Hindawi Publishing Corporation
ISRN Forestry
Volume 2013, Article ID 747591, 7 pages
http://dx.doi.org/10.1155/2013/747591

Research Article

Genetic Structure of a Loblolly Pine Breeding Population at Brazil

Juliane Rezende Mercer,[1] Milena de Luna Alves Lima,[2] Antonio Rioyei Higa,[2] Chirlei Glienke,[1] and Marina Isabel Mateus de Almeida[1]

[1] *Departamento de Genética, Universidade Federal do Paraná (UFPR), Setor de Ciências Biológicas, 81531-980 Curitiba, PR, Brazil*
[2] *Setor de Ciências Agrárias, Universidade Federal do Paraná (UFPR), Curso de Engenharia Florestal, Avenida Professor Lothario Meissner, 900, Jardim Botanico, 80210-170 Curitiba, PR, Brazil*

Correspondence should be addressed to Juliane Rezende Mercer; juju.mercer@hotmail.com

Received 1 April 2013; Accepted 23 April 2013

Academic Editors: G. Martinez Pastur, P. Newton, and H. Zeng

The genetic structure of a Brazilian loblolly pine (*Pinus taeda* L.) breeding population, represented by 120 open-pollinated families, was determined using Bayesian inference and genotypes of 15 microsatellite (simple sequence repeat (SSR)) *loci* in 1,130 seedling progeny. The 120 maternal parents had been phenotypically selected about 15 years ago for wood volume in five different forestry plantations (FPs) in the south of Brazil. Additional selection for wood volume, based on a previous progeny test, was applied to the first best (i) and second best (ii) tree per block within each family. We adopted a procedure of "learning samples" to find the most likely number of inferred genetic clusters (K) or ancestral populations. The first hypothesis that was rejected was that the most probable value of $K = 5$ was coincident with the five FPs, since the FPs were, *a priori*, assumed to be from 5 different backgrounds or origins. It was used the familiar structure of the population to infer the genotypes of maternal ancestors. It was concluded that the maternal generation is the most likely to have been planted by the mixture of three different seed sources or origins, that there are five genetic groups ($K = 5$) in the population of progeny, and that they have been formed from the occurrence of assortative mating and also from a strong pressure in the selection within families. The trees with the best genetic value (i) maintained a higher genetic variability when compared to the trees of second best performance (ii), with higher values of heterozygosity and of numbers of maternal alleles that were kept the same. The migration model that best explains the results is the contact zone model. The population differentiation (F_{ST}) was 2-3 times higher in offspring than in relation to the maternal generation. The relevancy of the results and the way they were explored may be of value both for studies of population genetics, as for plant breeding programs, since they help monitoring the population's genetic variability during generations of selection.

1. Introduction

Loblolly pine (*Pinus taeda* L.) is a monoecious conifer, diploid tree with a predominant cross-fertilization. Its native range is in southeastern United States, and it was first introduced in Brazil in the 1940s. Its wood is used for industry panels and sawn timber. It is a fast-growing tree and is well adapted to temperate climate that makes it a good candidate for planting forests in southern Brazil. Among 5.74 million hectares (ha) of all existing forest plantations in Brazil in 2006, *P. taeda* occupies 1.52 million hectares in the southern states [1]. Most of these plantations were established from genetically improved seeds produced in clonal seed orchards that were established by breeding programs using phenotypically selected trees.

The main objective of this research was to study the genetic structure of a breeding population that consists of open-pollinated progenies of multiple *P. taeda* families using microsatellite or simple sequence repeats (SSR) markers and a Bayesian clustering approach implemented in the STRUCTURE software (Pritchard et al. [2]) to infer groups or subpopulations in this population.

The software calculates log likelihoods for user-defined number (K) of subpopulations. Individuals in the sample are probabilistically assigned to a subpopulation "cluster" or partitioned between two or more subpopulations "clusters", if their genotypes indicate that they are hybrids or admixtures [3]. The model choice criterion implemented in structure to detect the true K is an estimate of the posterior probability of the data for a given K, called the $\Pr(X \mid K)$ [3] and called $\ln P(D)$ in Evanno et al. [4].

