To ensure sampling across environmental gradients, individuals were collected in each one of these physiognomies within each highland. We also scattered samples to cover the full extent of each highland (Fig. 2). A minimum distance of 20 meters between samples was used to minimize sampling related individuals. In addition to the soil samples gathered along with plant

tissue, we collected 50 extra soil samples from a small, highly degraded mine site (the mine pit from an exhausted mine) and seven soil samples from a large, unmined, but moderately disturbed site (scattered across Serra da Bocaina, Fig. 2).

Environmental data

Soil samples were air-dried and sieved using 2 mm mesh, and once dry they were sent to LABRAS (http://labrasambientaiseagricolas.com.br/) for chemical analyses. These included pH, organic matter, available P, K, and Na, exchangeable Ca, Mg, and Al, exchangeable S and available B, Cu, Fe, Mn and Zn (see details in supporting information Methods S1). To reduce the number of chemical parameters describing soil composition, we selected those known to most affect plant physiology in metalliferous ecosystems (organic matter, Fe, Mn, P, pH, S, and B (Whiting *et al.*, 2004; Bothe, 2011; Mitre *et al.*, 2018)), along with a set of orthogonal variables explaining most variation in soil composition across our study area. To identify this set of orthogonal variables we first used the function *imputePCA* from the *missMDA* R package (Josse & Husson, 2016) to impute missing data (20 samples contained missing data for at least one parameter) using the regularized iterative PCA algorithm recommended to avoid overfitting (Josse & Husson, 2016). We then ran separate principal component analyses (PCA) for each species using all of the centered and scaled chemical parameters and selected the three variables showing the strongest correlation with the first, second and third PC axes (which explained more than 50% of total variance and were organic matter, Zn and Na in both species).

We also retrieved climatic data from WorldClim (http://www.worldclim.com/bioclim), and followed a similar protocol to obtain the set of orthogonal variables explaining most climatic variance across our study region. The first three PCA axes explained 80% of total climatic variance in both species, and the bioclimatic variables most strongly correlated with those axes were isothermality (bio03), minimum temperature of coldest month (bio06) and precipitation of driest quarter (bio17) for *M. acutistipula*; and isothermality (bio03), minimum temperature of coldest month (bio06) and maximum temperature of warmest month (bio05) for *D. apurensis* (see Fig. S1 for maps of these bioclimatic layers). Geographic coordinates from the additional soil samples collected at the mining site and Serra da Bocaina were then used to extract the values from these climatic layers.

Phenotypic data

For each leaf sample we determined macro- and micro-nutrient content and the specific leaf area (SLA), using standard methods (see details in Methods S1). We selected those phenotypic variables

known to affect plant physiology in metalliferous ecosystems (SLA, N, B, Fe, Mn, P, N/P, (Bothe, 2011; Pérez-Harguindeguy *et al.*, 2013; Mitre *et al.*, 2018). As described above, we also selected the three orthogonal phenotypic variables explaining most of phenotypic variance (the first three PCA axes explained more than 50% of total variance in both species) after imputing missing data (11 samples contained missing data for at least one parameter). These orthogonal phenotypic variables explaining most variance were Zn, N and B for *M. acutistipula* and P, Mn and K for *D. apurensis.*

Genome size estimation, DNA extraction, genotype-by-sequencing and bioinformatic processing

We used flow cytometry to estimate genome size in both species (2C DNA content was 714 Mbp in *M. acutistipula* and 978 Mbp in *D. apurensis,* see details in Methods S1). Total DNA was extracted using Qiagen's DNeasy Plant Mini Kit. DNA concentration was quantified using Qubit High Sensitivity Assay kit (Invitrogen), and DNA integrity assessed through 1.2% agarose gel electrophoresis. Samples with concentrations below 5 ng/µL or showing no clean bands were excluded from all analyses, and selected samples were normalized to a concentration of 5 ng/µL and a total volume of 30 µL. These were then shipped to SNPSaurus (http://snpsaurus.com/) for sequencing and raw data bioinformatic processing (see details in Methods S1). All reads were mapped to the reference contig with an alignment identity threshold of 95% using *BBMap tools*. Genotype calling was done using *callvariants* (*BBMap tools*), and the resulting set of genotypes were filtered to remove alleles with population frequency of less than 3%. Loci that were heterozygous in all samples and loci that contained more than 2 alleles in a sample (suggesting collapsed paralogs) were removed. A total of 7,165 RAD-tag sequences were obtained for *M. acutistipula* and 4,325 for *D. apurensis.* Considering the genome size of each species and a linkage block size of 378 Mp (mean value for the Fabaceae family), we estimated a maximum proportion of genome coverage (assuming one RAD-tag per block) of 100% (McKinney *et al.*, 2017). From those RAD-tags, 17,403 SNPs were generated for *M. acutistipula* (sequencing depth between 14 and 2,756) and 9,857 SNPs for *D. apurensis* (depth between 9 and 2,136).

Genetic diversity and neutral genetic structure

The R package *r2vcftools* (https://github.com/nspope/r2vcftools) - a wrapper for VCFtools (Danecek *et al.*, 2011) - was used to perform final quality control on the genotype data. To assess neutral genetic structure and genetic diversity, we used a series of filters to obtain a set of neutral and independent loci. Filtering criteria included quality (Phred score > 30), read depth (20 – 800), minor allele frequency (MAF > 0.05), linkage disequilibrium ($r^2 < 0.8$, (Xuereb *et al.*, 2018)), Hardy-Weinberg Equilibrium (HWE, $p > 0.0001$), and loci and individuals with less than 20%

missing data (an example filtering script can be seen in https://github.com/rojaff/r2vcftools_basics). Additionally, we removed loci potentially under selection using genome scans. These accounted for population structure (assessed using the *snmf* function from the *LEA* package, as described below), and controlled for false discovery rates by adjusting *p*-values with the genomic inflation factor (λ) and setting false discovery rates to q=0.05, using the Benjamini-Hochberg algorithm (François *et al.*, 2016) (see details below).

We used two complementary genetic clustering approaches to assess neutral population structure: the *snmf* function from the *LEA* package (Frichot & François, 2015), and Discriminant Analysis of Principal Components - DAPC from the *adegenet* package (Jombart & Ahmed, 2011). The *snmf* model implements a fast yet accurate likelihood algorithm (Frichot et al., 2014), while DAPC is an assumption-free robust genetic clustering method (Jombart & Ahmed, 2011). For *snmf,* we tested from one to ten ancestral populations (k), with ten replicate runs for each value of k, choosing the most likely k based on minimized cross-entropy. In DAPC, we tested from one to ten clusters, inferring the optimal number of clusters from the lowest Bayesian Information Criterion (BIC) value. Considering the ancestry coefficients assigned by *snmf*, we then estimated expected heterozygosity (H_E), inbreeding coefficients (F), and nucleotide diversity (π) for each genetic cluster. We also estimated their effective population size (N_e) employing the linkage disequilibrium method implemented in NeEstimator 2.1 and a lowest allele frequency value of 0.05 (Do *et al.*, 2014). Finally, we assessed fine-scale spatial genetic structure within each genetic cluster through local polynomial fitting (LOESS) of Yang's genetic relatedness between pairs of individuals (Yang *et al.*, 2010) and pairwise geographic distance, as in (Carvalho *et al.*, 2019).

Assessing genotype-environment associations (GEA) and genotype-phenotype associations (GPA)

To assess GEA and GPA (Fig. 1) we first filtered loci for quality (Phred score > 30), read depth (20-800), minor allele frequency (MAF > 0.05), linkage disequilibrium (r^2 < 0.8), and loci and individuals with less than 20% of missing data. We then combined univariate and multivariate methods to assess both GEA and GPA, namely Latent Factor Mixed Models (LFMM) and Redundancy Analysis (RDA). While LFMM identifies associations between single loci and single predictors, RDA can detect multilocus signatures of selection as a function of a multivariate set of predictors (Forester *et al.*, 2018; Caye *et al.*, 2019). Both methods have been used extensively (Ahrens *et al.*, 2018), provide a good compromise between detection power and error rates, and are robust to a variety of sampling designs and underlying demographic models (Rellstab *et al.*, 2015; Forester *et al.*, 2018). Since both methods require complete data sets (without missing values), we performed an imputation of missing genotypes (7.6% and 7% missing genotypes for *M. acutistipula*

and *D. apurensis* respectively) based on the *snmf* population assignments from the previous step, using the *impute* function from the *LEA* package (Frichot & François, 2015).

LFMM analysis were performed using the *lfmm* package (Caye *et al.*, 2019) and ridge estimates, which minimize regularized least-squares with a L_2 penalty (see example script here: https://bcm-uga.github.io/lfmm/articles/lfmm). Instead of using raw predictor variables, we employed the first four axes resulting from a Principal Components Analysis (PCA) on all predictor variables in order to minimize the number of tests. We ran LFMM using the previously identified number of genetic clusters (k=3, see results) as latent factors, which accounts for the underlying neutral genetic structure. We then calculated the genomic inflation factor (λ) and modified it until a calibrated distribution of adjusted p-values was found, and set false discovery rates at a rate of q=0.05 using the Benjamini–Hochberg algorithm (François *et al.*, 2016).

