Tree Genetics & Genomes (2015) 11: 16
DOI 10.1007/s11295-015-0836-3

ORIGINAL PAPER

Multi-objective optimization for plant germplasm collection conservation of genetic resources based on molecular variability

Shana Schlottfeldt · Maria Emília M. T. Walter · André Carlos P. L. F. de Carvalho · Thannya N. Soares · Mariana P. C. Telles · Rafael D. Loyola · José Alexandre F. Diniz-Filho

Received: 21 September 2014 / Revised: 22 December 2014 / Accepted: 6 January 2015 / Published online: 5 February 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Germplasm collections play a significant role among strategies for conservation of diversity. It is common to select a core collection to represent the genetic diversity of a germplasm collection, in order to minimize the cost of conservation, while ensuring the maximization of genetic variation. We aimed to solve two main problems: (1) to select a set of individuals, from an in situ data set, that is genetically complementary to an existing germplasm collection, and (2) to define a core collection for a germplasm collection. We proposed a new multi-objective optimization (MOO) approach based on principles of systematic conservation planning (SCP) incorporating heterozygosity information; therefore, optimization takes genotypic diversity and variability patterns into account as well. As a case study, we used *Dipteryx alata* microsatellite loci information from two sources, an ex situ germplasm collection located at the Agronomy School of the Federal University of Goiás (UFG-AS), and an in situ data set composed of 642 sampled individual trees. We were able to identify within a population of several individuals, the exact accessions/samples that should be chosen in order to preserve the species diversity. We found that material from nine in situ individual trees are enough to complement the UFG-AS germplasm collection as it is, and that it is possible to define a core collection of 20 individual trees representing all studied genetic diversity. Moreover, we defined a method (a protocol) to deal with large amounts of accessions in the context of MOO. The proposed approach can be used to help constructing collections with maximal allelic richness and can also be extended to the in situ conservation. As far as we know, this is the first time that principles of SCP and the MOO approach are applied to the problem of complementing a germplasm collection and of finding a core collection for a germplasm collection.

Keywords Biodiversity · Conservation planning · Core collection · Genetic variability · Germplasm · Multi-objective optimization

Communicated by: J. Beaulieu

Electronic supplementary material The online version of this article (doi:10.1007/s11295-015-0836-3) contains supplementary material, which is available to authorized users.

S. Schlottfeldt (✉) · M. E. M. T. Walter
Department of Computer Science, University of Brasilia, Brasilia, DF, 70910-900 Brazil
e-mail: shanass@unb.br

M. E. M. T. Walter
e-mail: mariaemilia@unb.br

A. C. P. L. F. Carvalho
Department of Computer Science, University of São Paulo, SCC–ICMC–USP, CxP 668, São Carlos, SP, 13560-970 Brazil
e-mail: andre@icmc.usp.br

T. N. Soares · M. P. C. Telles
Department of General Biology, Federal University of Goiás, CxP 131, Goiânia, GO, 74001-970 Brazil
e-mail: tnsoares@gmail.com

M.P.C. Telles
e-mail: tellesmpc@gmail.com

R. D. Loyola · J. A. F. Diniz-Filho
Department of Ecology, Federal University of Goiás, CxP 131, Goiânia, GO, 74001-970 Brazil
e-mail: rdiasloyola@gmail.com

J.A.F. Diniz-Filho
e-mail: diniz@ufg.br

Introduction

Germplasm corresponds to the living tissues from which new plants can be grown, and thus are an important component for maintenance of plant genetic resources (Roederer et al. 2000). Plant germplasm can be stored as seed collections, pollen storage, in a nursery (field), in vitro (Engels et al. 2003) constituting what is called germplasm bank or collection. These germplasm collections are kept to represent the genetic diversity of plants, their wild relatives, and/or plants present and unique in a local region (Dawson and Were 1997). They play a significant role among strategies for conservation of diversity (Rao et al. 2006). Additionally, to commonly occurring species, rare, threatened, or endangered species are made available for study or for habitat restoration projects.

Core collections are useful tools for organizing and analyzing representative sets of genotypes in a germplasm collection and can be defined based on several criteria, including explicit evaluation using molecular markers. A core collection is a subset from a larger collection of a particular species that represents, with a minimum level of repetitiveness, the genetic diversity of that species and its wild relatives (Frankel 1984; Brown 1989). A core collection should not be considered a substitute of the whole collection, but it captures the complete diversity of the entire collection it was derived from. Therefore, a core collection should correspond to a set representing all of the species alleles, while ensuring minimum redundancy of those alleles and high reproducibility of entries. Core collections are being adopted as a useful instrument to improve conservation, accessibility and the use of plant genetic resources (Zhang et al. 2011). In this conservation context, one could understand that remaining natural populations of the species could be viewed as in situ genetic resources, whereas a core collection could be an ex situ sample that could be stored for further conservation applications. In this conservation context, the use of molecular markers to achieve the representativeness definition of a core collection is important because population persistence and resilience to environmental changes are usually positively correlated with genetic diversity. Methodologically, efforts to create germplasm core collections commonly use statistical and clustering methods (Holbrook et al. 1993; Li et al. 2004; Laghetti et al. 2008; Wang et al. 2011; Zhang et al. 2011; Belaj et al. 2012; Mei et al. 2012; Rao et al. 2012).

There is always a cost involved in maintaining germplasm collections such as maintenance of storage space, controlled temperature, and relative humidity (Rao et al. 2002). Various studies for estimating the costs of conservation have been carried out adopting different methodologies (Gupta et al. 2002). Germplasm collection operations and facilities vary substantially in size, capacity, types, and amount of equipments. These factors depend on the quantities of germplasm to be stored in the germplasm collection, which will in turn depend upon several circumstances that vary from objectives of the germplasm collection, range of species and breeds to be conserved, to financial resources available for the conservation program.

Cerrado is a large biome in Central Brazil, occupying about $1,500,000$ km^2 and includes significant environmental heterogeneity. It is a typical tropical savanna environment, but actually has different types of vegetation, ranging from open grasslands to dense woodlands and dry forests. Cerrado has been considered one global biodiversity hot spot (Myers et al. 2000), because of the strong plant endemism and high rates of habitat loss and fragmentation due to a recent expansion of soybean cultures and cattle ranching (Diniz-Filho et al. 2008). The conservation of indigenous plant species germplasm is important since considerable material has been identified as unique to the Brazilian Cerrado biodiversity. In this context, germplasm collections exist to conserve, increase utilization, and catalogue germplasm of plants that might otherwise be lost.

