Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Optimizing reproducibility evaluation for random amplified polymorphic DNA markers

J.R. Ramos[1], M.P.C. Telles[1,2], J.A.F. Diniz-Filho[3], T.N. Soares[2,4], D.B. Melo[2,4] and G. Oliveira[3]

[1]Mestrado em Genética, Universidade Católica de Goiás, Goiânia, GO, Brasil
[2]Departamento de Biologia Geral, Laboratório de Genética and Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, GO, Brasil
[3]Departamento de Biologia Geral, Laboratório de Ecologia Teórica e Síntese, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, GO, Brasil
[4]Programa de Pós-graduação em Agronomia, Universidade Federal de Goiás, Goiânia, GO, Brasil

Corresponding author: M.P.C. Telles
E-mail: tellesmpc@gmail.com

Genet. Mol. Res. 7 (4): 1384-1391 (2008)
Received September 15, 2008
Accepted October 13, 2008
Published December 4, 2008

ABSTRACT. The random amplified polymorphic DNA (RAPD) technique is often criticized because it usually shows low levels of repeatability; thus it can generate spurious bands. These problems can be partially overcome by rigid laboratory protocols and by performing repeatability tests. However, because it is expensive and time-consuming to obtain genetic data twice for all individuals, a few randomly chosen individuals are usually selected for *a priori* repeatability analysis, introducing a potential bias in genetic parameter estimates. We developed a procedure to optimize repeatability analysis based on RAPD data, which was applied to evaluate genetic variability in three local populations of *Tibochina papyrus*, an endemic Cerrado plant found in elevated rocky fields in Brazil. We used a simulated annealing procedure to select the smallest number of individuals that contain all bands and repeated the analyses only

for those bands that were reproduced in these individuals. We compared genetic parameter estimates using HICKORY and POPGENE softwares on an unreduced data set and on data sets in which we eliminated bands based on repeatability of individuals selected by simulated annealing and based on three randomly selected individuals. Genetic parameter estimates were very similar when we used the optimization procedure to reduce the number of bands analyzed, but as expected, selecting only three individuals to evaluate the repeatability of bands produced very different estimates. We conclude that the problems of repeatability attributed to RAPD markers could be due to bias in the selection of loci and primers and not necessarily to the RAPD technique *per se*.

Key words: Random amplified polymorphic DNA; Repeatability; Optimization; Simulated annealing; Band selection; Primer selection

INTRODUCTION

Despite enormous advances in molecular techniques that have been used to obtain genetic data, random amplified polymorphic DNA (RAPD) remains a useful approach (Williams et al., 1990) for several reasons (Lacerda et al., 2002; Magalhães et al., 2007; Brahmane et al., 2008; Dutra et al., 2008; Soares et al., 2008). First, it can be used to evaluate populations for which no specific marker has been developed, allowing then a quick screening of genetic variability. Second, it is a less expensive technique, does not involve complex procedures of DNA sequencing, requires only a simple laboratory with minimum equipment able to perform polymerase chain reaction (PCR); in addition, despite the impossibility to distinguish heterozygotes, several approaches can be used to estimate genetic parameters from dominant data (Meunier and Grimont, 1993; Holsinger et al., 2002; Holsinger and Wallace, 2004; Rjiput et al., 2006).

RAPD, however, has been criticized because it usually shows low levels of repeatability, and because of the basic design of the technique it can potentially generate spurious bands (Pérez et al., 1998; Rabouam et al., 1999). These problems can be overcome, in principle, by using rigid laboratory protocols and doing repeatability tests. The purpose of such tests is to repeat the analyses and retain for further analyses only the bands that appear in both initial and later screenings (Telles and Soares, 2007; Santos et al., 2007).

Repeatability has been usually performed with a small number of individuals and as an *a priori* procedure, as part of a preliminary or pilot study to select the primers and the bands that will be used for further analyses of larger samples (MacPherson et al., 1993; Jones et al., 1997). However, this is not always a good solution because it is unlikely that these very few individuals will possess all possible bands to be found and, consequently, the genetic parameters to be estimated may be biased by the loci that are present in these individuals (although bias can be less if these individuals come from different populations, this is not always the case and will depend on the relative amount of interpopulational variation). On the other hand, it is expensive and time-consuming to obtain genetic data twice for all individuals. Selecting bands with strong staining only

is also difficult, because the same band can appear with different intensities in different individuals (Holsinger et al., 2002; Hardy, 2003; Hill and Weir, 2004).