We performed RDA using the *rda* function from the *vegan* package (Oksanen *et al.*, 2019) as implemented in Forester *et al.* (2018), modeling genotype as a function of predictor variables, and producing as many constrained axes as predictors (see example script here: https://popgen.nescent.org/2018-03-27_RDA_GEA.html). Multicollinearity between predictors was assessed using the variance inflation factor (VIF) and since all predictor variables showed VIF < 3 none were excluded. Raw predictor variables were scaled and centered prior to analyses and the population assignments from *snmf* (population ID) were used to control for population structure by running a partial RDA. Candidate loci were then identified using a Mahalanobis distance-based approach (Capblancq *et al.*, 2018), which made RDA result comparable with those obtained in LFMM, since it allowed adjusting p-values using the genomic inflation factor (λ) and setting false discovery rates to q=0.05, as described above (calculated and modified genomic inflation factors and p-value distributions for LFMM and RDA tests are provided in Figures S2-S7). Finally, to visualize patterns of GEA and GPA, we ran additional RDA models excluding neutral loci, using the combined candidate adaptive loci detected using RDA and LFMM analyses.

In order to search for the proteins coded by the genes contained in the flanking regions of our candidate SNPs (found in GEA and GPA analyses), contig sequences containing candidate loci were first submitted to the EMBOSS Transeq (http://www.ebi.ac.uk/Tools/st/emboss_transeq/) to obtain corresponding protein sequences. We used all six frames with standard code (codon table), regions (start-end), trimming (yes), and reverse (no). We then ran a functional analysis using InterPro (https://www.ebi.ac.uk/interpro/; interproscan.sh -dp –appl PfamA, TIGRFAM, PRINTS, PrositePatterns, Gene3d –goterms –pathways -f tsv -o MySequences.tsv -i MySequences.faa), searching for gene ontology terms and pathways along a variety of annotation databases (i.e., Interpro, Pfam, Tigrfam, Prints, PrositePattern and Gene3d).

Mapping adaptive genetic variation

To map adaptive genetic variation, we used the *adegenet* package (Jombart & Ahmed, 2011) to run a Spatial Principal Component Analysis (sPCA) on the combined candidate adaptive loci detected using RDA and LFMM analyses in GEA and GPA (results for intersected loci are presented in supporting information). sPCA is a spatially explicit multivariate method that yields scores summarizing genetic variability and spatial structure among individuals (Jombart *et al.*, 2008). Spatial structure is estimated using a Moran's Index that relies on the comparison of allelic frequencies observed in one individual to the values observed in neighboring individuals. These neighboring individuals can be defined by distinct connection networks, which in our case was set to a distance-based neighborhood, as indicated for aggregated distributions (Jombart *et al.*, 2008). The Moran's Index generates two types of spatial structuring: global structure, which reflects positive spatial autocorrelation, and local structure, that reflects negative spatial autocorrelation (Jombart *et al.*, 2008). To decide if global and/or local structures should be interpreted and thus retained in sPCA analyses, we used the global and local tests proposed by Jombart & Ahmed (2011). The first three retained axes where then interpolated on 10 meter resolution grids covering our study area, and the resulting rasters used to create an RGB composite, using the Merge function in QGIS 3.4 (see example scripts here: https://github.com/rojaff/LanGen_pipeline). The resulting color patterns represent the similarity in adaptive genetic composition.

To predict the adaptive genotypes associated with environmental data collected from restoration sites (highly degraded exhausted mine and moderately disturbed Serra da Bocaina, Fig. 2), we employed the GEA-RDA models fitted on the combined candidate adaptive loci detected by LFMM and RDA (see previous section), and ran the *predict.cca* function from the *vegan* package. Environmental samples (soil and climate) from these sites were thus used to predict RDA scores, based on the fitted GEA-RDA models. We then performed a *k*-means cluster analysis (using a Euclidean k-means algorithm) on observed and predicted RDA scores for individuals from both species across the entire study area, using all significant constrained axes from RDA analyses and allowing the number of clusters to vary between two and five. We used the *NbClust* package (Charrad *et al.*, 2014) to obtain the optimal number of clusters chosen by 30 different algorithms.

Results

Genetic diversity and neutral genetic structure

After filtering for quality, read depth, minor allele frequencies, missing data, linkage disequilibrium, Hardy-Weinberg Equilibrium, and outlier loci, we retained 7,376 and 3,496 neutral and independent SNPs and 177 and 163 individuals for *M. acutistipula* and *D. apurensis* respectively, which were then used to assess genetic diversity and population structure. Both genetic clustering approaches (*snmf* and DAPC) indicated the presence of three clusters in the two study species (Fig. S8). Admixture levels were low, and all individuals were correctly assigned to their source Canga highland (Fig. 2). Expected heterozygosity and nucleotide diversity were similar in both species, but inbreeding coefficients were lower and effective population sizes larger in *M. acutistipula* (Table 1). Both species showed significant inbreeding coefficients in all genetic clusters and exhibited the largest effective population sizes in Serra Sul (Table 1). We detected spatial autocorrelation in genetic relatedness within genetic clusters in each species (Fig. S9). In both, the strength of spatial autocorrelation was highest in Serra Norte, where genetic neighborhood size was larger (~10km, Fig. S9).

Genotype-environment and genotype-phenotype associations

The first four principal component axes of the predictor variables explained more than 60% of total variance for both environment and phenotype in both species. These PCA axes were strongly correlated ($|r| > 0.7$) with organic material, B, Fe, Bio06, Bio17, Zn, S and Na (environmental variables), and N/P, P, Fe (phenotypic variables), in both species. After filtering for quality, read depth, minor allele frequencies, missing data, and linkage disequilibrium we retained 9,480 and 4,720 SNPs and 177 and 163 individuals for *M. acutistipula* and *D. apurensis*, respectively. Using LFMM we identified a total of 198 and 154 contigs (RAD-tags) containing GEA, and 94 and 185 contigs containing GPA in *M. acutistipula* and *D. apurensis*, respectively (Tables S1 and S2). Only the first two constrained axes from RDA analyses were significant (ANOVA's $p < 0.05$) in GEA and GPA analyses, so they were retained and used to identify candidate loci in both species (Fig. S10). RDA revealed a total of 403 and 225 contigs containing significant GEA and 281 and 119 contigs containing significant GPA in *M. acutistipula* and *D. apurensis* respectively (Fig. 3 and Tables S1 and S2). Combining both methods (LFMM and RDA), we found a total of 588 contigs showing GEA in *M. acutistipula* and 360 in *D. apurensis*, and 368 contigs showing GPA in *M. acutistipula* and 288 in *D. apurensis*. Only 108 contigs contained both GEA and GPA in *M. acutistipula* and 65 in *D. apurensis.*

Subsequent RDA models using the combined candidate adaptive loci detected using RDA and LFMM analyses revealed population-level patterns of GEA and GPA (Fig. 4). In *M. acutistipula* GEA and GPA models explained 17% and 5% of total variance respectively, while in *D. apurensis*

GEA and GPA models explained 31% and 9% of total variance. In both species, axes loadings were higher for climatic variables (0.01-0.89 for *M. acutistipula* and 0.01-0.83 for *D. apurensis*) than for soil variables (0.005-0.47 for *M. acutistipula* and 0.003-0.59 for *D. apurensis*). In *M. acutistipula*, the first and second axes split individuals into three large GEA groups corresponding to their sampling location. While individuals from Serra Norte showed associations with higher isothermality (bio03) and higher winter temperatures (bio06), individuals from Serra Sul showed associations with lower winter temperatures and higher dry season precipitation (bio17). Individuals from Serra da Bocaina exhibited associations with higher pH (less acidic soils) and lower dry season precipitation (Fig. 4a). Interestingly, *Dioclea apurensis* showed similar GEA patterns based on isothermality (bio03), winter temperatures (bio06), and pH, despite using a slightly different set of predictors (Fig. 4b). On the other hand, the first and second constrained axes divided individuals into two large GPA groups in *M. acutistipula* (Fig. 4c), the first one encompassing individuals from Serra Norte (which showed associations with higher SLA and Mn, and lower P), and the second individuals from Serra Sul and Serra da Bocaina (showing associations with lower SLA and Mn). In *D. apurensis*, the first and second axes split individuals into three GPA groups, with individuals from Serra Norte showing associations with a higher leaf content of Fe and Mn and a lower content of P, while those from Serra Sul showed associations with higher N and those from Serra da Bocaina with lower SLA and N/P (Fig. 4d).

A subset of the contigs containing candidate SNPs showed InterPro annotations (105 contigs in *M. acutistipula* and 59 in *D. apurensis*). Candidate adaptive genes were associated to different functions, including intracellular transport, catalytic activity, synthesis of hormones, metabolic and oxidation-reduction processes, and plant defense response (a full list of candidate genes with InterPro annotations is presented in Table S3). Only 17 putative adaptive genes containing InterPro annotations were shared between both species (Table S4).