At the Agronomy School of the Federal University of Goiás (UFG-AS), Brazil, there is a large accessible germplasm collection for *Dipteryx alata* (also known as *baru*, a Fabaceae tree species widely distributed and endemic to Brazilian Cerrado), composed of 178 fully grown adult trees. Besides, data from *D. alata* was collected in situ for a total of 642 individual trees sampled in 25 local populations throughout species' geographic range (Fig. 1).

In this context, using data from *D. alata* as a case study for our method, we aim to solve two main general problems:

1. To select a set of individuals from in situ collected data that is genetically complementary to an existing germplasm collection;
2. To define a core collection for a germplasm collection.

In both problems, we look for minimizing the overall cost of conservation while maximizing the allele representation. On the one hand, these problems can be mapped to the systematic conservation planning (SCP), a widely accepted biodiversity-focused approach to selecting priority areas for protection. In a simplified way, SCP is the problem of finding a minimum set of sites (in this case, individual trees) with the maximum representation of some feature (Margules and Pressey 2000). On the other hand, the SCP problem can be modeled by the minimum set covering problem, that was shown to be NP-hard (Cormen et al. 2001), i.e., there is no known efficient exact solution for the problem, therefore, when the input grows arithmetically, the time to find a solution increases exponentially. Clearly, there are two conflicting objectives (minimize selected individuals

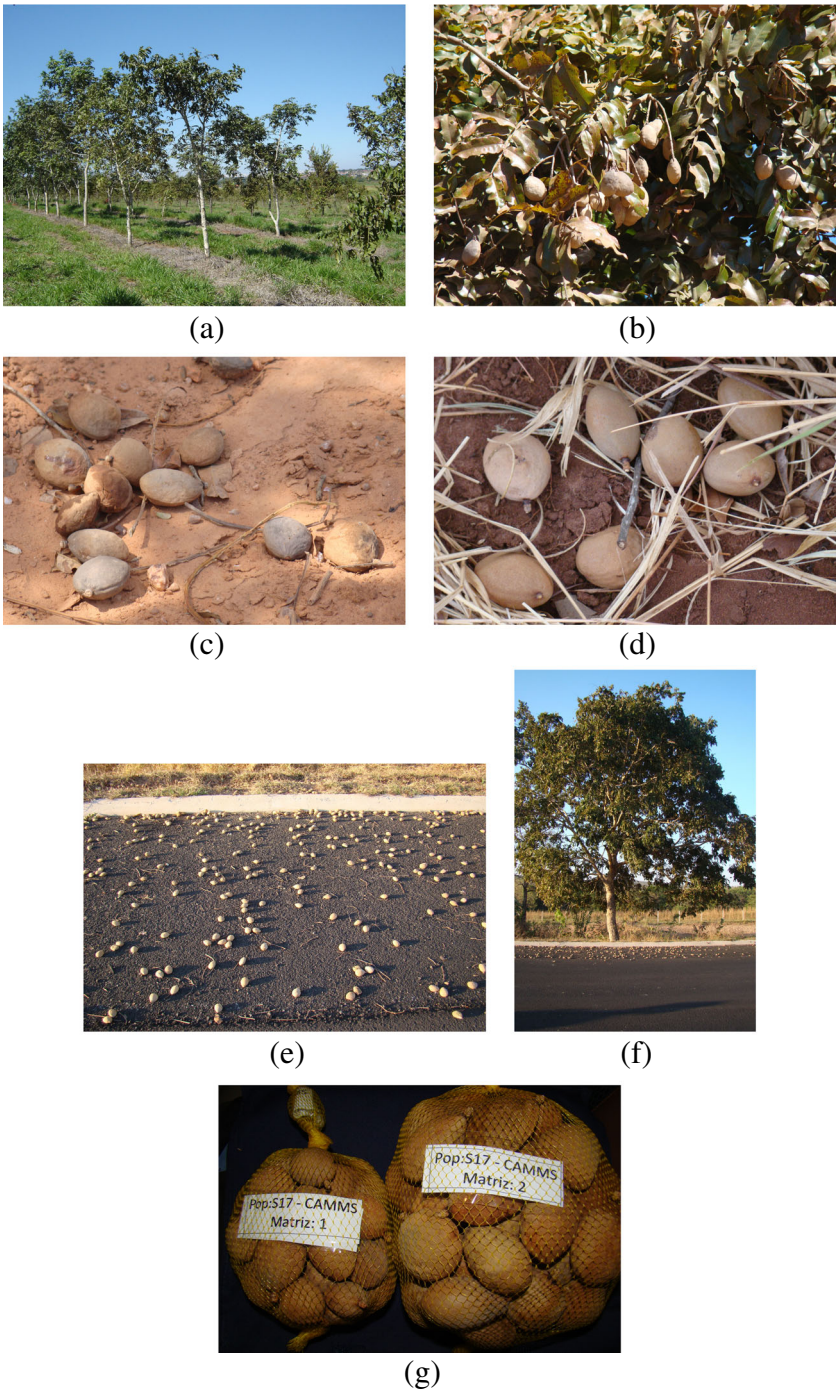


Fig. 1 *Dipteryx alata*, also known as *baru*, is a widely distributed tree species in the Cerrado biome, Central Brazil, although restricted to seasonal savannas that grow in eutrophic and drained soil. Fruits have a very woody endocarp with an edible nut that is eaten and dispersed by mammals, e.g., bats and monkeys. The *D. alata* is used as lumber, for charcoal production, shade in pasture, and it is a source of raw material for handicraft, cosmetics, and food industries, playing an important role in the local economy (Correa et al. 2008; Collevatti et al. 2013). **a** The UFG-AS *D. alata* germplasm collection (nursery), composed of 178 fully grown adult trees. **b** *Baru* fruit in situ. **c**–**e** Ripe baru fruits are collected from the ground. **f** Individual tree in situ. **g** In situ collected samples (samples were collected for a total of 642 individual trees in 25 local populations distributed throughout species' geographic range)

while maximizing allele representation), making the problem a perfect candidate for multi-objective optimization (MOO).