Here, we present a new procedure to optimize repeatability and avoid bias in sampling loci for genetic analyses based on RAPD data, and demonstrate its application in the evaluation of genetic variability in *Tibouchina papyrus*, an endemic Cerrado plant with a very restricted geographic distribution in elevated rocky fields. Our aim was to propose a new protocol to perform *a posteriori* reproducibility evaluation and to compare genetic parameter estimates using a different approach to select bands.

MATERIAL AND METHODS

A total of 207 individuals were collected from three local samples in Goiás State (Brazil), one situated in Serra dos Pirineus, municipality of Pirenópolis, and two in Serra Dourada, in municipality of Goiás. DNA was extracted from young leaves, using the CTAB (cetyltrimethylammonium bromide) method described by Hillis et al. (1996). The quality of the extracted DNA was evaluated on 1% agarose gels stained with ethidium bromide and quantified by comparison with a DNA ladder (low molecular weight DNA, Invitrogen™). DNA was diluted to a working concentration of approximately 5 ng/µL.

Six RAPD primers (Operon Technologies, Inc.; Alameda, USA) (Table 1) were used for the analysis. PCR was performed in a PCT-100, MJ Research PCR system, with a total volume of 20 µL containing 9.34 µL water, 3 µL genomic DNA (5.0 ng/µL), 2.60 µL 10X PCR Buffer (Invitrogen™), 2.08 µL dNTPs (2.5 mM), 0.78 µL $MgCl_2$ (2.5 mM), 2 µL primer, and 0.2 µL Taq Polymerase (5 U-Invitrogen™). PCR profiles consisting of an initial denaturation of DNA for 3 min at 96°C, 40 cycles of 1 min denaturation at 92°C, 1 min annealing at 37°C and 1 min extension at 72°C, followed by 3 min at 72°C for a final extension.

The amplification products were separated on 1.5% agarose gels stained with ethidium bromide, run in 1X TBE buffer at 120 V for 4 h, in groups with 23 samples (individuals) separated by 100-bp DNA ladder (Amersham Pharmacia Biotech™). Digital images of gels were captured using EDAS120-KODAK, and individual profiles for all gels were scored for the presence (1) or absence (0) of bands (loci), using the 100-bp DNA ladder as a reference for aligning bands of different gels to different loci and to minimize ambiguity in the coding procedure.

Since it is expensive and time-consuming to repeat the analyses for all individuals and check the reproducibility of each band, the problem is how to determine the bands that are "real" markers and that can then be safely used for further statistical analyses. If reproducibility tests were performed *a priori*, the question is irrelevant of course, but our aim here was to check the consequences of *a priori* and *a posteriori* band selection. In the latter case, the question becomes how can the selection of individuals for reproducibility tests be better performed knowing all possible bands? The answer is to select the minimum number of individuals that contain, in combination, all bands that were found and for which it is necessary to check the reproducibility. Obviously, if one individual contains all bands, the solution is trivial and it alone will provide a solution for the problem, whereas if there is a band that appears in a single individual, this individual will thereby be part of the solution. This kind of approach is the same that has been widely used in

conservation biology to find the smallest number of areas to conserve all species of interest, forming an optimum reserve network (see Margules and Pressey, 2000; Diniz-Filho and Telles, 2006).

This problem can be solved using methods of Operation Research (see Possingham et al., 2000; Russel and Norving, 2004), and there are currently several approaches to find the minimum solution. The simplest algorithm is to perform a sequential search, starting with the individual that possesses the most bands and selecting it. The bands that this initial individual contains can be considered as already selected, and then the next individual to be selected must contain the largest number of bands that are not found in the first individual, maximizing the complementarity between them. The procedure continues until all individuals are selected, forming a solution of length *k* (number of individuals). This simple method, although easy to implement, does not provide necessarily the optimum solution (i.e., finding the smallest possible number of individuals), and more sophisticated methods, using exact mathematical solutions or intense computational procedures based on artificial intelligence, can be used. One of them is "Simulated Annealing", which is a global optimization meta-heuristic method that, starting with a random configuration of objects (i.e., individuals), probabilistically decides to add or delete individuals from this configuration and iterates the procedure minimizing a given quantity (in this case, the loss in number of bands) (Russel and Norving, 2004).