It allows to infer the most likely number of K based on maximum likelihood estimates $\ln P(D)$. However, this analysis greatly depends on the modeling assumptions, such as Hardy-Weinberg equilibrium and others. It was observed as in Evanno et al. [4] simulations that, in most cases, once the real K is reached, $\ln P(D)$ stabilizes at larger Ks plateaus or continues increasing slightly. This phenomenon is also reported in the structure's manual (Pritchard et al. [3]), in which the variance between runs increases. Evanno et al. [4] proposed an *ad hoc* statistic (ΔK) based on the rate of change of maximum likelihood estimates computed by the STRUCTURE software. The rational for this ΔK is to make salient the break in slope for the distribution of $\ln P(D)$ at the true K.

However, selfing can lead to spurious clustering of structured population using the standard STRUCTURE algorithm that can bias admixture estimates. Therefore, Gao et al. [5] extended the Bayesian clustering approach of STRUCTURE [3] to simultaneous inference of inbreeding or selfing rates and classification of the source population using multilocus genetic data. The Hardy-Weinberg equilibrium is no longer assumed within groups, and the expected genotype frequencies are estimated based on rates of inbreeding or selfing.

Unaccounted population structure can lead to false-positive spurious associations between traits and polymorphic alleles in association mapping studies. That is why it is very important to analyze population structure and take it into account in association mapping studies. As a next goal in this research aims to conduct an association mapping to detect genetic variation between strong candidate genes and some wood quality traits, the present approach assumes great relevance.

2. Material and Methods

Breeding population used in this study consisted of 12-year-old seedlings representing open-pollinated progenies from 120 families planted in a complete block design with five repetitions. The total numbers of trees studied were 1.130. Seeds for these seedlings were collected about 15 years ago in one of five forestry plantations (FPs) in Santa Catarina, a Brazilian southern state. Geographic origins of these FPs are unknown. The seeds were collected from 120 maternal trees that had been phenotypically selected for wood volume at the five FPs—the seeds represent the first round of selection within families.

The needles were collected from 1130 individual seedlings, stored at −20°C and then used to isolate DNA using the CTAB method [6]. *Microsatellite loci* were genotyped using MegaBace 1000 (GE Healthcare), and the multiplex system protocol was established during this research. The following microsatellites loci (and their respective Genbank accession numbers) were genotyped in the study: *PtTX3026* (AF143971), *PtTX3002* (AF277846), *PtTX3088* (AF277843), *PtTX3091* (AF277848), *PtTX3098* (AF277847), *PtTX3118* (AF277845), *PtTX2037* (AF143959), *PtTX3025* (AF143970), *RPtest5* (BV728798), *2n11e* (AA556153), *PRtest8* (BV728799), *PRtest11* (BV728796), *8934 M* (AA739818), *9317 M* (AA740072), and *PRtest1* (BV728795).

2.1. Statistical Analyses. The segregation of microsatellite loci was tested by inferring the maternal genotype for each open-pollinated family. Linkage and Hardy-Weinberg disequilibrium analyses were done using GDA 1.1 http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php) [7]. The population structure was studied using Bayesian cluster inference based on 15 microsatellite loci (SSRs) genotyped in 1130 seedlings, under two scenarios, assumin Hardy-Weinberg equilibrium (software STRUCTURE 2.3; http://pritch.bsd.uchicago.edu/structure.html) and assuming inbreeding (software INSTRUCT; http://cbsuapps.tc.cornell.edu/InStruct.aspx). The permutation analysis for testing assignment of individual trees to clusters was done using CLUMPP 1.1.2, a Cluster Matching and Permutation Program [8]. An *ad hoc* statistic suggested by Evanno et al. [4] was used to analyze the STRUCTURE results.