Mapping adaptive genetic variation

The combination (union) of candidate adaptive loci detected through GEA and GPA resulted in 914 loci for *M. acutistipula* and 614 loci for *D. apurensis.* Since none of the sPCA local structure tests were significant, we retained the first three positive global axes, which explained the most variance in both species (Fig. S11). These revealed a similar adaptive genetic structure in both species (Fig. 5), with two adaptive units in Serra Norte and one in Serra da Bocaina. *Mimosa acutistipula* nevertheless exhibited a clinal adaptive pattern in Serra Sul, whereas *D. apurensis* did not. Similar spatial patterns were found when using the intersected loci (i.e. those shared by GEA and GPA; Fig. S12). Finally, predicted genotypes associated with the environmental samples collected from a

highly degraded mining site did not cluster together with any of our study populations in either species (Fig. 6a and 6b). In contrast, most predicted genotypes for environmental samples from the moderately disturbed Serra da Bocaina clustered together with individuals collected in the same location, revealing local provenances are best adapted to local environmental conditions (Fig. 6c and 6d).

Discussion

The delineation of seed sourcing areas requires accounting for evolutionary history, genetic diversity, and how likely individuals will adapt to the environmental conditions of the targeted restoration sites (Breed *et al.*, 2019). Here we employ a comprehensive landscape genomic approach to characterize neutral and adaptive genetic variation and provide insights to assist the restoration of a highly degraded mining site and a moderately disturbed Canga highland from the Carajás Mineral Province. We discuss how our results can help define site-adjusted provenancing strategies and argue that our methods can be more broadly applied to assist other restoration and rehabilitation initiatives.

Several studies have stressed the importance of avoiding inbreeding, increasing genetic diversity to maintain evolutionary potential, and minimizing outbreeding depression in restored populations (Broadhurst *et al.*, 2008; Weeks *et al.*, 2011; Hufbauer *et al.*, 2015; Mijangos *et al.*, 2015). The assessment of neutral genetic structure provides information on how to minimize outbreeding depression by avoiding mixing individuals from different evolutionary lineages (Mijangos *et al.*, 2015). Estimates of the genetic neighborhood size, on the other hand, provide clues on how to sample unrelated individuals within seed sourcing areas to increase genetic diversity and reduce the risk of inbreeding depression in restored populations (Krauss & Koch, 2004; Breed *et al.*, 2019). Our initial assessment of neutral genetic structure identified three demographically independent units (or Management Units *sensu* (Funk *et al.*, 2012)), which could be considered distinct seed sourcing zones to minimize the risk of outbreeding depression (Frankham *et al.*, 2017). Within these zones, our estimates of genetic neighborhood size provide information on within-cluster seed sourcing strategies to maximize genetic diversity. In Serra Sul, for example, seed sources located 5 Km apart are not expected to be related (Fig. S9), and would comprise a better representation of standing genetic variation than individuals collected across smaller spatial scales. Effective population size estimates (Table 1) nevertheless indicate that none of our observed genetic clusters is likely to experience inbreeding depression in the near future based on the 50/500 rule (Jamieson & Allendorf, 2012). Inbreeding levels observed in both species

were nonetheless significantly different from zero, suggesting some level of selfing or mating between related individuals is taking place.

Patterns of local adaptation will ultimately determine the ability of plants to effectively colonize and quickly recover disturbed sites (Mijangos *et al.*, 2015). By using the combined candidate loci detected in GEA and GPA, using both univariate and multivariate methods, we improved the detection of single-locus and multi-locus adaptive signals, circumventing the need to carry out time-consuming common garden experiments. (Talbot *et al.*, 2016; Vangestel *et al.*, 2018; Mahony *et al.*, 2019). Interestingly, more intersections between GEA and GPA were found when using RDA than when using LFMM (Fig. 3), indicating that most adaptations to local environmental conditions expressing differential phenotypes are polygenic (Forester *et al.*, 2018). Indeed most fitness-related traits in plants have a polygenic basis (Falke *et al.*, 2013), including tolerance to soil with phytotoxic levels of heavy metals (Arnold *et al.*, 2016). Interestingly, the most frequent amongst our identified candidate proteins are involved in plant defense and stress responses (including reverse transcriptase, ribonuclease H-like domain, P-loop NTPase fold, leucine-rich repeat, and thaumatin) or basic metabolic processes (pentatricopeptide repeat, protein kinase domain, Nitrite/Sulfite reductase ferredoxin-like domain, major intrinsic protein, and kinesin motor domain), suggestive of adaptations to harsh Canga environments (Tables S3 and S4). Some of these putative adaptive genes were shared between the two species as well as with other co-occurring species (Table S4), indicating convergent evolution to similar environmental pressures (Arnold *et al.*, 2016; Yeaman *et al.*, 2016). These shared genes thus constitute primary targets for functional studies investigating the molecular basis of adaptation to Canga environments and minelands.

To visualize and better understand the mechanisms behind the observed GEA and GPA we ran additional RDA analyses using the combined candidate adaptive loci detected by LFMM and RDA. As expected, we found similar patterns of GEA in both species (Figs. 4a and 4b). Interestingly, the strongest GEA were found with climatic variables in both species, in spite of the coarse resolution of WorldClim data and the narrow climatic variation found across our study area (Fig. S1). Our results thus suggest that local climate constitutes an important environmental filter driving local adaptation, as found in other species from Canga environments (Lanes *et al.*, 2018) and temperate climates (Pluess *et al.*, 2016; Pais *et al.*, 2017). In *M. acutistipula,* Serra Norte populations showed associations with higher SLA, suggesting climatic or soil conditions in Serra Norte are more favorable to plant growth (He *et al.*, 2018). Individuals from Serra da Bocaina and Serra Sul showed associations with lower SLA and lower concentration of several micro- and macro-nutrients, suggesting that increasing leaf thickness in these individuals avoids dissection or better preserves scarce nutrients (Costa-Saura *et al.*, 2016). In contrast, SLA-associations in *D. apurensis*

did not separate Canga highlands, showing that the influence of climatic variation on SLA is different across species (Liu *et al.*, 2017; Gong & Gao, 2019). In *D. apurensis*, different genotype associations with leaf micro and macronutrients separated highlands (Fig. 4d), suggesting different physiological requirements or nutrient availability at each site. Low correlations between leaf and soil Fe and Mn concentrations (Pearson's correlation coefficients ranged between 0.02 and 0.20), suggest our study species are controlling nutrient absorption, which makes them suitable for the restoration of areas with a high concentration of these metals. Controlled common-garden or reciprocal transplant experiments are nevertheless needed to assess growth and overall performance of different genotypes (sources) in different soils and climates (Rellstab *et al.*, 2015; Aitken & Bemmels, 2016).

Our local adaptation maps reveal areas containing similar local adaptations (colors) in each species (Fig. 5, Fig. S12), which could be used to delineate seed sourcing strategies. In contrast to the commonly employed Generalized Dissimilarity Models (GDM) (Shryock *et al.*, 2015; Gugger *et al.*, 2018; Supple *et al.*, 2018; Rossetto *et al.*, 2019), our mapping approach based on sPCA allows incorporating GPA and predicting adaptive genetic variation from site-level data, which is particularly useful in areas lacking high-resolution environmental layers. Moreover, sPCA explicitly account for spatial autocorrelation in genetic composition, which is likely to play an important role explaining patterns of local adaptation (Lesica & Allendorf, 1999b; Richardson *et al.*, 2014) (see Fig. S13 for alternative adaptation maps generated using GDM). Our adaptation maps showed similar adaptations across Serra da Bocaina (i.e. a single adaptive unit), and most predicted genotypes associated with the environmental samples collected there clustered together with individuals sampled in the same location. This result indicates that local provenances are probably best adapted to local environmental conditions at this site (Fig. 6c and 6d), and supports the recommendations made by Lesica & Allendorf (1999) for restoration of large and moderately disturbed sites. Since genetic neighborhood size in Serra da Bocaina was roughly 3 Km for both species, our results suggest that local seeds collected in areas separated by at least 3 Km would maximize genetic diversity at this location.

In contrast, predicted genotypes for the environmental data collected at the degraded mine site did not cluster with any of our study populations in either species (Fig. 6a and 6b). This indicates that none of the genotypes we sampled from natural habitats overlap with the multivariate environmental conditions present at the mine site. In this case, mixing genotypes containing different local adaptations could be regarded as the best option to maximize evolutionary potential and facilitate adaptation to novel environments (Lesica & Allendorf, 1999b). Seeds could be sourced from all the identified adaptive units (colors from sPCA maps); and within these units they

could be sampled in areas separated by the genetic neighborhood size to further enhance genetic diversity. Although mixing individuals from different management units could result in outbreeding depression (Hufford & Mazer, 2003; Weeks *et al.*, 2011), the risk is likely marginal for these study species, which are widely distributed across the continent (Dutra & Morim, 2015; Queiroz, 2015). Moreover, environmental conditions and plant communities show remarkable similarities across the Carajás Mineral Province when compared to other *campo rupestre* formations (Zappi *et al.*, 2019). Such a regional admixture provenancing approach (Bucharova *et al.*, 2019) would represent a significant improvement over introducing exotic species, which are currently being used in mine reclamation programs due to the availability of seeds in large quantities and their ability to quickly colonize mine environments to prevent soil erosion (D'Antonio & Meyerson, 2002; Silva *et al.*, 2018; Gastauer *et al.*, 2019). Our results could guide the establishment of seed production areas for both native species, aiming at overcoming shortfalls in seed availability while capturing standing neutral and adaptive genetic variation (Nevill *et al.*, 2016).