Optimization problems with more than one objective are called vector optimization or multi-objective problems (MOP) (Brockhoff 2009; Deb 2008; Coello Coello 2001; Coello Coello et al. 2007; Zitzler 2002). In these cases, there is no single optimal solution, but rather, a set of solutions that should be considered equivalents in the absence of information about the relevance of each objective related to the others (Fonseca and Fleming 1995). These solutions are known as Pareto optimal solutions (Fig. 2), and their plot form what is called Pareto front (Fig. 3) (Coello Coello et al. 2007).

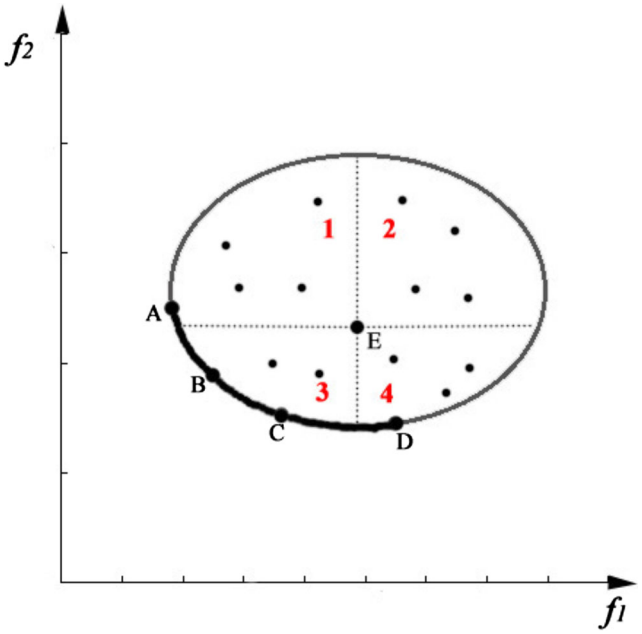


Fig. 2 Pareto optimality. In multiobjective problems (MOP) there is not only one objective function to optimize, but many, thus, the aim is to find good trade-offs rather than a single solution. The notion of optimum is based on the Pareto dominance: considering two objective functions (f_1 and f_2), a solution (a point) can be better, worse, equal, or also indifferent to another solution with respect to the objective values computed from functions f_1 and f_2. "Better" means that a solution is not worse in any objective and it is better in at least one objective; this solution is also said to dominate the other ones. Using this concept, an optimal solution is not dominated by any other solution. Such a solution is called Pareto optimal, and the entire set of optimal trade-offs is called the Pareto optimal set, which is represented here by points *A*, *B*, *C*, and *D*. Considering a minimization problem, taking solution *E* as reference, solutions located in: area *2* are dominated by *E* (are worse than *E*); area *3* dominate *E* (are better than *E*); areas *1* and *4* are indifferent (it is not possible to compare them, since if *E* is better in f_1, it is worse in f_2 and vice versa)

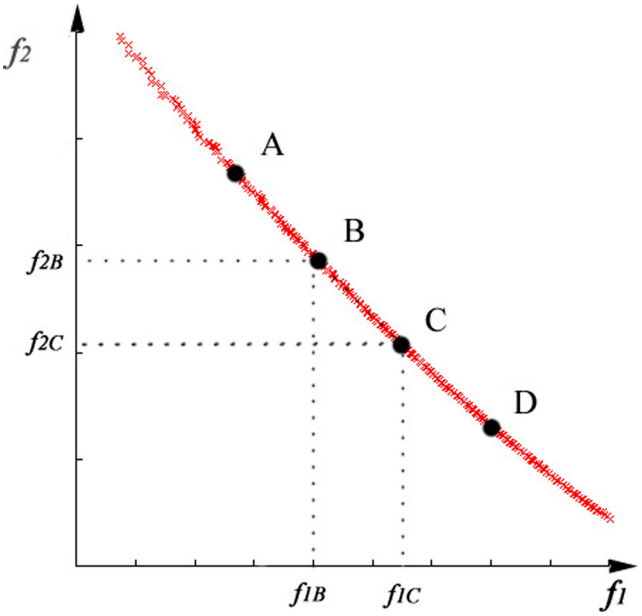


Fig. 3 Pareto front. An example of a problem with two objective functions to be simultaneously minimized: f_1 and f_2. Observe that for points *B* and *C*, it is not possible to obtain an improvement in one objective without a degradation of the other objective, i.e., $f_{1B} < f_{1C}$, $f_{2B} > f_{2C}$. The solutions (points) *A*, *B*, *C*, and *D* are optimal when f_1 and f_2 are considered. The Pareto front (or trade-off surface) is indicated by the *curve*

SCP has been generally used at species level (or hierarchically higher), but has also been applied to conservation genetics, aiming to maximize molecular variation within populations (Diniz-Filho and Telles 2002; 2006; Diniz-Filho et al. 2012b). To properly use this method to organize germplasm collections, it would be important to improve the approach to analyze individuals and their genotypes.

In a previous work, our group solved a problem that looked for the smallest set of local populations of *D. alata* that might be preserved in order to represent the genetic diversity of this species to its in situ conservation (Schlottfeldt-Santos et al. 2012; Schlottfeldt et al. 2014). That study was pioneer in the use of information about allele frequency, heterozygosity, and Hardy-Weinberg equilibrium as objectives for simultaneous optimization.

Diniz-Filho et al. (2012b) used a mono-objective approach (in particular, simulated annealing) to find the minimum amount of populations, not specifically individuals, needed to represent all diversity in in situ conservation of *D. alata*. Here, we propose a multi-objective optimization approach for the conservation of genetic resources based on individual molecular variability aiming to inform ex situ conservation strategies, and to guide sampling for germplasm collections. We represented the known alleles, but incorporated individual heterozygosity information, thus optimization also takes genotypic diversity and variability patterns into account. By including these characteristics, individuals can better represent the genetic diversity, allowing to identify sets of individuals with a higher probability of persistence throughout time.