A preliminary analysis with 20 independent runs was done to determine the number of burn-in and the number of repetitions for the STRUCTURE software using the admixture model and correlated allele frequencies assumptions. It was observed that the posterior distributions, respectively, converge with burn-in and MCMC of 10,000 and 50,000 cycles, respectively. The *ad hoc* statistic of Evanno et al. [4] was also computed for these 20 runs for $K = 1$ to 21.

Before implementing the EVANNO *ad hoc* statistic, an exploratory study was conducted to determine the set of values for the logarithms of likelihood represented by the posterior probabilities $[\Pr(X \mid K)]$. It was concluded that the arithmetic mean was not a good estimator for the observed data dispersion, and it was concluded that the median values provided a better fit (data will be soon published). Thus, a change in the authors' equations [4] was implemented, which basically consisted in using the median values instead of the arithmetic mean of $[\Pr(X \mid K)]$. Therefore, instead of using the ΔK introduced by the authors, it was referred here as ΔK_m.

The equations used by the STRUCTURE software to measure population substructure F_{ST}, expected heterozygosity (H_e), coancestry coefficients (q), and genetic distance using the allelic frequencies are fully described in [3, 9]. The equations and the inference of genetic parameters used by the INSTRUCT software are fully described in Gao et al. [5].

The CLUMPP 1.1.2 software with 10,000 permutations for the 30 Q matrices generated in 30 runs of the STRUCTURE 2.2 software was used to perform permutations that test assignment of each seedling to a particular cluster and also to determine admixtures.

Based on the initial analysis, the mode 5 of the INSTRUCT software proved to have the best performance for

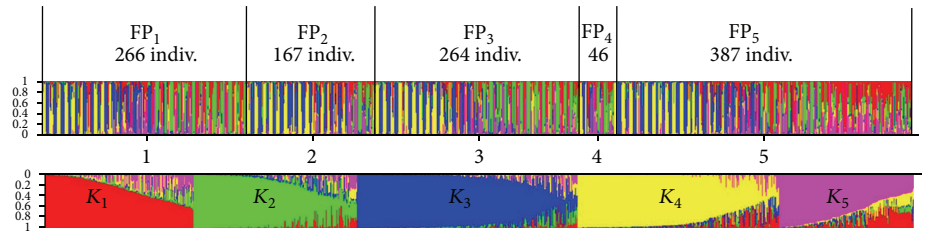

Figure 1: Distribution of the five genetic clusters (K) in each of the five forestry plantations (FPs). Upper graph: individual seedlings arranged in five FPs; lower graph: distribution of coancestry (Q) for five clusters in 1.130 seedlings.

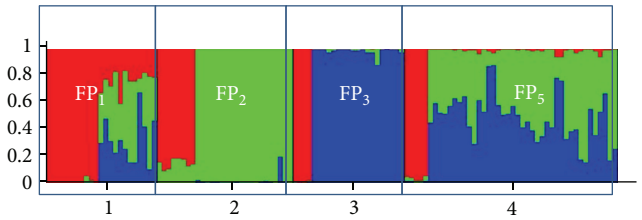

Figure 2: Analyses for the admixtures detected in previous analysis for $K = 5$. The *nonadmixture* seedlings and data from FP_4 were excluded. For the present analysis, the best posterior probability was $K = 3$.

Table 1: Descriptive statistics for the five forestry plantations (FPs).