Our work illustrates how neutral and adaptive genetic variation can be used to provide evidence-based recommendations for provenance schemes aimed to effectively restore moderately disturbed and highly degraded sites. In our two study species, local provenances where found optimal to restore a moderately disturbed site, whereas a mixture of genotypes was suggested as the most promising strategy to recover a highly degraded mining site, to which local provenances were not adapted. Our proposed methodological approach (Fig. 1) can be more broadly applied to define site-adjusted provenance strategies in other locations and for other disturbance regimes. As genomic data becomes available for more species exhibiting different life-history characteristics, restoration genomic initiatives using similar approaches are likely to substantially improve restoration outcomes.

Acknowledgments

Funding was provided by Instituto Tecnológico Vale, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) grants 301616/2017-5 (RJ), 153535/2018-0 (MG) and 316067/2018-0 (LCRM), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) grant 88887.156652/2017-00 (CSC). We thank Cesar Neto and Eder Lanes for assistance in the field, Nelson Carvalho for assistance in genome size estimation and Eder Lanes and Manoel Lopes for help in the laboratory.

Author contributions

RJ conceived, designed and coordinated the project. CSC, MG and RJ coordinated the field work and sampling. CSC, SM and MG performed laboratory work. CSC, LCRM, LT, BF, MG and RJ performed the data analysis. The first draft of the paper was written by CSC and RJ and all authors contributed to discussing the results and editing the paper.

Conflict of interest

The authors declare no competing interests.

Data Availability Statement

Geographic coordinates, genotypes in Variant Call Format, and sequences in FASTA format for both species will be deposited in Figshare and a url address provided upon the acceptance of this manuscript. All the mentioned R scripts have been deposited in GitHub and their url addresses provided in the text.

References

Ahrens CW, Rymer PD, Stow A, Bragg J, Dillon S, Umbers KDL, Dudaniec RY. **2018**. The search for loci under selection: trends, biases and progress. *Molecular Ecology* **27**: 1342–1356.

Aitken SN, Bemmels JB. **2016**. Time to get moving: Assisted gene flow of forest trees. *Evolutionary Applications* **9**: 271–290.

Allendorf FW, Luikart GH, Aitken SN. **2013**. *Conservation and the genetics of populations*. Oxford: Wiley-Blackwell.

Arnold BJ, Lahner B, DaCosta JM, Weisman CM, Hollister JD, Salt DE, Bomblies K, Yant L. **2016**. Borrowed alleles and convergence in serpentine adaptation. *Proceedings of the National Academy of Sciences* **113**: 8320–8325.

Balkenhol N, Dudaniec RY, Krutovsky K V, Johnson JS, Cairns DM, Segelbacher G, Selkoe KA, von der Heyden S, Wang IJ, Selmoni O, *et al.* 2017. Landscape Genomics: Understanding Relationships Between Environmental Heterogeneity and Genomic Characteristics of Populations. In: Cham: Springer International Publishing, 261–322.

Bothe H. **2011**. Plants in Heavy Metal Soils. In: Detoxification of Heavy Metals. 35–57.

Boyer S, Wratten SD. **2010**. The potential of earthworms to restore ecosystem services after opencast mining - A review. *Basic and Applied Ecology* **11**: 196–203.

Breed MF, Harrison PA, Blyth C, Byrne M, Gaget V, Gellie NJC, Groom SVC, Hodgson R, Mills JG, Prowse TAA, *et al.* 2019. The potential of genomics for restoring ecosystems and biodiversity. *Nature Reviews Genetics*.

Broadhurst LM, Lowe A, Coates DJ, Cunningham SA, McDonald M, Vesk PA, Yates C. **2008**. Seed supply for broadscale restoration: Maximizing evolutionary potential. *Evolutionary Applications* **1**: 587–597.

Bucharova A, Bossdorf O, Hölzel N, Kollmann J, Prasse R, Durka W. **2019**. Mix and match: regional admixture provenancing strikes a balance among different seed-sourcing strategies for ecological restoration. *Conservation Genetics* **20**: 7–17.

Capblancq T, Luu K, Blum MGB, Bazin E. **2018**. Evaluation of redundancy analysis to identify signatures of local adaptation. *Molecular Ecology Resources* **18**: 1223–1233.

Carvalho C da S, Lanes ECM, Silva AR, Caldeira CF, Carvalho-Filho N, Gastauer M, Imperatriz-Fonseca VL, Nascimento W, Oliveira G, Siqueira JO, *et al.* 2019. Habitat loss does not always entail negative genetic consequences. *Frontiers in Genetics* **10**.

Caye K, Jumentier B, Lepeule J, François O. **2019**. LFMM 2: fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular Biology and Evolution* **36**: 852–860.

Charrad M, Ghazzali N, Boiteux V, Niknafs A. **2014**. NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* **61**: 1–36.

Coates DJ, Byrne M, Moritz C. **2018**. Genetic Diversity and Conservation Units: Dealing With the Species-Population Continuum in the Age of Genomics. *Frontiers in Ecology and Evolution* **6**: 1–13.

Costa-Saura JM, Martínez-Vilalta J, Trabucco A, Spano D, Mereu S. **2016**. Specific leaf area and hydraulic traits explain niche segregation along an aridity gradient in Mediterranean woody species. *Perspectives in Plant Ecology, Evolution and Systematics* **21**: 23–30.

D'Antonio C, Meyerson LA. **2002**. Exotic Plant Species as Problems and Solutions in Ecological Restoration: A Synthesis. *Restoration Ecology* **10**: 703–713.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Do C, Waples RS, Peel D, Macbeth GM, Tillett BJ, Ovenden JR. **2014**. NeEstimator v2: Re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. *Molecular Ecology Resources* **14**: 209–214.

Durka W, Michalski SG, Berendzen KW, Bossdorf O, Bucharova A, Hermann J-M, Hölzel N, Kollmann J. **2017**. Genetic differentiation within multiple common grassland plants supports seed transfer zones for ecological restoration (S Wan, Ed.). *Journal of Applied Ecology* **54**: 116–126.

Dutra VF, Morim MP. **2015**. Mimosa in Lista de Espécies da Flora do Brasil. Jardim Botânico do Rio de Janeiro. *http://floradobrasil.jbrj.gov.br/jabot/floradobrasil/FB116253*.

Falke KC, Glander S, He F, Hu J, de Meaux J, Schmitz G. **2013**. The spectrum of mutations controlling complex traits and the genetics of fitness in plants. *Current Opinion in Genetics & Development* **23**: 665–671.

Forester BR, Lasky JR, Wagner HH, Urban DL. **2018**. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology* **27**: 2215–2233.

François O, Martins H, Caye K, Schoville SD. **2016**. Controlling false discoveries in genome scans for selection. *Molecular Ecology* **25**: 454–469.

Frankham R, Ballou JD, Ralls K, Eldridge M, Dudash MR, Fenster CB, Lacy RC, Sunnucks P. **2017**. *Genetic management of fragmented animal and plant populations*. Oxford University Press.

Frichot E, François O. **2015**. LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution* **6**: 925–929.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. **2014**. Fast and Efficient Estimation of Individual Ancestry Coefficients. *Genetics* **196**: 973–983.

Funk WC, McKay JK, Hohenlohe PA, Allendorf FW. **2012**. Harnessing genomics for delineating conservation units. *Trends in Ecology & Evolution* **27**: 489–496.

Garris HW, Baldwin SA, Van Hamme JD, Gardner WC, Fraser LH. **2016**. Genomics to assist mine reclamation: A review. *Restoration Ecology* **24**: 165–173.

Gastauer M, Souza Filho PWM, Ramos SJ, Caldeira CF, Silva JR, Siqueira JO, Furtini Neto AE. **2019**. Mine land rehabilitation in Brazil: Goals and techniques in the context of legal requirements. *Ambio* **48**: 74–88.

Giannini TC, Giulietti AM, Harley RM, Viana PL, Jaffe R, Alves R, Pinto CE, Mota NFO, Caldeira Jr CF, Imperatriz-Fonseca VL, *et al.* 2017. Selecting plant species for practical restoration of degraded lands using a multiple-trait approach. *Austral Ecology* **42**: 510–521.

Gong H, Gao J. **2019**. Soil and climatic drivers of plant SLA (specific leaf area). *Global Ecology and Conservation* **20**: e00696.