As far as we know, this is the first time that a SCP approach dealing with multi-objective optimization is applied to the problems of finding a core collection for a germplasm collection and complementing this core collection using heterozygosity information as well.

Material and methods

Data

As a case study for our method, we used data from *D. alata* (data available on Supplementary_File_S2, and Supplementary_File_S3):

1. An ex situ data set corresponding to the germplasm collection for *D. alata*, composed of 178 fully grown adult trees, located at the Agronomy School of the Federal University of Goiás. These 178 individuals came from several locations in the state of Goiás, Brazil.
2. An in situ data set composed of 642 individual trees sampled in 25 local populations throughout *D. alata*'s geographic range (Fig. 4). Sample sizes within local populations ranged from 12 to 32 (see Diniz-Filho et al. (2012a) and Soares et al. (2012) for sampling methodological details).

The *D. alata* samples were genotyped for nine microsatellite loci (Bm164, DaE06, DaE12, DaE20, DaE34, DaE41, DaE63, DaE67, DaE46) (Soares et al. 2012; Diniz-Filho et al. 2012b) finding a total of 55 distinct alleles. All 55 alleles are represented among the in situ sampled trees. The ex situ germplasm collection has 40 from the 55 studied alleles.

Based on these data, we produced three matrices:

1. Matrix A: an allele-by-tree matrix for the ex situ data set. In matrix A_{txa}, $t = 178$ (UFG-AS adult trees) and $a = 55$ (alleles), a_{ij} represents the occurrence of allele j in tree i.
2. Matrix B: an allele-by-tree matrix for the in situ data. In matrix C_{txa}, $t = 642$ (sampled individual trees) and $a = 55$ (alleles), b_{ij} represents the occurrence of allele j in tree i.
3. Matrix C: an allele-by-population presence-absence matrix for the in situ data. In matrix C_{pxa}, $p = 25$ (populations) and $a = 55$ (alleles), c_{ij} represents the occurrence of allele j in population i.

In these matrices, alleles in homozygosity received lower score compared to alleles in heterozygosity. By doing this, we benefited solutions with higher content of heterozygosity.

The problem

The overall problem is to maximize the number of alleles while minimizing the number of individuals required to represent all alleles, maximizing at the same time heterozygosity. A candidate solution for the problem is a vector

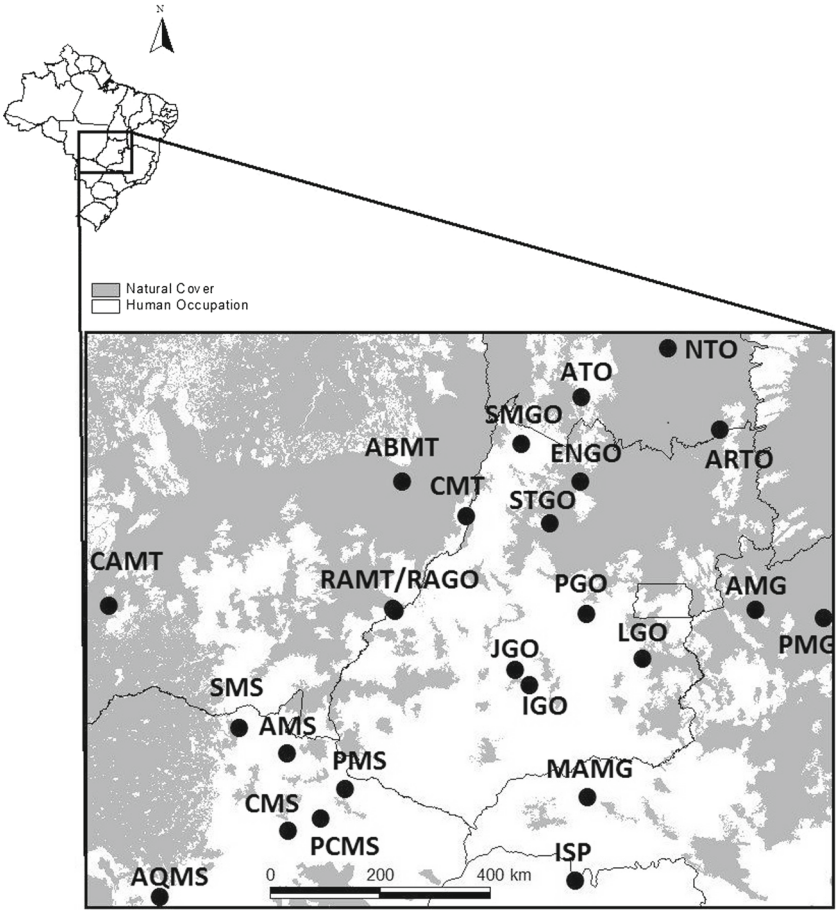


Fig. 4 Geographic location of the 25 sampled local populations of *Dipteryx alata* in Central Brazil. Regions shown in *dark tone* are still covered by natural remnants of Cerrado vegetation. It is worth to note that the studied region is comprised of small fragments or isolated individuals in a matrix of pastures and crops, notably maize and soybean. The only populations connected by continuous original vegetation are *RAGO* and *RAMT*. The names and geographical coordinates of the populations can be found in Supplementary file S1. Modified from Diniz-Filho et al. (2012b)

$\overrightarrow{x} = x_1, ..., x_k$, where k is the number of accessions (individual trees), $x_i \in \{0, 1\}$, such that $x_i = 1$, if the individual tree i is selected to compose the solution; or 0, otherwise.

The aim is to obtain:

$$\min\left(\sum_{i=1}^{k} x_i\right) \quad (1)$$

Subject to:

$$\forall j \in \{1, 2, ..., n\}, \sum_{i=1}^{k} a_{ij} x_i \geq 1 \quad (2)$$

Where $n = 55$ (alleles).