We applied the Simulated Annealing procedure as implemented in SITES, a software used in reserve design procedure (see Possingham et al., 2000) to find the smallest number of individuals that (when combined) contain all bands. Since PCR will have to be done independently for each primer for reproducibility tests, the analyses were performed by primer. Most importantly, in large matrices and with relatively simple representation problems, there is usually no unique solution to the problem, i.e., there may be more than one combination of *k* individuals that contain all bands. Although all solutions are mathematically equivalent (i.e., will allow achieving the goal of representing all bands with smallest possible number of individuals), it is also possible to calculate the frequency of each individual in the *p* minimum solutions, a quantity usually called irreplaceability. As previously pointed out, if a band appears in a single individual, it will appear necessarily in all *p* minimum solutions if the optimum length *k* is to be achieved, so that irreplaceability of the individual is equal to 1 (or 100%).

We applied the simulated annealing to presence-absence of bands for the six primers, and analyzed the data before and after removing such bands (unreduced and reduced data sets). For comparison, we also "simulated" the standard approach for band selection by analyzing only those bands that show reproducibility for 3 individuals, one randomly chosen in each population.

We analyzed the three RAPD datasets using the HICKORY v. 1.0 program (Holsinger and Lewis, 2003), which estimates the genetic parameters from dominant markers using the Bayesian approach proposed by Holsinger et al. (2002), avoiding explicit assumptions about Hardy-Weinberg equilibrium in the local populations. Parameters were approximated numerically through a Markov Chain Monte Carlo simulation, which tends to converge to a β distribution of estimators. We estimated in HICKORY v. 1.0 (Holsinger and Lewis, 2003) the divergence among local populations using θ^B (the estimate of F_{ST}), the inbreeding coefficient *f* (the F_{IS}) and the genetic diversity Hs for each local population. All these quantities were estimated by the HICKORY full model, whose performance was checked by deviance information criterion. We also analyzed the same data sets using the POPGENE 1.32

program (population genetic analysis) (Yeh and Boyle, 1997) and estimated the proportion of polymorphic loci (P%), Nei's genetic diversity (He) and Shannon genetic diversity (S), assuming Hardy-Weinberg equilibrium. Although these estimates are not entirely adequate for dominant markers, they are still widely used (e.g., Artiukova et al., 2004; Ferreyra et al., 2004; Basavaraju et al., 2007)

RESULTS AND DISCUSSION

The six RAPD primers selected (OPB-04, OPB-05, OPB-10, OPM-03, OPM-05, and OPM-13) provided initially a total of 176 bands (loci), ranging from 29 to 32 per primer. The simulated annealing found that the minimum number of individuals to represent all loci varied between 3 and 5 (Table 1), and the number of solutions found for each one ranged from 11 to 146. As expected, most individuals had low irreplaceabilities and only 27% of the irreplaceabilities were different from zero (and only three individuals had irreplaceability equal to 1.0), and thus, a strongly right-skewed distribution of these values appeared. One of the solutions (i.e., the one with the most irreplaceable individuals) was then chosen and PCR and RAPD analyses were repeated only for these few individuals that contained all bands for the primer, and by comparing the bands in the two RAPD solutions, the original matrix was reduced from 176 to 147 bands (84% reproducibility), with variable band reduction for primer (Table 1). When randomly selecting three individuals for *a priori* reproducibility analysis, 120 bands remained in the data set (68% reproducibility).

Table 1. Total number of bands (TNB) originally found for each primer, smallest number of individuals necessary to test all bands (*k*) and number of minimum solutions found by simulated annealing (*p*) and number of bands retained for analyses after repeatability using individuals selected based on simulated annealing (NBSA) and based on three randomly selected from each population (NBR).

Primer	Sequence 5'→ 3'										TNB	*k*	*p*	NBSA	NBR
OPM-03	G	G	G	G	G	A	T	G	A	G	29	4	55	25	23
OPM-05	G	G	G	A	A	C	G	T	G	T	30	5	11	23	18
OPM-13	G	G	T	G	G	T	C	A	A	G	27	3	146	23	18
OPB-04	G	G	A	C	T	G	G	A	G	T	29	4	145	26	19
OPB-05	T	G	C	G	C	C	C	T	T	C	32	4	45	24	20
OPB-10	C	T	G	C	T	G	G	G	A	C	29	4	79	26	22
Total											176	–	–	147	120

The genetic parameters estimated for the three different data sets are very similar, mainly because they actually originated from the same full-data set (Table 2). Anyway, parameters estimated for the reduced data set based on simulated annealing are much more similar to the full-data set than those estimated based on the three individuals only, and the percentage of change when using selection by simulated annealing is half the one using the three individuals for θ^B and *f*, and much higher for within-population metrics for genetic diversity. In the Shannon estimate, for example, the average change when reducing to 147 bands (based on simulated annealing) was equal to 1.7% versus a change of 11.2% when selection was based on three individuals. Another interesting point is that estimates from HICKORY seem to be less affected than those from POPGENE.