Gugger PF, Liang CT, Sork VL, Hodgskiss P, Wright JW. **2018**. Applying landscape genomic tools to forest management and restoration of Hawaiian koa (Acacia koa) in a changing environment. *Evolutionary Applications* **11**: 231–242.

He D, Chen Y, Zhao K, Cornelissen JHC, Chu C. **2018**. Intra- and interspecific trait variations reveal functional relationships between specific leaf area and soil niche within a subtropical forest. *Annals of Botany* **121**: 1173–1182.

Hufbauer RA, Szucs M, Kasyon E, Youngberg C, Koontz MJ, Richards C, Tuff T, Melbourne BA. **2015**. Three types of rescue can avert extinction in a changing environment. *Proceedings of the National Academy of Sciences* **112**: 10557–10562.

Hufford KM, Mazer SJ. **2003**. Plant ecotypes: Genetic differentiation in the age of ecological restoration. *Trends in Ecology and Evolution* **18**: 147–155.

Jamieson IG, Allendorf FW. **2012**. How does the 50 / 500 rule apply to MVPs? *Trends in Ecology & Evolution* **27**: 578–584.

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

Jombart T, Devillard S, Dufour A-B, Pontier D. **2008**. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity* **101**: 92–103.

Josse J, Husson F. **2016**. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software* **70**: 1–31.

Krauss SL, Koch JM. **2004**. Rapid genetic delineation of provenance for plant community restoration. *Journal of Applied Ecology* **41**: 1162–1173.

Krauss SL, Sinclair EA, Bussell JD, Hobbs RJ. **2013**. An ecological genetic delineation of local seed-source provenance for ecological restoration. *Ecology and Evolution* **3**: 2138–2149.

Lanes ÉC, Pope NS, Alves R, Carvalho Filho NM, Giannini TC, Giulietti AM, Imperatriz-Fonseca VL, Monteiro W, Oliveira G, Silva AR, *et al.* 2018. Landscape Genomic Conservation Assessment of a Narrow-Endemic and a Widespread Morning Glory From Amazonian Savannas. *Frontiers in Plant Science* **9**: 532.

Lesica P, Allendorf FW. **1999a**. Ecological Genetics and the Restoration of Plant Communities: Mix or Match? *Restoration Ecology* **7**: 42–50.

Lesica P, Allendorf FW. **1999b**. Ecological Genetics and the Restoration of Plant Communities: Mix or Match? *Restoration Ecology* **7**: 42–50.

Liu M, Wang Z, Li S, Lü X, Wang X, Han X. **2017**. Changes in specific leaf area of dominant plants in temperate grasslands along a 2500-km transect in northern China. *Scientific Reports* **7**: 10780.

Lu M, Krutovsky K V, Loopstra CA. **2019**. Predicting adaptive genetic variation of loblolly pine (Pinus taeda L.) populations under projected future climates based on multivariate models. *Journal of Heredity*.

Mahony CR, MacLachlan IR, Lind BM, Yoder JB, Wang T, Aitken SN. **2019**. Evaluating genomic data for management of local adaptation in a changing climate: A lodgepole pine case study. *Evolutionary Applications*: eva.12871.

Martins K, Gugger PF, Llanderal-Mendoza J, González-Rodríguez A, Fitz-Gibbon ST, Zhao JL, Rodríguez-Correa H, Oyama K, Sork VL. **2018**. Landscape genomics provides evidence of climate-associated genetic variation in Mexican populations of Quercus rugosa. *Evolutionary Applications* **11**: 1842–1858.

McKinney GJ, Larson WA, Seeb LW, Seeb JE. **2017**. RADseq provides unprecedented insights into molecular ecology and evolutionary genetics: comment on Breaking RAD by Lowry et al . (2016). *Molecular Ecology Resources* **17**: 356–361.

Mijangos JL, Pacioni C, Spencer PBS, Craig MD. **2015**. Contribution of genetics to ecological restoration. *Molecular Ecology* **24**: 22–37.

Mitre SK, Mardegan SF, Caldeira CF, Ramos SJ, Furtini Neto AE, Siqueira JO, Gastauer M. **2018**. Nutrient and water dynamics of Amazonian canga vegetation differ among physiognomies and from those of other neotropical ecosystems. *Plant Ecology* **219**: 1341–1353.

Mota NF de O, Silva LVC, Martins FD, Viana PL. **2015**. Vegetação sobre sistemas ferruginosos da Serra dos Carajás. *Geossistemas Ferruginosos no Brasil. Instituto Pr{\'\i}stino, Belo Horizonte*: 289–315.

Mota NF de O, Watanabe MTC, Zappi DC, Hiura AL, Pallos J, Viveros RS, Giulietti AM, Viana PL. **2018**. Cangas da Amazônia: a vegetação única de Carajás evidenciada pela lista de fanerógamas. *Rodriguésia* **69**: 1435–1488.

Nevill PG, Tomlinson S, Elliott CP, Espeland EK, Dixon KW, Merritt DJ. **2016**. Seed production areas for the global restoration challenge. *Ecology and Evolution* **6**: 7490–7497.

Nunes JA, Schaefer CEGR, Júnior WGF, Neri A V., Correa GR, Enright NJ. **2015**. Soil-vegetation relationships on a banded ironstone 'island', Carajás Plateau, Brazilian Eastern Amazonia. *Anais da Academia Brasileira de Ciencias* **87**: 2097–2110.

Oksanen J, Blanchet F, Friendly, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara R, Simpson G, Solymos P, *et al.* 2019. vegan: Community Ecology Package.

Pais AL, Whetten RW, Xiang Q-YJ. **2017**. Ecological genomics of local adaptation in Cornus florida L. by genotyping by sequencing. *Ecology and Evolution* **7**: 441–465.

Pérez-Harguindeguy N, Diaz S, Gamier E, Lavorel S, Poorter H, Jaureguiberry P, Bret-Harte MS, Comwell WK, Craine JM, Gurvich DE, *et al.* 2013. New handbook for stand-ardised measurement of plant functional traits worldwide. *Australian Journal of Botany* **61**: 167–234.

Pluess AR, Frank A, Heiri C, Lalagüe H, Vendramin GG, Oddou-Muratorio S. **2016**. Genome-environment association study suggests local adaptation to climate at the regional scale in Fagus sylvatica. *New Phytologist* **210**: 589–601.

Poveromo JJ. **1999**. Iron Ores. In: The Making, Shaping, and Treating of Steel: Ironmaking volume. Pittsburg, PA: The AISE Steel Foundation, 547–550.

Preite V, Sailer C, Syllwasschy L, Bray S, Ahmadi H, Krämer U, Yant L. **2019**. Convergent evolution in Arabidopsis halleri and Arabidopsis arenosa on calamine metalliferous soils. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**.

Prober SM, Byrne M, McLean EH, Steane DA, Potts BM, Vaillancourt RE, Stock WD. **2015**. Climate-adjusted provenancing: a strategy for climate-resilient ecological restoration. *Frontiers in Ecology and Evolution* **3**.

Queiroz LP. **2015**. Dioclea in Lista de Espécies da Flora do Brasil. Jardim Botânico do Rio de Janeiro. *http://floradobrasil.jbrj.gov.br/jabot/floradobrasil/FB83091*.

Ramos SJ, Caldeira CF, Gastauer M, Costa DLP, Furtini Neto AE, de Souza FBM, Souza-Filho PWM, Siqueira JO. **2019a**. Native leguminous plants for mineland revegetation in the eastern Amazon: seed characteristics and germination. *New Forests*.

Ramos SJ, Gastauer M, Mitre SK, Caldeira CF, Silva JR, Furtini Neto AE, Oliveira G, Souza Filho PWM, Siqueira JO. **2019b**. Plant growth and nutrient use efficiency of two native Fabaceae species for mineland revegetation in the eastern Amazon. *Journal of Forestry Research*.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R. **2015**. A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology* **24**: 4348–4370.

Richardson JL, Urban MC, Bolnick DI, Skelly DK. **2014**. Microgeographic adaptation and the spatial scale of evolution. *Trends in Ecology and Evolution* **29**: 165–176.

Rossetto M, Bragg J, Kilian A, McPherson H, van der Merwe M, Wilson PD. **2019**. Restore and Renew: a genomics-era framework for species provenance delimitation. *Restoration Ecology* **27**: 538–548.

Shryock DF, Havrilla CA, DeFalco LA, Esque TC, Custer NA, Wood TE. **2015**. Landscape genomics of Sphaeralcea ambigua in the Mojave Desert: a multivariate, spatially-explicit approach to guide ecological restoration. *Conservation Genetics* **16**: 1303–1317.

Shryock DF, Havrilla CA, DeFalco LA, Esque TC, Custer NA, Wood TE. **2017**. Landscape genetic approaches to guide native plant restoration in the Mojave Desert: *Ecological Applications* **27**: 429–445.