Population	N	n	n, %	P	A	H_e	H_o	f
FP_1	266	163.46	61.5	1.00	4.62	0.49	0.33	0.32
FP_2	167	108.69	65.1	1.00	4.92	0.51	0.38	0.26
FP_3	264	173.54	65.7	1.00	4.54	0.51	0.35	0.31
FP_4	46	29.38	63.9	1.00	3.92	0.51	0.34	0.34
FP_5	387	245.62	63.5	1.00	4.77	0.48	0.31	0.35
Mean	226	144.14	63.8	1.00	4.55	0.50	0.34	0.32
Total	1130	720.69						

N: number of individuals in each FP; n: number of admixture individuals or individuals that had less than 95% of coancestry to a determined cluster; P: the proportion of polymorphic loci; A: the mean number of alleles per locus; H_e: the expected heterozygosity; H_o: the observed heterozygosity; f: estimate of the fixation index using the method of moments in GDA.

this population. This mode has estimated rates of selfing or inbreeding using back reflection. A hundred runs of the INSTRUCT were performed to estimate the parameters of inbreeding and clustering that provide the best probabilities of likelihood for the posterior K. The numbers of burn-in and MCMC iterations were 10,000 and 50,000, respectively, and 10 for the thinning period.

A series of hypothesis for the subsequent analysis was constructed, similar to a methodology of learning samples. These exploratory data analyses were an approach to increase the analyzing data sets, to observe their main characteristics in an easy-to-understand form, in order to formulate hypotheses that could be tested on new data sets.

3. Results and Discussion

The STRUCTURE analysis was run first using the admixture model with correlated allele frequencies, with $K = 5$, and no *a priori* information about the population. The objective was to test the hypothesis that FPs are distinct genetic groups. The results did not indicate the homogeneity of FPs. Each FP consisted of five groups or clusters. The analyses that were performed by informing software about the FP origin confirmed the previous results (Figure 1).

These results indicate that each FP consists of five genetic clusters in different proportions. There is almost no difference in genetic diversity between the FPs, numbers of polymorphic alleles, and expected and observed heterozygosities were similar between FPs (Table 1).

To verify the gene flow between the FPs, the analyses considered the FP origin information, but it was settle as not informed for the individuals classified as being admixture in the first analysis. The results estimated the confidence interval of the admixture individuals' coancestry to each of the five genetic clusters and also the probabilities for the number of past generations in which the hybridization event have happened (parents, grandparents, and great grandparents). The results indicated that 60% of admixtures were generated in the parental generation during the maternal ancestral pollination in the forestry plantations.

To study the admixtures effects, all the nonadmixture individuals resulted from previous analysis, and also the smallest forestry plantation population (FP_4) were excluded from the analysis. The best posterior probability for this dataset resulted in $K = 3$ (Figure 2). The most distinct forestry populations were FP_2 and FP_3; the majority of FP_5 are hybrid between FP_2 and FP_3. Although half of the individuals in FP_1 are the same hybrids, this population have a distinct genetic group, which also appears in the other four forestry plantations and almost did not hybridize among the other genetic groups. However, deeper analyses in order to test this hypothesis of gene flow could not be stated since no information was provided about the geographical origin of the forestry plantations.

The results of the STRUCTURE are presented in Table 2. The best [$\Pr(X \mid K)$] was detected for $K = 19$ (−12, 238.4), but the most appropriate value, according to the literature, although arbitrary, would be the smallest number of cluster that captured most of the genetic structure; after that, it occurs an increasing pattern for the next K's [$\Pr(X \mid K)$]

Table 2: The posterior probability [Pr ($X \mid K$)] on average for 10 consecutive analyses of each K *a priori*. (K = 1 to 20). Wright's F_{ST} and inferred values of alpha.