Regarding the multi-objective optimization (MOO), there are three objectives to be optimized:

1. Minimize the number of selected individuals (3);

$$\min(f_1(\overrightarrow{x})) = \min(\text{individual_trees}(\overrightarrow{x})) \quad (3)$$

2. Maximize the number of alleles (4). For simplicity, this second objective function was transformed into an equivalent minimization objective function by using the number of lacking alleles (those that are not present in the solution);

$$\begin{aligned} \max(f_2(\overrightarrow{x})) &= \max(\text{alleles}(\overrightarrow{x})) \Leftrightarrow \min(\text{lacking_alleles}(\overrightarrow{x})) \\ &= \min(55\text{-alleles}(\overrightarrow{x})) \end{aligned} \quad (4)$$

3. Maximize the heterozygosity (5).

$$\max(f_3(\overrightarrow{x})) = \max(\text{heterozygosity}(\overrightarrow{x})) \quad (5)$$

Next, we describe the experiments used to select individuals from the in situ sample in order to complement the germplasm collection and to select entries from the existing germplasm collection to form a core collection.

Experiment 1: Complementing the UFG-AS germplasm collection. The optimization problem defined here is to find the smallest set of *D. alata* individuals that better complement the genetic variability already preserved in the UFG-AS germplasm collection, thus, representing the genetic diversity of this species, aiming at its conservation and persistence. This problem is similar to the one stated in Schlottfeldt-Santos et al. (2012), but instead of treating populations as a unit, here we solved the problem at individual level.

As mentioned before, from the 55 studied alleles, 40 were already presented in the germplasm collection, therefore, individual trees belonging to the in situ data set should be selected in order to represent those 15 still lacking alleles.

Experiment 2: Defining the core collection (reducing the existing germplasm collection). One of the challenges of maintaining a germplasm collection is to keep the genetic variability while reducing the maintenance costs, and this is the objective of this experiment.

We have two scenarios:

1. *Scenario 1*: we addressed the problem of defining a core collection by identifying, within the UFG-AS germplasm collection, the minimal set of individuals needed to represent all the genetic variability exhibited by the germplasm collection.

   For this first scenario, we optimized the three objectives stated in Eqs. 3 to 5 on the ex situ data set.
2. *Scenario 2*: considering that maybe within the universe of all sampled trees (ex situ and in situ individual trees) it could be found an even smaller core collection, we performed the optimization joining information from the ex situ data set (178 trees) and the in situ data set (642 trees), in a total of 820 individual trees.

   For this scenario, we tested two approaches in order to define a protocol to deal with this optimization in a more efficient way, aiming the attainment of better results (a smaller, still representative, core collection):

   (a) *Method 1*: to perform the MOO directly on the 820 individual tree data set.

   (b) *Method 2*:

   Step 1: to perform the MOO in population level, by considering the 25 in situ data populations and assuming that the germplasm collection was the 26th population. This step identifies the smallest set of populations needed to represent all 55 alleles.

   Step 2: considering only individual trees belonging to populations selected in step 1, to perform a new MOO.

The null model. We executed a null model in order to verify if the same results would be found without the use of MOO.

Results

Table 1 summarizes the obtained results. It is worth noting that here, we are dealing with solutions having a minimum set of individuals, but since there was more than one optimized objective, we obtained a larger portfolio of solutions that are equally good in the sense that none of them is better when all the objectives are simultaneously considered (they are all Pareto optimal).

We performed experiment 1 aiming at complementing the UFG-AS germplasm collection. We found that the

Table 1 Sets of selected populations identified by simultaneously optimizing three objectives, where the first objective is the minimization of selected population/individuals set, the second objective is the minimization of lacking alleles, and the third objective is the maximization of heterozygosity

					Results	
		Data source	Number of accessions	Number of alleles	Core collection	Pop. minimum
Experiment 1		In situ data set	642	15	9	8
	Scenario 1	Ex situ data set	178	40	18	-
	Scenario 2	(In situ + ex situ)	820	55	42	20
Experiment 2	Method 1	data set				
	Scenario 2	Selected pop. from	336	55	20	6
	Method 2	(In situ + ex situ)	(in 8 pop.)			
		data set				

The column "Core collection" shows the minimal quantity of individuals found as solution to represent all the diversity (all the alleles) indicated in the experiment. The column "Pop. minimum" shows the number of populations to which the individuals found as minimum solution belong

smallest set of individuals needed to complement the UFG-AS germplasm collection has nine individuals selected from eight populations (AMS, ABMT, CAMT, CMT, MAMG, PMS, RAMT, SMS) (Fig. 5).

In experiment 2, we had the objective of finding a core collection for the UFG-AS germplasm collection, trying to reduce the number of individuals (and therefore, the maintenance costs), but preserving the overall diversity. For scenario 1, we found that 18 individuals were enough to represent all 40 alleles from the UFG-AS germplasm collection

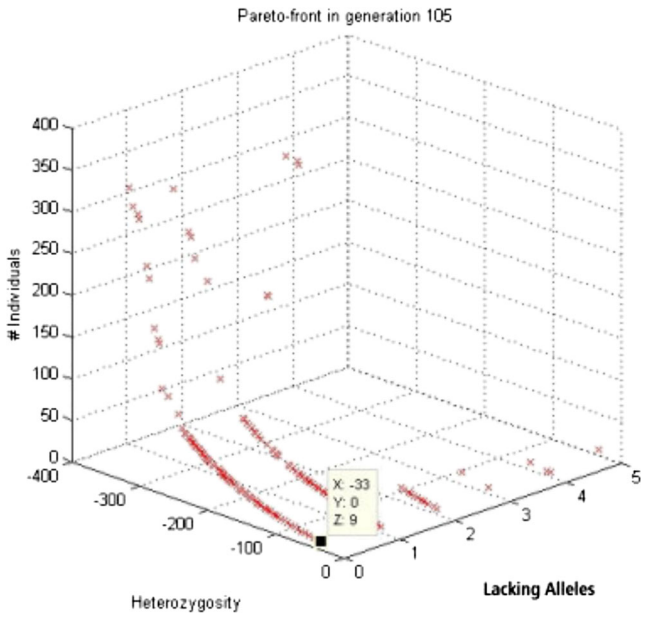


Fig. 5 Pareto front for experiment 1 (complementing the UFG-AS germplasm collection with selected individuals from the in situ data set)

(Fig. 6). For scenario 2 - method 1, we found that 42 individuals were enough to represent all 55 alleles taking into account the 820 individuals trees from joined in situ and ex situ data set (Fig. 7a). Finally, for scenario 2 - method 2, we found, in step 1, that the smallest set of *populations* needed to represent all 55 alleles was 5, but among all the obtained solutions, from the initial universe of 26 populations, 8 appeared in at least one solution with no lacking alleles (AMS, CAMT, CMT, ENGO, PMS, RAMT, STGO, UFG-AS). In step 2, MOO was performed considering only individual trees belonging to those eight populations. In the end of the process, we found that the smallest set of individuals needed to represent all 55 alleles is composed of 20 individuals (within six populations: AMS, CAMT, CMT, PMS, RAMT, UFG-AS) (Fig. 7b).