Silva JR, Gastauer M, Ramos SJ, Mitre SK, Furtini Neto AE, Siqueira JO, Caldeira CF. **2018**. Initial growth of Fabaceae species: Combined effects of topsoil and fertilizer application for mineland revegetation. *Flora: Morphology, Distribution, Functional Ecology of Plants* **246–247**: 109–117.

Skirycz A, Castilho A, Chaparro C, Carvalho N, Tzotzos G, Siqueira JO. **2014**. Canga biodiversity, a matter of mining. *Frontiers in Plant Science* **5**: 653.

Souza-Filho PWM, Giannini TC, Jaffé R, Giulietti AM, Santos DC, Nascimento Jr. WR, Guimarães JTF, Costa MF, Imperatriz- Fonseca VL, Siqueira JO. **2019**. Mapping and quantification of ferruginous outcrop savannas in the Brazilian Amazon: A challenge for biodiversity conservation. *PLOS ONE* **14**: e0211095.

Steane DA, Potts BM, McLean E, Prober SM, Stock WD, Vaillancourt RE, Byrne M. **2014**. Genome-wide scans detect adaptation to aridity in a widespread forest tree species. *Molecular Ecology* **23**: 2500–2513.

Supple MA, Bragg JG, Broadhurst LM, Nicotra AB, Byrne M, Andrew RL, Widdup A, Aitken NC, Borevitz JO. **2018**. Landscape genomic prediction for restoration of a Eucalyptus foundation species under climate change. *eLife* **7**: 1–22.

Talbot B, Chen T-W, Zimmerman S, Joost S, Eckert AJ, Crow TM, Semizer-Cuming D, Seshadri C, Manel S. **2016**. Combining Genotype, Phenotype, and Environment to Infer Potential Candidate Genes. *Journal of Heredity* **108**: esw077.

Vangestel C, Eckert AJ, Wegrzyn JL, St. Clair JB, Neale DB. **2018**. Linking phenotype, genotype and environment to unravel genetic components underlying cold hardiness in coastal Douglas-fir (Pseudotsuga menziesii var. menziesii). *Tree Genetics & Genomes* **14**: 10.

Weeks AR, Sgro CM, Young AG, Frankham R, Mitchell NJ, Miller KA, Byrne M, Coates DJ, Eldridge MDB, Sunnucks P, *et al.* 2011. Assessing the benefits and risks of translocations in changing environments: A genetic perspective. *Evolutionary Applications* **4**: 709–725.

Whiting SN, Reeves RD, Richards D, Johnson MS, Cooke JA, Malaisse F, Paton A, Smith JAC, Angle JS, Chaney RL, *et al.* 2004. Research Priorities for Conservation of Metallophyte Biodiversity and their Potential for Restoration and Site Remediation. *Restoration Ecology* **12**: 106–116.

Williams A V., Nevill PG, Krauss SL. **2014**. Next generation restoration genetics: Applications and opportunities. *Trends in Plant Science* **19**: 529–537.

Xuereb A, Kimber CM, Curtis JMR, Bernatchez L, Fortin MJ. **2018**. Putatively adaptive genetic variation in the giant California sea cucumber (Parastichopus californicus) as revealed by environmental association analysis of restriction-site associated DNA sequencing data. *Molecular Ecology* **27**: 5035–5048.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, *et al.* 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42**: 565–569.

Yeaman S, Hodgins KA, Lotterhos KE, Suren H, Nadeau S, Degner JC, Nurkowski KA, Smets P, Wang T, Gray LK, *et al.* 2016. Convergent local adaptation to climate in distantly related conifers. *Science* **353**: 1431–1433.

Zappi DC, Moro MF, Walker B, Meagher T, Viana PL, Mota NFO, Watanabe MTC, Nic Lughadha E. **2019**. Plotting a future for Amazonian canga vegetation in a campo rupestre context (N Cellinese, Ed.). *PLOS ONE* **14**: e0219753.

Tables

Table 1: Genetic diversity measures for *Mimosa acutistipula* var. *ferrea* and *Dioclea apurensis* within genetic clusters (highlands). Sample sizes (*N*) are followed by mean expected heterozygosity (H_E), mean inbreeding coefficient (*F*), nucleotide diversity (π) and effective population size (N_e). Values represent 95% confidence intervals.

Species	Highland	N	H_E	F	π	N_e
M. acutistipula	Serra Sul	80	[0.28/0.28]	[0.12/0.15]	[0.26/0.26]	[1601.0/1778.0]
	Serra Norte	61	[0.28/0.28]	[0.10/0.14]	[0.23/0.24]	[554.7/581.2]
	Serra da Bocaina	36	[0.28/0.28]	[0.12/0.17]	[0.25/0.26]	[1008.0/1243.5]
D. apurensis	Serra Sul	81	[0.25/0.25]	[0.10/0.14]	[0.22/0.23]	[1070.1/1278.5]
	Serra Norte	45	[0.29/0.29]	[0.14/0.20]	[0.24/0.25]	[193.2/204.6]
	Serra da Bocaina	37	[0.27/0.27]	[0.22/0.31]	[0.21/0.22]	[193.7/212.2]

Figures

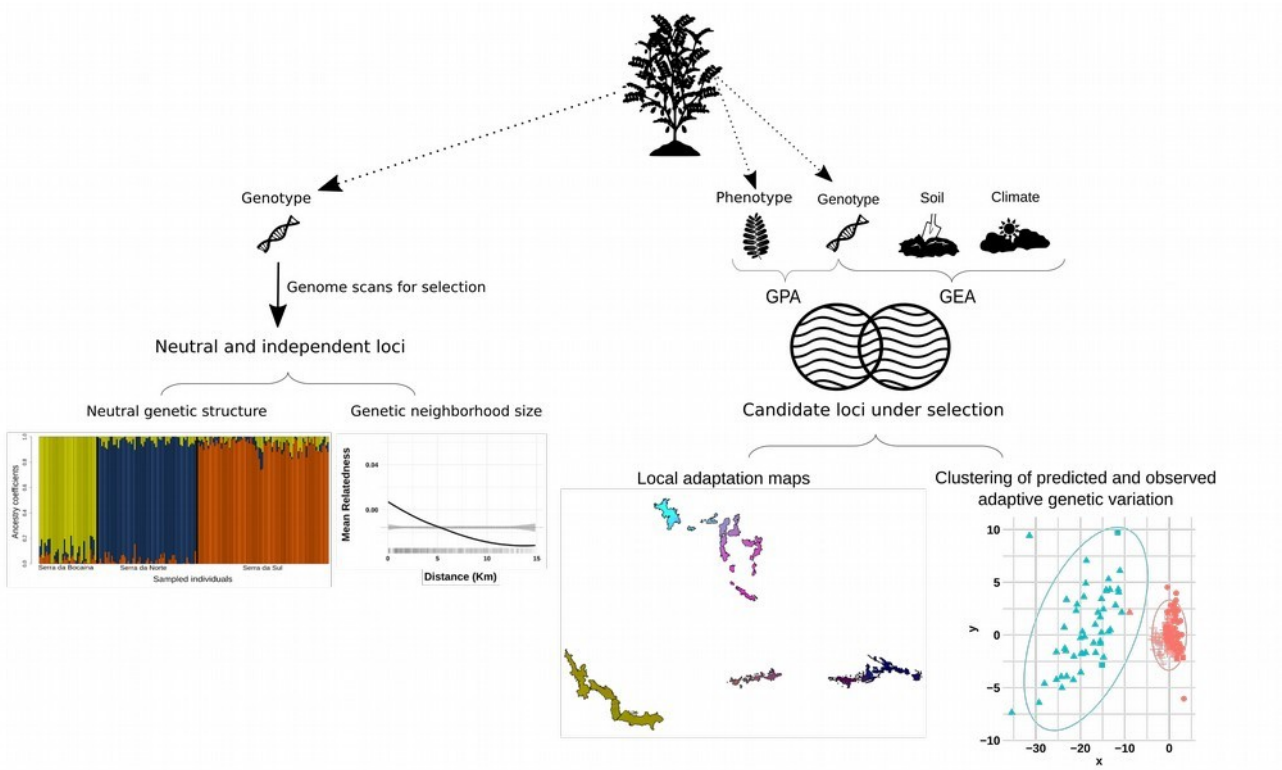

Figure 1: Diagram summarizing our methodological approach. We used a subset of neutral and independent loci to assess broad and fine-scale genetic structure and determine the genetic neighborhood size. Subsequently, we combined genotype, phenotype, and environmental data to identify loci under selection and then employed all candidate loci to map patterns of adaptive genetic variability and predict adaptive genotypes for restoration sites.