K	ln Pr ($X \mid K$)	Alpha	F_{ST1}	F_{ST2}	F_{ST3}	F_{ST4}	F_{ST5}	F_{ST6}
1	−16672.1	—	0.1080					
2	**−15056.9***	**0.0666**	**0.1620**	**0.1878**				
3	−14416.8	0.0473	0.192	0.169	0.295			
4	**−13541.4**	**0.0427**	**0.3256**	**0.2054**	**0.3522**	**0.2757**		
5	**−13337.3**	**0.0415**	**0.3186**	**0.3885**	**0.3017**	**0.3778**	**0.1964**	
6	−13084.8	0.0418	0.3037	0.1942	0.3892	0.3861	0.3701	0.3786
7	−12814.6	0.0402	0.4003	0.4197	0.3179	0.4532	0.3536	0.3944
8	**−12700.5**	**0.0395**	**0.2619**	**0.2847**	**0.4235**	**0.5584**	**0.4340**	**0.4157**
9	−12553.3	0.0380	0.5705	0.3574	0.4223	0.4284	0.3811	0.5048
10	−12937.5	0.0384	0.4098	0.3624	0.5221	0.4361	0.4866	0.4091
11	−12469.0	0.0374	0.4400	0.3308	0.4720	0.4553	0.5905	0.4737
15	−12319.5	0.0367	0.5479	0.4228	0.4794	0.5518	0.5449	0.3863
19	**−12238.4**	**0.0372**	**0.3775**	**0.5804**	**0.5108**	**0.4279**	**0.4106**	**0.5439**
20	−12693.2	0.0374	0.4783	0.5684	0.5109	0.4521	0.4646	0.5389

*Bold numbers are the best posterior values.

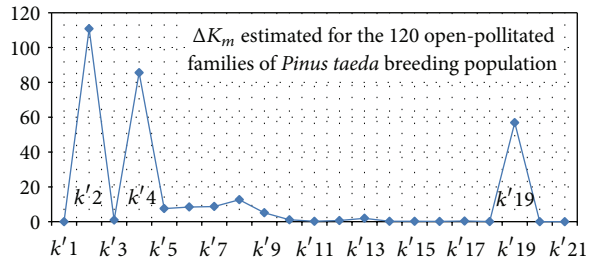

Figure 3: The graph with the results for ΔK_m and the highest peak detected for K_2.

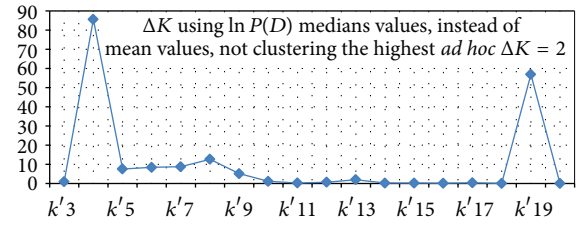

Figure 4: ΔK_m using medians values. Not including K_2 (the strongest signal) but including the second higher peaks at K_4, K_{19}, and K_8, in analyses of ΔK_m.

until a plateau is reached. Nonetheless, K = 4 and K = 5 had similar results; here, K = 5 (−13337.3) was considered. The adjustments of alpha and F_{ST} values during the burn-in MCMC, resulted in the best posterior distribution analysis for K_1, K_5, and K_{19}. Note that for K = 5, they showed the best adjustments.

Evanno et al. [4] proposed a statistic to determine the most probable number of K that could be graphically interpreted. In Figure 3, the graph depicts the results for ΔK_m, which demonstrates that the highest peak is for K = 2. In Figure 4, there is the same graphic with K_2 not included, in order to have a better visualization of the second highest peaks detected for K_4, K_{19}, and K_8.

The greatest magnitude of ΔK_m was at K_2 with smaller peaks at K_4 and K_{19}. Figure 5 summarizes the results. These results came from the analysis that leads to a stepping stone model, expressing the distance in allele frequency divergence among all the 19 substructures detected. This graphic was created by plotting the distances among all the 19 substructures that comprise the five major genetic groups. On the x-axis, were plotted the genetic distances among the three most similar genetic groups, and on the y-axis, the distances of the two most divergent genetic groups in relation to those three similar groups. This same procedure was performed within each of the five main groups. This artifice generated a better interpretation of the genetic distances results, making it more visually comprehensible.

Though the results from the two *ad hoc* statistics seem different, after performing an analysis within groups and subgroups, the conclusion is that both of them lead to the same results. Hence, it was assumed that five may be the most real number of genetic groups to represent the breeding population.