Table 2. Percentage of polymorphic loci (P%) and genetic diversity (He) estimated based on different bands, selected by different criterion.

Dataset	Population	P (%)	Hs (± SD)	He (± SD)	Shannon (± SD)	f	θ^B
TNB	1	93	0.3056 (± 0.0038)	0.245 (± 0.159)	0.386 (± 0.212)	0.9711 (± 0.0293)	0.1731 (± 0.0107)
	2	95	0.3372 (± 0.0034)	0.286 (± 0.158)	0.438 (± 0.209)		
	3	97	0.3491 (± 0.0036)	0.317 (± 0.163)	0.477 (± 0.209)		
NBSA	1	93	0.3018 (± 0.0041)	0.246 (± 0.159)	0.387 (± 0.212)	0.9624 (± 0.0361)	0.1811 (± 0.122)
	2	95	0.3429 (± 0.0037)	0.295 (± 0.155)	0.451 (± 0.204)		
	3	97	0.3493 (± 0.0039)	0.326 (± 0.161)	0.487 (± 0.206)		
NBR	1	94	0.324 (± 0.0043)	0.281 (± 0.163)	0.431 (± 0.215)	0.928 (± 0.0545)	0.1905 (± 0.0132)
	2	98	0.3599 (± 0.0042)	0.323 (± 0.143)	0.488 (± 0.181)		
	3	98	0.3683 (± 0.0042)	0.358 (± 0.150)	0.527 (± 0.192)		
NBSA - TNB (%)	-	0.000	0.168	2.131	1.775	-	-
NBSA - NBR (%)	-	1.755	6.084	13.522	11.185	-	-

For abbreviations, see legend to Table 1.

Because of the large number of loci retained after repeatability analyses, the genetic parameter estimates are very similar when using the optimization procedure reducing the number of bands analyzed. Even when using *a priori* reduction based on 3 individuals, a total of 120 bands are still retained for analysis. Even so, as expected, selecting only three randomly chosen individuals to evaluate the repeatability of bands generates more underestimated genetic parameters because not all possible bands are considered. These results show that, at least in part, the problems of repeatability attributed to RAPD markers, in terms of variation in estimates of genetic parameters when using all loci and only those with high repeatability, can be due to bias in the selection of loci and primers and not necessarily to RAPD *per se*. More importantly, when methods especially designed to deal with dominant markers are used, bias is less and thus this reinforces that, although more sophisticated molecular markers are available, RAPD data are still a viable solution to genetic analysis if data are carefully obtained, checked and adequately analyzed. Thus, our procedure provides an unbiased way to select primers and improve the quality of RAPD analyses, which can still be safely used until more sophisticated markers are available for more species and can be produced at much lower costs.

REFERENCES

Artiukova EV, Kholina AB, Kozyrenko MM and Zhuralev I (2004). Analysis of genetic variation in rare endemic species *Oxytropis chankaensis* Jurtz. (Fabaceae) using RAPD markers. *Genetika* 40: 877-884.

Basavaraju Y, Prasad DT, Rani K, Kumar SP, et al. (2007). Genetic diversity in common carp stocks assayed by random-amplified polymorphic DNA markers. *Aquac. Res.* 38: 147-155.

Brahmane MP, Mitra K and Mishra SS (2008). RAPD fingerprinting of the ornamental fish *Badis badis* (Hamilton 1822) and *Dario dario* (Kullander and Britz, 2002) (Perciformes, Badidae) from West Bengal, India. *Genet. Mol. Biol.* 31: 789-792.

Diniz-Filho JAF and Telles MPC (2006). Optimization procedures for establishing reserve networks for biodiversity conservation taking into account population genetic structure. *Genet. Mol. Biol.* 29: 207-214.

Dutra NC, Telles MP, Dutra DL and Silva Junior NJ (2008). Genetic diversity in populations of the viper *Bothrops moojeni* Hoge, 1966 in Central Brazil using RAPD markers. *Genet. Mol. Res.* 7: 603-613.