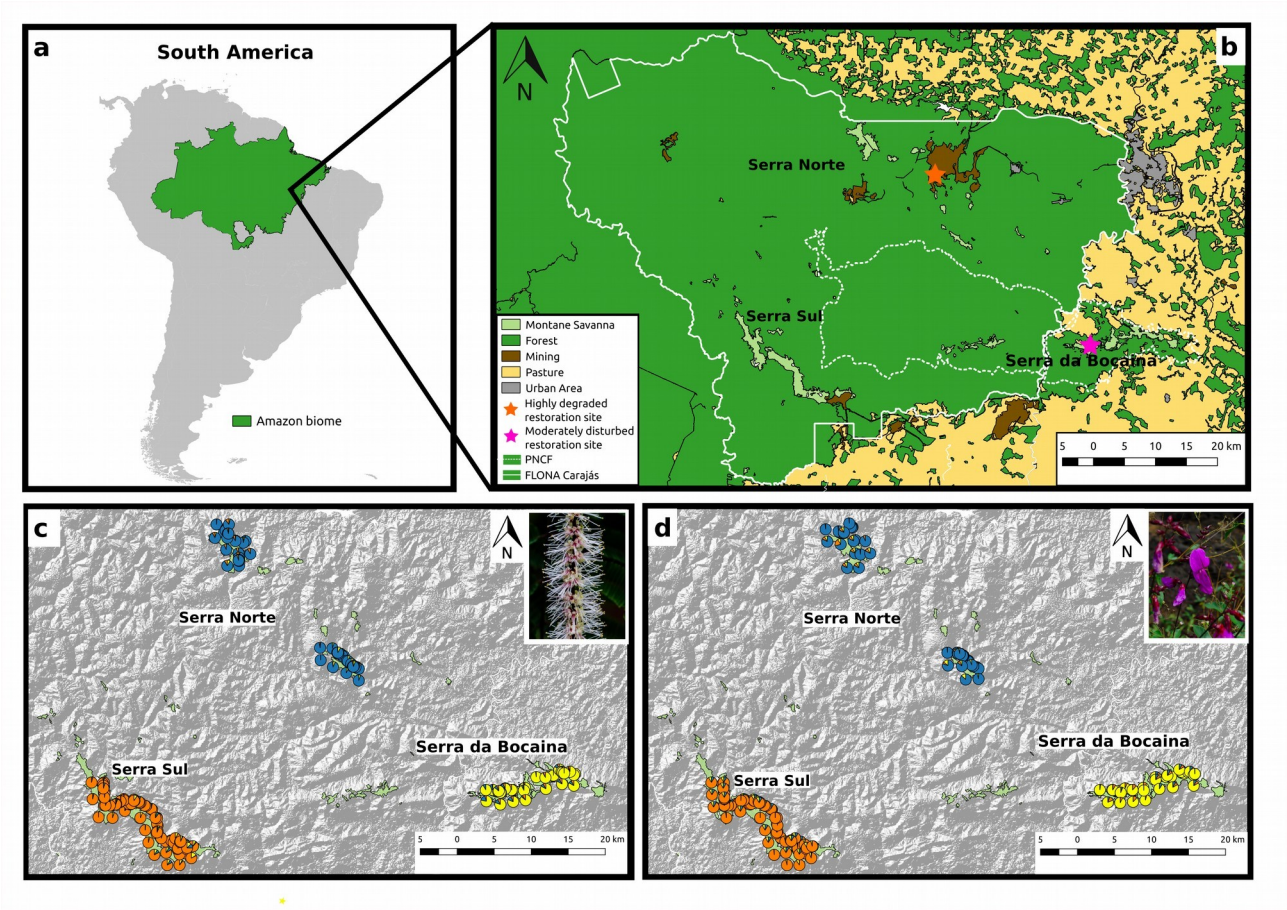

Figure 2: Maps of our study region depicting restoration sites and neutral genetic structure for *Mimosa acutistipula* var. *ferrea* and *Dioclea apurensis*. **a**: Location of the Carajás Mineral Province within the Amazon biome. **b:** Location of the major Canga highlands (Montane Savanna) where samples were collected. White continuous lines represent the Carajás National Forest (FLONA Carajás), white dashed lines the Campos Ferruginosos National Park (PNCF), and stars depict restoration sites. Ancestry coefficients for all samples from *M. acutistipula* var. *ferrea* (**c**) and *D. apurensis* (**d**), determined using the *snmf* function from the LEA package.

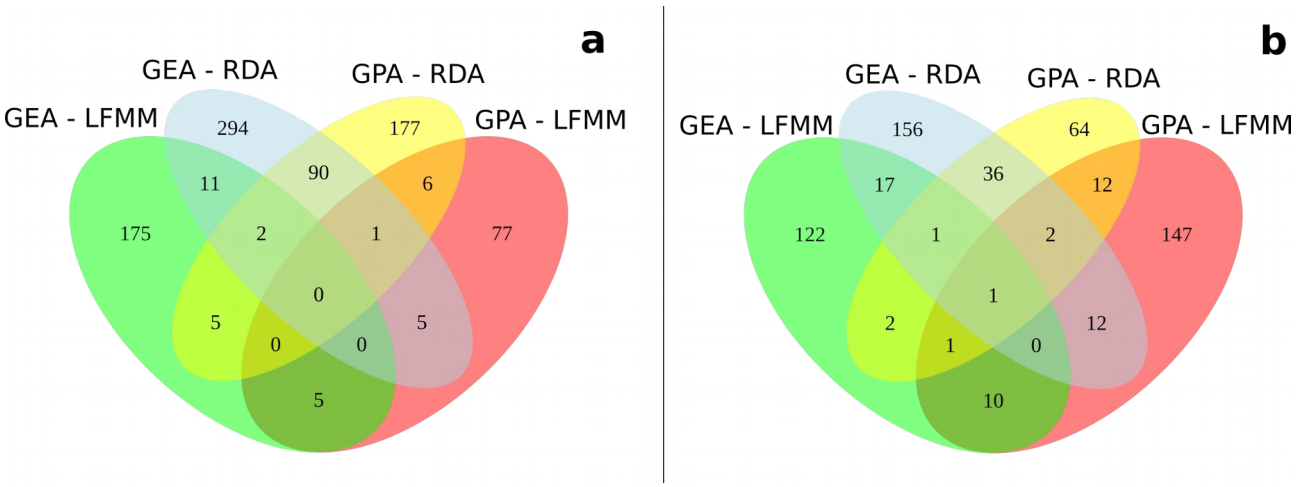

Figure 3: Venn diagram showing the intersection of contigs (RAD-tags) containing candidate SNPs for *Mimosa acutistipula* var. *ferrea* (**a**) and *Dioclea apurensis* (**b**). Genotype-environment associations (GEA) and genotype-phenotype association (GPA) were assessed using Redundant Analysis (RDA) and Latent Factor Mixed Model (LFMM).