3.1. Genetic Structure of the Maternal Generation. The F_{ST} coefficient for the five clusters of the breeding population ranged from 0.19 to 0.42 (Table 3) indicating a very differentiated population. For the genetic structure detected in the maternal generation it was found that the most real number of genetic groups is K = 3, and F_{ST} coefficient ranged from 0.01 to 0.15 (Table 3).

3.2. Bayesian Analysis Assuming Inbreeding. The best posterior probability in the analyses of INSTRUCT software and assuming inbreeding was, again, obtained for K = 5, with

Table 3: F_{ST} values and expected heterozygosity (H_e) between the three maternal clusters and the five genetic clusters of the breeding population.

F_{ST1}	H_{e1}	F_{ST2}	H_{e2}	F_{ST3}	H_{e3}	F_{ST4}	H_{e4}	F_{ST5}	H_{e5}
120 mothers ($K = 3$) [$\ln P(D) = -3184.9$; Var [$\ln P(D)$] = 242.7; $\alpha = 0.188$]									
0.01	0.5	0.13	0.53	0.16	0.45				
1130 OP indv. ($K = 5$) [$\ln P(D) = -13263.4$; Var [$\ln P(D)$] = 1120.2; $\alpha = 0.046$[3]]									
0.19	0.5	0.42	0.38	0.32	0.38	0.24	0.44	0.30	0.41

the lower value for the log likelihood in mean and in variance (mean and variance were −6855.3 and 13710.5, resp.).

The result of the Bayesian analysis under inbreeding that had the best convergence for the Markov chain is being considered. This is determined by the value of the Gelman-Rubin (GR) statistic. This statistic measures the convergence of the log likelihood. The value for this chain was 1.022.

The distribution of the five resulting clusters using INSTRUCT was similar to the STRUCTURE, indicating again that forestry plantations are a mosaic of all five groups and that were coincident in analysis assuming Hardy-Weinberg (STRUCTURE) and assuming inbreeding (INSTRUCT).

STRUCTURE generates clusters based on both transient Hardy-Weinberg disequilibrium (HWD) and linkage disequilibrium (LD) caused by admixture between populations. The program works by clustering individuals in groups in which both linkage and HWD are minimized, and, as a result, the presence of LD in the data improves clustering results [9]. On the other hand, strong LD or departure from Hardy-Weinberg equilibrium could lead to an overestimation of the number of detected clusters [9]. Since almost all microsatellite *loci* in this study departure from Hardy-Weinberg equilibrium, it was expected that those $K = 5$ were overestimated and, if so, INSTRUCT would detect a lower number of genetic clusters. However, it did not occur. In fact, the results indicated the same most probable K in both methods.

In the STRUCTURE analysis the F_{ST} varied from 0.19 to 0.42 for 1.130 individuals and 15 microsatellite *loci*. In the INSTRUCT analysis that assumes this inbreeding, the F_{ST} values ranged from 0.022 to 0.227; however, these results were analyzed with only 738 individuals (that have the lower number of missing data for the genotyped alleles) from the existing 1.130, and with a reduction of 15 to 10 in the number of microsatellite *loci* (by excluding the *loci*: 01, 06, 08, 17, and 21). This has occurred because the stability of the convergence for the log of likelihood of the posterior distributions of K occurred for this dataset configurations in the INSTRUCT analysis. Table 4 summarizes this result. The F_{ST} values assuming inbreeding were lower than assuming Hardy-Weinberg.

Thus, this result confirms that the population consists of five groups or clusters under both assumptions, Hardy-Weinberg equilibrium and inbreeding. These results may indicate that the reduction in expected heterozygosity is probably caused by the occurrence of higher rates of selfing and assortative mating.

In a publication [10], as well as in other references [11], F_{ST} values from 0.01 to 0.05 would be expected for open-pollinating populations of *conifer* of open cross-pollinating under natural conditions and from 0.05 to 0.30 for divergent populations [11]. The values obtained in this study indicate the high degree of differentiation in this population of breeding.