Finally, we generated a *null model* as follows. We randomly produced the same amount of solution generated for each previously described experiment. For each randomly generated solution, we computed values for heterozygosity, number of lacking alleles, and number of populations/individuals, following the computations performed for the MOO. For all the randomly generated solutions, the comparison values were worse than solutions found with MOO, as well as the number of lacking alleles (meaning that less alleles were presented in the null model solutions).

Discussion

The proposed MOO approach based on principles of SCP incorporating heterozygosity information can be used to help construct collections with maximal allelic richness.

The most important contribution of this work is to identify, in the context of SCP, within a population of several

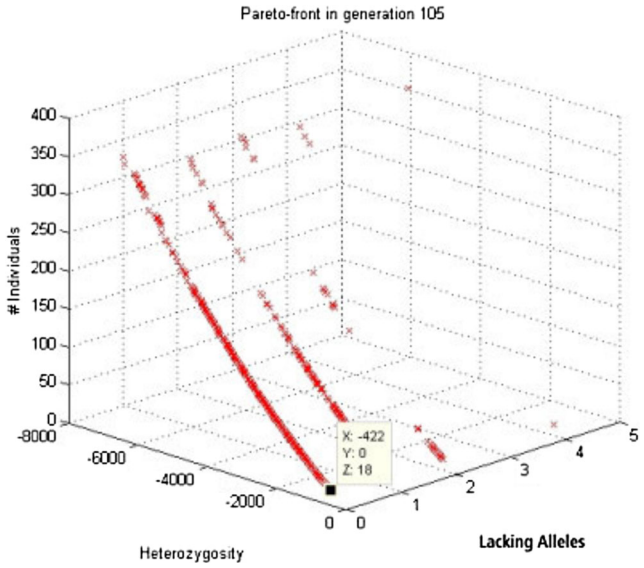
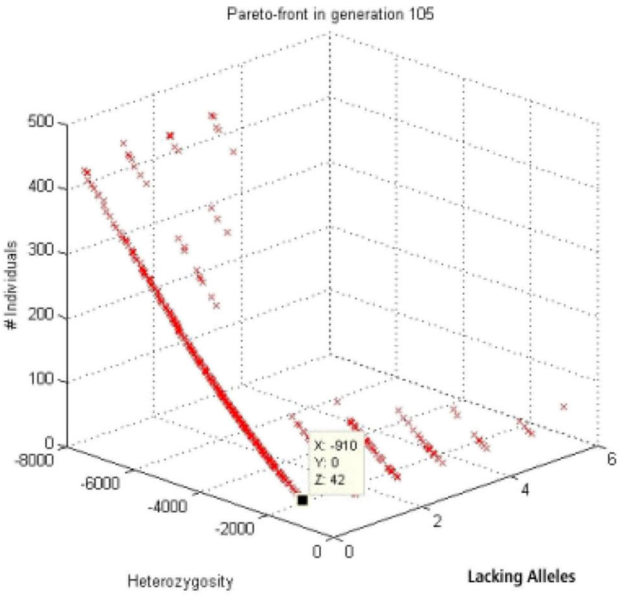
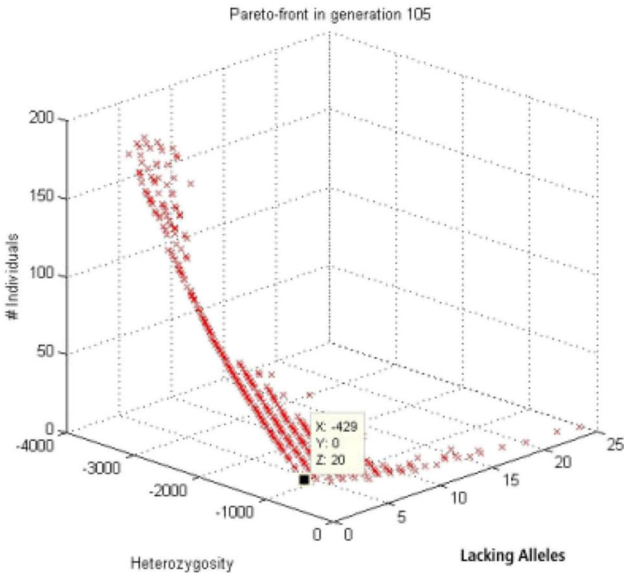


Fig. 6 Pareto front for experiment 2 - scenario 1 (defining the minimal set of individuals needed to represent all the genetic variability exhibited within the UFG-AS germplasm collection (ex situ data set))



(a) Method 1: performing MOO directly on all 26 populations (*in-situ* + *ex-situ* data set = 820 individual trees).



(b) Method 2: using previously selected populations.

Fig. 7 Pareto front for experiment 2 - scenario 2 (ex situ and in situ data set altogether)

individuals, the exact samples that should be chosen in order to preserve the species diversity.

This approach can be extended to the in situ conservation. Previous approaches generally indicate a population to be preserved; the proposed method indicates exactly which individuals within the population should be sampled/kept.

Even if the aim is to obtain a minimum set, this method identifies a portfolio of solutions, indicating sets with more individuals that better fulfill the stated objectives, providing decision makers with more options for achieving their conservation targets. These sets are optimal in the sense that none of them can be considered the best when all the objectives are simultaneously considered.

In the context of the case study carried out, most *D. alata* diversity is found only in nature, and many such populations are increasingly threatened by habitat reduction. In nature, there are many potential useful populations, yet, for practical purposes, only a fraction of this material can be afforded protection or maintenance in germplasm collections or in protected areas. Additionally, individual trees are often geographically wide ranging, making it costly to collect representative samples of material.