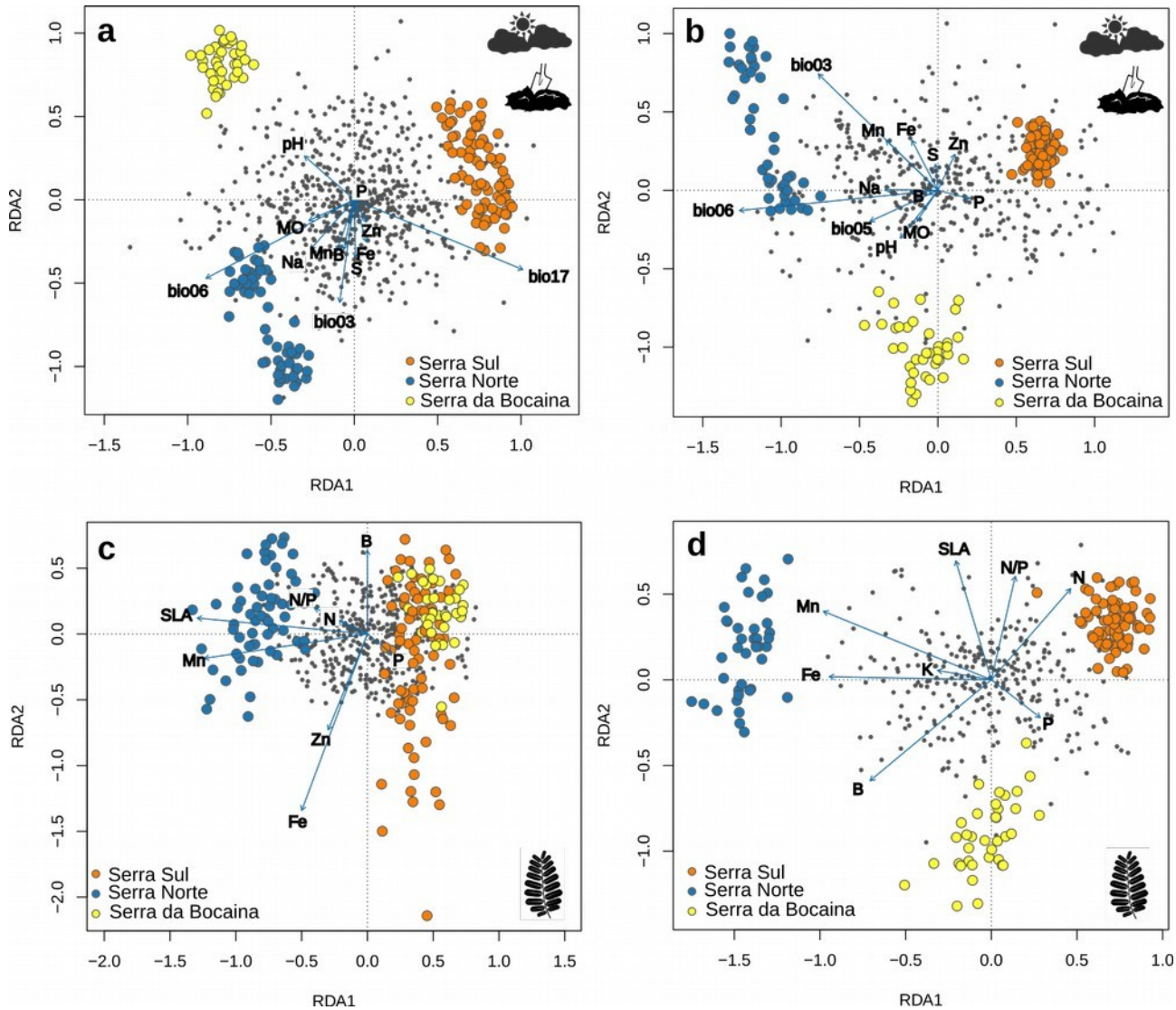

Figure 4: Redundancy analysis (RDA) using only the candidate loci identified in genotype-environment (upper panels) and genotype-phenotype association analyses (lower panels) in *Mimosa acutistipula* var. *ferrea* (**a**, **c**) and *Dioclea apurensis* (**b**, **d**). The plots show the first and second constrained axes from RDA, with SNPs represented as gray filled circles, environment and phenotype variables as blue arrows, and individuals from different Canga highlands in color circles (bio03: isothermality; bio05: maximum temperature of warmest month; bio06: minimum temperature of coldest month; bio17: precipitation of driest quarter; MO: organic material; and SLA: specific leaf area).

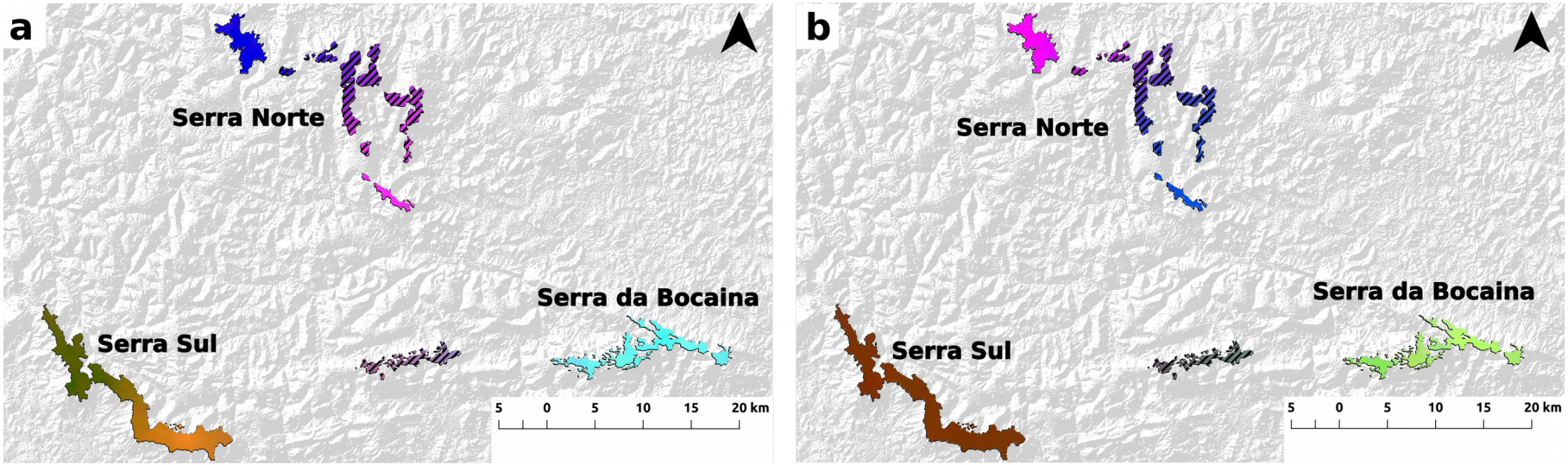

Figure 5: Spatial distribution of adaptive genetic variation in *Mimosa acutistipula* var. *ferrea* (**a**) and *Dioclea apurensis* (**b**). The maps represent an RGB composite made using interpolated principal components from a sPCA, ran on the combined candidate loci found in GEA and GPA. Regions with similar colors within each panel represent analogous genetic composition and areas with diagonal lines were not sampled (i.e., adaptive genetic composition was interpolated from neighboring areas).

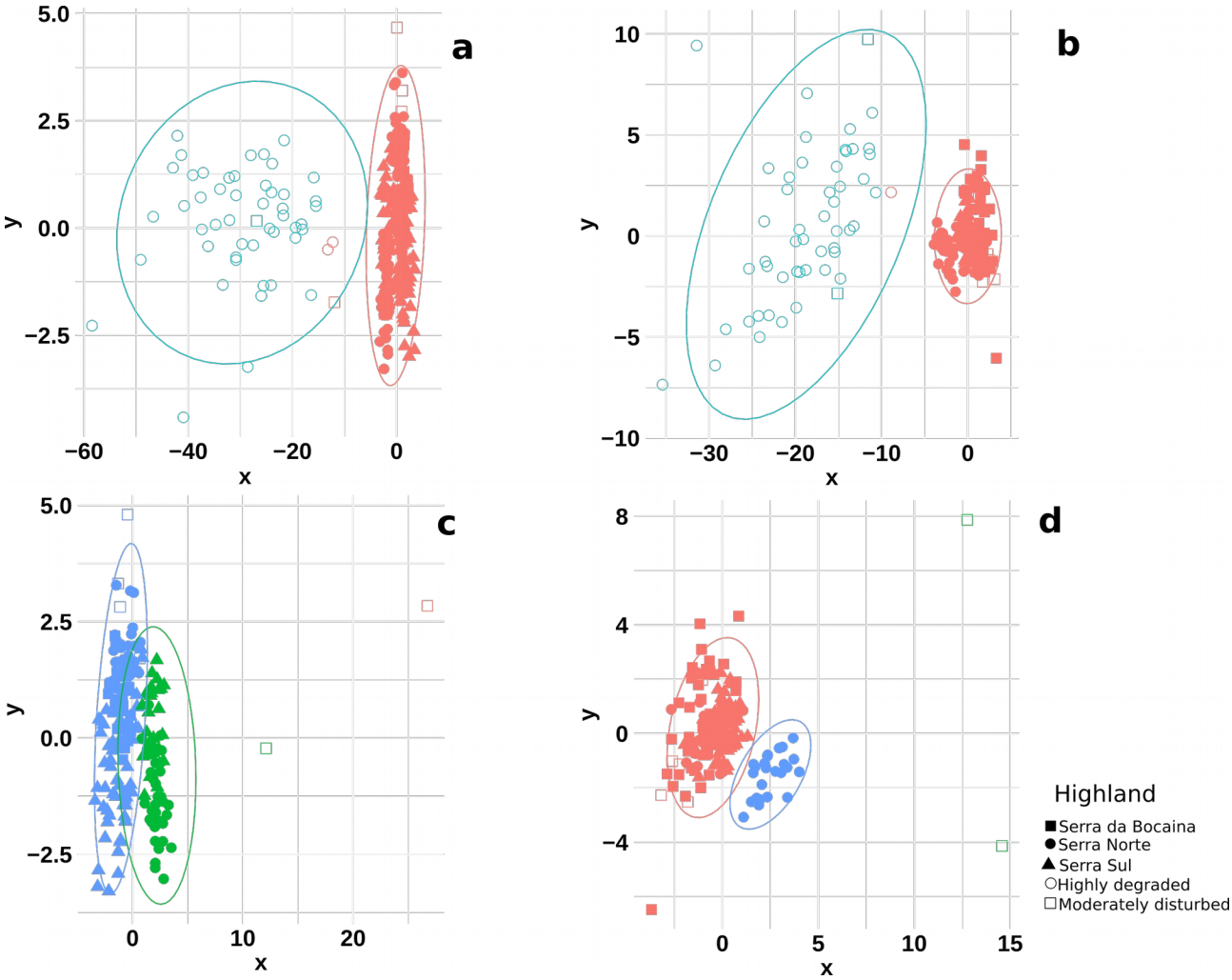

Figure 6: Clustering of predicted and observed adaptive genetic variation in *Mimosa acutistipula* var. *ferrea* (a, c) and *Dioclea apurensis* (b, d). A *k*-means clustering approach was employed to examine the similarity between observed (filled symbols) and predicted genotypes (open symbols) associated to the environmental conditions of highly degraded (upper panels) and moderately disturbed (lower panels) sites. Colors indicate different clusters and symbols the Canga highlands where individuals were sampled.