Pinus taeda has historically large, interconnected populations which extend along the US Atlantic coast, from Maryland to Florida and from westward to central Texas. The loblolly pine area was divided also into five regions with a high level of gene flow within each region obtained in earlier studies [12]. The origin of the maternal trees is unknown, and the five Ks detected in the present study cannot be directly linked to the five regions detected in [12]. However, such coincidence is unlikely to occur by chance alone and needs further investigation.

4. Conclusions

The Bayesian method of clustering analysis based on the 15 microsatellite *loci* helped to study the genetic structure of the breeding population. Based on the observed results, it is possible to conclude that the five forestry plantations, used as the seed sources to establish a breeding population in this study, were a mosaic of 3 genetic groups (the maternal generation), and after one generation of open-pollination at the FPs, the genetic differentiation became even higher, leading to the five genetic groups determined at the breeding population.

Both Bayesian analyses that assumed Hardy-Weinberg equilibrium and inbreeding resulted in five consensus genetic groups that most likely represent the breeding population.

The F_{ST} values that assumed inbreeding were lower than the ones that assumed Hardy-Weinberg equilibrium. Hence, the definition for the genetic parameters of F_{ST} and H_e will be considered from the Hardy-Weinberg equilibrium analyses, which is $F_{ST} = 0.194–0.384$ and $H_e = 0.38–0.50$.

It is possible that the highly fragmented breeding population detected in the Bayesian analyses has first occurred not only due to the selection pressure within open-pollinated progeny when the program selected the first and second best trees for wood volume but also due to the occurrence of assortative mating phenomenon.

If the higher inbreeding levels detected are real, this phenomenon should be observed with caution, as they are not expected in outcrossing plants, and narrowing the genetic base of the population may represent a risk in advancing

Table 4: Comparing F_{ST} values of the inferred maternal genotypes and the open-pollinated progenies assuming Hardy-Weinberg equilibrium (STRUCTURE, 1st and 2nd lines, resp.) and assuming inbreeding (INSTRUCT, 3rd line).

Independent structure analyses	K	ln $P(D)$	Var[ln $P(D)$]	F_{ST1}	F_{ST2}	F_{ST3}	F_{ST4}	F_{ST5}
120 maternal trees	3	−3184	242.7	0.099	0.126	0.156		
1130 seedlings	5	−13353	1108.0	0.375	0.326	0.384	0.194	0.290
738 seedlings	5	−6905,2	13810.4	0.170	0.227	0.212	0.022	0.045

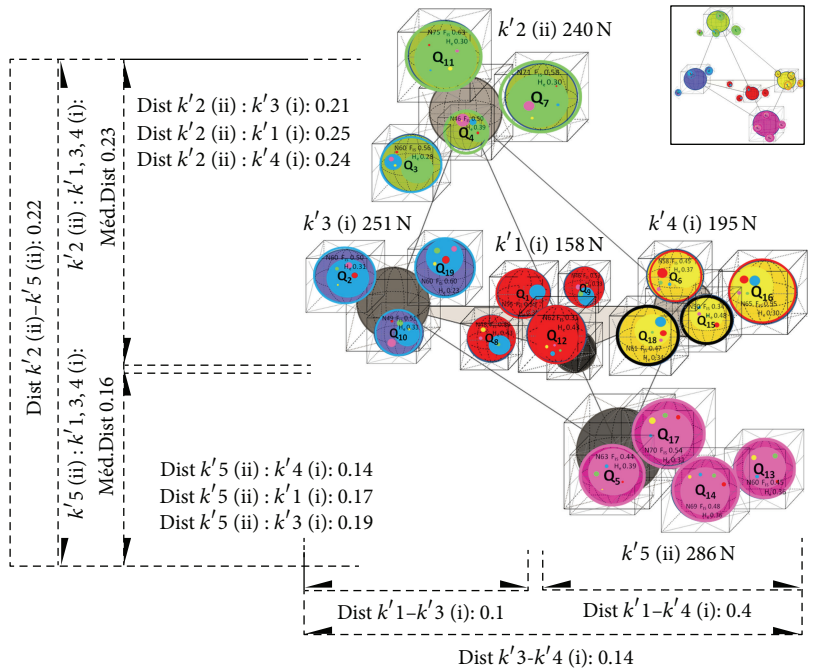

Figure 5: Migration model proposed based on the distribution of the five genetic groups ($K = 5$) and how is it distributed within 19 subpopulations.

generations of selection, particularly in the long-term genetic breeding.