Ferreyra LI, Bessega C, Vilardi JC and Saidman BO (2004). First report on RAPDs patterns able to differentiate some Argentinean species of section Algarobia (Prosopis, Leguminosae). *Genetica* 121: 33-42.

Hardy OJ (2003). Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. *Mol. Ecol.* 12: 1577-1588.

Hill WG and Weir BS (2004). Moment estimation of population diversity and genetic distance from data on recessive markers. *Mol. Ecol.* 13: 895-908.

Hillis DM, Moritz C and Mable BK (1996). Molecular Systematics. Sinauer Associates, Sunderland.

Holsinger KE and Lewis PO (2003). HICKORY v. 1.0. Department of Ecology & Evolutionary Biology, University of Connecticut, Storrs. Available at [http://www.eeb.uconn.edu/].

Holsinger KE and Wallace LE (2004). Bayesian approaches for the analysis of population genetic structure: an example from *Platanthera leucophaea* (Orchidaceae). *Mol. Ecol.* 13: 887-894.

Holsinger KE, Lewis PO and Dey DK (2002). A Bayesian approach to inferring population structure from dominant markers. *Mol. Ecol.* 11: 1157-1164.

Jones CJ, Edwards KJ, Castaglione S, Winfield MO, et al. (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. *Mol. Breed.* 3: 381-390.

Lacerda DR, Acedo MDP, Filho JPL and Lovato MB (2002). A técnica de RAPD: uma ferramenta molecular em estudos de conservação de plantas. *Lundiana* 3: 87-92.

MacPherson JM, Eckstein PE, Scoles GJ and Gajadhar AA (1993). Variability of the random amplified polymorphic DNA assay among thermal cyclers, and effects of primer and DNA concentration. *Mol. Cell. Probes* 7: 293-299.

Magalhães M, Martinez RA and Gaiotto FA (2007). Genetic diversity of *Litopenaeus vannamei* cultivated in Bahia State, Brazil. *Pesq. Agropec. Bras.* 42: 1131-1136.

Margules CR and Pressey RL (2000). Systematic conservation planning. *Nature* 405: 243-253.

Meunier JR and Grimont PA (1993). Factors affecting reproducibility of random amplified polymorphic DNA

fingerprinting. *Res. Microbiol.* 144: 373-379.
Pérez T, Albornoz J and Domínguez A (1998). An evaluation of RAPD fragment reproducibility and nature. *Mol. Ecol.* 7: 1347-1357.
Possingham H, Ball I and Andelman S (2000). Mathematical Methods for Identifying Representative Reserve Networks. In: Quantitative Methods for Conservation Biology (Ferson S and Burgman M, eds.). Springer-Verlag, New York, 291-306.
Rabouam C, Comes AM, Bretagnolle V, Humbert JF, et al. (1999). Features of DNA fragments obtained by random amplified polymorphic DNA (RAPD) assays. *Mol. Ecol.* 8: 493-503.
Rajput SG, Wable KJ, Sharma KM, Kubde PD, et al. (2006). Reproducibility testing of RAPD and SSR markers in Tomato. *African J. Biotechnol.* 5: 108-112.
Russel SJ and Norving P (2004). Inteligência Artificial. Elsevier, Rio de Janeiro.
Santos RP, Angelo PCS, Quisen RC, Oliveira CL, et al. (2007). RAPD em Pau-rosa (*Aniba rosaeodora* Ducke): adaptação do método para coleta de amostras *in situ*, ajuste das condições de PCR e apresentação de um processo para selecionar bandas reprodutíveis. *Acta Amazonica* 37: 253-260.
Soares TN, Chaves LJ, de Campos Telles MP, Diniz-Filho JA, et al. (2008). Landscape conservation genetics of *Dipteryx alata* ("baru" tree: Fabaceae) from Cerrado region of central Brazil. *Genetica* 132: 9-19.
Telles MPC and Soares TN (2007). DNA Fingerprinting no Estudo de Populações de Plantas do Cerrado. In: Recursos Genéticos e Conservação de Plantas Medicinais do Cerrado (Pereira AMS, ed.). Editora Legis Summa; FAPESP, Ribeirão Preto, 109-145.
Williams JG, Kubelik AR, Livak KJ, Rafalski JA, et al. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* 18: 6531-6535.
Yeh FC and Boyle TJB (1997). Population genetic analysis of co-dominant and dominant markers and quantitative traits. *Belg. J. Bot.* 129: 157. Popgene version 1.32. Available at [http://www.ualberta.ca/~fyeh/download.htm]. Accessed March 2007.