Experiment 1 results show that samples of only nine ex situ individual trees are necessary to complement the existing UFG-AS germplasm collection, completing the already represented 40 alleles with the 15 lacking ones.

Moreover, using the proposed approach, it is possible to include additional optimization objectives, such as the distance from in situ individual trees to the UFG-AS, reducing, therefore, displacement costs associated with collection of samples for complementing the germplasm collection.

Experiment 2 - scenario 1 found that from the 178 trees in the UFG-AS germplasm collection, it is possible to preserve the allele diversity (40 alleles) by keeping a core collection of 18 individual trees (only 10 % from the current germplasm collection). Considering that the lifespan of a *D. alata* tree is 60 years, the proposed method is important to

define strategies to provide a set of genetically diverse material while selecting the most representative individual trees. By maximizing genetic diversity in germplasm collections, resources available for conservation of biodiversity can be allocated to a larger number of species.

Associating results of experiment 1 (minimum set of 9 individual trees) and experiment 2 - scenario 1 (minimum set of 18 individual trees), we obtained a core collection of 27 individual trees ($9 + 18 = 27$) representing the entire allelic richness of the nine coded microsatellite *loci*.

Considering that within the universe of all the available data (ex situ and in situ data sets, e.g., 820 individual trees) a smaller core collection could be found, we executed experiment 2 - scenario 2 in order to verify this hypothesis. Experiment 2 - scenario 2 - method 1 found a minimum set of 42 individual trees representing all the allelic richness. Likely, this result was not better because we have done the same number of executions performed in the previous experiments, but with a bigger input (820 individual trees). As an alternative to overcome this issue, we proposed a refining method that initially selected the main populations integrating the optimum solutions, and within these populations we looked for the specific individuals that would compose our minimum set. By doing so, experiment 2 - scenario 2 - method 2 was able to find a core collection with only 20 individual trees representing all 55 alleles; less than the numbers obtained for experiment 2 - scenario 2 - method 1 (42 individual trees) and experiment 1 associated with experiment 2 - scenario 1 (27 individual trees), suggesting that this is an adequate method to be applied.

We verified that, for all the experiments, there was a significant improvement in the retention of alleles found in selected accessions.

Results from the null model assured that solutions obtained with MOO have not emerged randomly.

An effective genetic strategy should maximize the retention of genetic variation associated with long-term species survival, yet it is not easy to assess the level and properties of such variation, not to mention which traits might become important in future.

The new proposed approach using MOO and SCP for the conservation of genetic resources based on individual molecular variability helps design ex situ conservation strategies, and guide in situ sampling for germplasm core collections aiming for effective conservation and future species utilization.

Conclusion

Genetic diversity is basic for meaningful and effective conservation for the species-specific traits. This paper showed how principles of SCP and the MOO approach associated to microsatellite loci information can be applied to successfully help construct germplasm collections having maximum allelic richness and minimum number of accessions.

Having established that the approach is viable, our future work will focus on applying this approach to other kinds of genomic information (e.g., single nucleotide polymorphism – SNP) in order to verify its feasibility for this kind of data.

Acknowledgments SS acknowledges the support from CNPq, through the Science without Borders Program. The research program in molecular ecology and conservation of Cerrado plants has been continuously supported by several grants and fellowships to the research network GENPAC (Geographical Genetics and Regional Planning for Natural Resources in Brazilian Cerrado) from CNPq/MCT/CAPES (projects#564717/2010-0 and #563624/2010-8) and by the "Núcleo de Excelência em Genética e Conservação de Espécies do Cerrado"-GECER (PRONEX/FAPEG/CNPq CP 07-2009). Field work has been supported by Systema Naturae Consultoria Ambiental LTDA and work by MEMTW, ACPLFC, TNS, MPCT, RDL, and JAFDF have been continuously supported by productivity fellowships from CNPq.

Data Archiving Statement

Raw data available from the Dryad Digital Repository: http://doi.org/10.5061/dryad.hq8pd. (Dryad informs that we must wait until our manuscript has been accepted before submitting data to the repository). Supplementary_File_S1 (geografic_location): Geographic location of *D. alata* populations. Supplementary_File_S2 (ex-situ_data): Excel file containing *D. alata ex-situ* genotype data (UFG-AS germplasm collection). Individuals are identified in each row. Loci are designated by column with the nine microsatellite listed. Supplementary_File_S3 (in-situ_data): Excel file containing *D. alata in-situ* genotype data. Individuals are identified by population in each row. Loci are designated by column with the nine microsatellite listed. (?) refers to missing data.

References

Belaj A, del C Dominguez-Garcia M, Atienza SG, Urdiroz NM, de la Rosa R, Satovic Z, Martin A, Kilian A, Trujillo I, Valpuesta V, Del Rio C (2012) Developing a core collection of olive (Olea europaea L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. Tree Genet Genomes 8(2):365–378. doi:10.1007/s11295-011-0447-6

Brockhoff D (2009) Theoretical aspects of evolutionary multiobjective optimization—a review. Rapport de Recherche RR-7030, INRIA Saclay—Île-de-France

Brown A (1989) Core collections: a practical approach to genetic resources management. Genome 31(2):818–824

Coello Coello CA (2001) A short tutorial on evolutionary multiobjective optimization. In: Proceedings of the first international conference on evolutionary multi-criterion optimization. Springer-Verlag, London. EMO '01, pp 21–40. http://dl.acm.org/citation.cfm?id=647889.736510

Coello Coello CA, Lamont GB, Van Veldhuizen DA (2007) Evolutionary algorithms for solving multi-objective problems, 2nd edn. Springer-Verlag, New York. ISBN 978-0-387-33254-3

Collevatti R, Telles M, Nabout J, Chaves L, Soares T (2013) Demographic history and the low genetic diversity in *Dipteryx alata* (Fabaceae) from Brazilian Neotropical Savannas. Heredity doi:10.5061/dryad.1cd80

Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms, 2nd edn. The MIT Press

Correa GC, Naves RV, Rocha MR, Chaves LJ, Borges JD (2008) Physical determinations in fruit and seeds of Baru (*Dipteryx alata* Vog.), Cajuzinho (*Anacardium othonianum* Rizz.) and Pequi (*Caryocar brasiliense* Camb.), aiming genetic breeding. Biosci J 24(4):42–47