What draws attention in this study was the relevance of the applied methodology on the data analyses, which means that, by implementing both the Bayesian algorithms in a learning sample approach, it has allowed a richness in the data interpretation that might be undetected, if routine analysis was applied instead, as commonly observed in similar studies.

Conflict of Interests

The authors declare that they have no competing interests.

Disclosure

The population samples used in the study are legal property of a Brazilian private program and were previously integrated in a major forestry breeding funded project, coordinated by HIGA, A. R.; doubly supported partly by private initiative, partly by the funding agency FINEP, executed at LAMEF (Laboratory of Forest Genetics and Breeding, Department of Forestry, UFPR).

Authors' Contributions

Juliane Rezende Mercer conceived and directed this study, also performed all experimental work, biometric analysis, and hypothesis statements, and wrote the paper. Milena de Luna Alves Lima, Antonio Rioyei Higa, and Marina Isabel Mateus de Almieda provided advice. All authors read and approved the final version of the paper.

Acknowledgments

The tree samples facility are provided by the private company program. This work was supported by FINEP and Fundação Araucária. The authors gratefully acknowledge the advices of Konstantin V. Krutovsky, Craig S. Echt, Juliana V. M. Bittencourt, Paula Rachel R. Corrêa and Thomas Kirk Bonds (in memorian).

References

[1] Sociedade Brasileira de Silvicultura, *Fatos e Números do Brasil Florestal*, Sociedade Brasileira de Silvicultura, São Paulo, Brazil, 2007.

[2] J. K. Pritchard, X. Wen, and D. Falush, "Documentation for structure software: version 2.3," 2010, http://pritch.bsd.uchicago.edu/structure.html.

[3] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using *multilocus* genotype data," *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[4] G. Evanno, S. Regnaut, and J. Goudet, "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study," *Molecular Ecology*, vol. 14, no. 8, pp. 2611–2620, 2005.

[5] H. Gao, S. Williamson, and C. D. Bustamante, "A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from *multilocus* genotype data," *Genetics*, vol. 176, no. 3, pp. 1635–1651, 2007.

[6] J. J. Doyle and J. L. Doyle, "Isolation of plant DNA from fresh tissue," *Focus*, vol. 13, pp. 13–15, 1990.

[7] P. O. Lewis and D. Zaykin, "Genetic Data Analysis: Computer Program for the Analysis of Allelic Data (version 1.1)," 2001, http://hydrodictyon.eeb.uconn.edu/people/plewis/software.php.

[8] M. Jakobsson and N. A. Rosenberg, "CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure," *Bioinformatics*, vol. 23, no. 14, pp. 1801–1806, 2007.

[9] D. Falush, M. Stephens, and J. K. Pritchard, "Inference of population structure using *multilocus* genotype data: linked *loci* and correlated allele frequencies," *Genetics*, vol. 164, no. 4, pp. 1567–1587, 2003.

[10] L. L. Cavalli-Sforza, P. Menozzi, and A. Piazza, *The History and Geography of Human Genes*, Princeton University Press, Princeton, NJ, USA, 1994.

[11] J. Yu, Z. Zhang, and C. Zhu, "Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping," *The Plant Genome*, vol. 2, no. 1, pp. 63–77, 2009.

[12] M. A. Al-Rabab'ah and C. G. Williams, "Population dynamics of *Pinus taeda* L. based on nuclear microsatellites," *Forest Ecology and Management*, vol. 163, no. 1–3, pp. 263–271, 2002.