Dawson I, Were J (1997) Collecting germplasm from trees—some guidelines. Agrofor Today 9(2):6–9

Deb K (2008) Multiobjective optimization, Springer-Verlag, chap introduction to evolutionary multiobjective optimization, pp 59–96

Diniz-Filho J, Telles M (2006) Optimization procedures for establishing reserve networks for biodiversity conservation taking into account population genetic structure. Genet Mol Biol 29:207–214

Diniz-Filho J, Bini LM, Vieira C, Blamires D, Terribile L, Bastos R, Oliveira G, Souza B (2008) Spatial patterns of terrestrial vertebrate species richness in the Brazilian Cerrado. Zool Stud 47(2):146–157

Diniz-Filho JAF, Telles MPC (2002) Spatial autocorrelation analysis and the identification of operational units for conservation in continuous populations. Conserv Biol 16(4):924–935

Diniz-Filho JAF, Collevatti RG, Chaves LJ, Soares TN, Nabout JC, Rangel TF, Melo DB, Lima JS, Telles MPC (2012a) Geographic shifts in climatically suitable areas and loss of genetic variability in *Dipteryx alata* ("Baru" Tree; Fabaceae). Genet Mol Biol 11(2):1618–1626

Diniz-Filho JAF, Melo DB, de Oliveira G, Collevatti RG, Soares TN, Nabout JC, de S Lima J, Dobrovolski R, Chaves LJ, Naves RV, Loyola RD, de C Telles MP (2012b) Planning for optimal conservation geographical genetic variability within species. Conserv Genet 13:1085–1093

Engels J, Visser L, International Plant Genetic Resources Institute (2003) A guide to effective management of germplasm collections. IPGRI handbooks for genebanks, international plant genetic resources institute

Fonseca CM, Fleming PJ (1995) An overview of evolutionary algorithms in multiobjective optimization. IEEE T Evolut Comput 3:1–16

Frankel OH (1984) Genetic manipulation: impact on man and society, Cambridge University Press, chap Genetic perspectives of germplasm conservation, pp 161–170

Gupta A, Sanjeev S, Vikas C, B GS, Riya S, Neeru J (2002) Cost of conservation of agrobiodiversity. IIMA Working Papers WP2002-05-03, Indian Institute of Management Ahmedabad, Research and Publication Department, http://EconPapers.repec.org/RePEc:iim:iimawp:wp00015

Holbrook CC, Anderson WF, Pittman RN (1993) Selection of core collection from the U.S. Germplasm collection of peanut. Crop Sci 33(4):859–861

Laghetti G, Pignone D, Sonnante G (2008) Statistical approaches to analyze Gene Bank data using a lentil germplasm collection as a case study. Agric Conspec Sci 73(3):175–181

Li CT, Shi CH, Wu JG, Xu HM, Zhang HZ, Ren YL (2004) Methods of developing core collections based on the predicted genotypic value of rice (Oryza sativa L.) Theor Appl Genet 108(6):1172–1176. doi:10.1007/s00122-003-1536-1

Margules CR, Pressey RL (2000) Systematic conservation planning. Nature 405(6783):243–253. doi:10.1038/35012251

Mei Y, Zhou J, Xu H, Zhu S (2012) Development of sea island cotton (*Gossypium barbadense* L.) core collection using genotypic values. Aust J Crop Sci 6(4):673–680

Myers N, Mittermeier RA, Mittermeier CG, da Fonseca GAB, Kent J (2000) Biodiversity hot spots for conservation priorities. Nature 403:853–858

Rao ES, Kadirvel P, Symonds RC, Geethanjali S, Ebert AW (2012) Using SSR markers to map genetic diversity and population structure of *Solanum pimpinellifolium* for development of a core collection. Plant Genet Resour 10:38–48. doi:10.1017/S1479262111000955

Rao NK, Bonwoo K, Sastry DVSSR (2002) Pod and seed storage: cost-benefit study for groundnut germplasm conservation. International Arachis Newsletter 22:10–12

Rao NK, Hanson J, Dulloo ME, Ghosh K, Nowell A, Larinde M (2006) Manual of seed handling in genebanks. IPGRI Handbooks for Genebanks, Renouf Publishing Company Limited

Roederer C, Nugent R, Wilson P (2000) Economic impacts of genetically modified crops on the agri-food sector: a synthesis, working document, vol 1. Directorate-General for Agriculture, European Commission, http://ec.europa.eu/agriculture/publi/gmo/gmo.pdf

Schlottfeldt S, Timmis J, Walter MEMT, Carvalho ACPLF, Diniz-Filho JAF, Simon LM, Loyola RD, Telles MPC (2014) Multiobjective optimization applied to systematic conservation planning and spatial conservation priorities under climate change. In: Proceedings of the 2014 conference companion on genetic and evolutionary computation companion. GECCO Comp'14, pp 177–178, doi:10.1145/2598394.2598404

Schlottfeldt-Santos S, Walter MEMT, Diniz-Filho JAF, de C Telles MP (2012) Multiobjective optimization in systematic conservation planning to represent genetic variability within species. In: Proceedings of the 8th International Conference on Ecological Informatics

Soares TN, Melo DB, Resende LV, Vianello RP, Chaves LJ, Collevatti RG, Telles MPC (2012) Development of microsatellite markers for the neotropical tree species *Dipteryx alata* (Fabaceae). Am J Bot 99:e72–e73. doi:10.3732/ajb.1100377

Wang Y, Zhang J, Sun H, Ning N, Yang L (2011) Construction and evaluation of a primary core collection of apricot germplasm in China. Sci Hort 128:311–319. doi:10.1016/j.scienta.2011.01.025

Zhang P, Li J, Li X, Liu X, Zhao X, Lu Y (2011) Population structure and genetic diversity in a rice core collection (*Oryza sativa* L.) Investigated with SSR Markers. PLoS One 6(12):e27,565+ doi:10.1371/journal.pone.0027565

Zitzler E (2002) Evolutionary algorithms for multiobjective optimization. In: Evolutionary methods for design, optimisation, and control (EUROGEN 2001), CIMNE, pp 19–26