OPTIMIZING THE CHOICE OF MICROSATELLITE MARKERS FOR FINGERPRINTING *EUCALYPTUS*

L. Sánchez[1], M.M. Ribeiro[2], C. Ribeiro[3], J.A. Araújo[3], N. Borralho[3] and C.M. Marques[3]

[1] INRA Centre d'Orléans, Unité Amélioration, Génétique et Physiologie Forestière, 45166 Olivet, France; [2] Unidade Departamental de Silvicultura e Recursos Naturais, Escola Superior Agrária, 6001-909 Castelo Branco, Portugal; [3] RAIZ-Direcção de Investigação Florestal, ITQB II, 2781-901 Oeiras, Portugal.(leopoldo.sanchez@orleans.inra.fr)

In this study we have analyzed the information provided by 17 publicly available *Eucalyptus* microsatellite (SSR) markers (Brondani *et al.* 1998, 2002; Jones *et al.* 2002; Steane *et al.* 2001) in the context of genetic identification within a sample of 140 individuals from an elite collection (denoted hereafter *base*) of RAIZ genetic improvement population. This collection involves some groups of individuals of distant origin, and mostly of unknown pedigree. For the study, we developed a software tool based on Monte-Carlo and optimization techniques for the selection of the Minimum Set of fully Informative Markers (MSIM) for genetic identification in this elite collection. This software (*Zeta*) can be used for any genotyped population. Main results are presented on the assessment of the risk of confusion (D) between any two given genotypes, including the event of null alleles among the segregation of the available microsatellite markers.

In order to compare the informativeness of the markers we calculated the discriminant power (D) of each marker, taken as the frequency of pairwise comparisons for which a marker presents distinguishable genotypic patterns over the assemble of comparisons across the sample under study (Table 1). For a given value of confusion (i.e. C = 1 – D) fixed by the user, the newly developed software found the MSIM for the sample of genotypes. This process involved an optimization routine to minimize the number of markers observing a constraint on the global discrimination (i.e. confusion) in the sample under study. Such routine was recursively run: i) for bootstrapped sub-samples within the existing *base*, or the *base* population itself; and ii) for bootstrapped hypothetical offspring (selfs, full-sibs and half-sibs) obtained from recursive recombination and segregation out of the parental genotypes in *base*. By bootstrapping, empirical D's and their corresponding standard errors were obtained for an ample range of combination of markers, and results used in the subsequent risk assessment.

Table 1: Discriminant power of 17 SSR relative to the base population

SSR	D	Effective number of alleles
EMBRA12	0.9897	12.7334
EMCRC8	0.9894	13.2878
EMBRA23	0.9894	12.6306
EMBRA18	0.9864	11.0309
EMBRA11	0.9816	9.4793
EMCRC11	0.9801	8.8242
EMBRA6	0.9754	8.8014
EMCRC10	0.9721	8.3177
EMBRA2	0.9533	5.7777
EMBRA8	0.9526	5.949
EMCRC7	0.9321	4.7886
EMCRC12	0.9267	4.555
EMBRA20	0.9229	4.7337
EMBRA5	0.9206	4.8311
EMCRC5	0.9194	5.627
EMCRC2	0.9085	4.244
EMBRA19	0.8636	3.3986

Considering the 140 individuals in *base*, the MSIM that discriminated between any two of them was composed by 4 highly polymorphic SSR markers (confusion error $\leq 1.07\times10^{-4}$: EMBRA 12, EMCRC 8, EMBRA 23 and EMBRA 18). The same MSIM was subsequently tested in sub-samples of distant origins within *base*, and results showed negligible variations in C between the different origins. The common attributes to these highly informative markers were: i) a very low variation in allele frequencies (directly related to the Effective Number of alleles, see Table 1), and ii) high levels of heterozygosity (H). With confusion errors $\leq 10^{-2}$, the discrimination between hypothetical selfs, full-sibs, half-sibs and unrelated offspring

derived from simulated matings between the *base* parents was obtained with MSIM's of 6, 4, 2 and 2 SSR, respectively.

In the course of this study, the impact of the presence of null alleles on D was also studied by Monte-Carlo simulation. Our analyses revealed important deficits in H from what is expected from allelic frequencies for some of the markers in *base* (average H deficit over loci of 17.2%, range 1.6% ~ 47.6%). We assumed an extreme simulation scenario consisting in each putative homozygote in *base* being actually a carrier of one null allele. These hypothetical *base* genotypes were subsequently used to simulate a very large number (10^6) of new offspring (i.e. no significant change in allele frequencies from parents to offspring). We found in the offspring generation a similar pattern of H deficit across loci to that found in *base*, the average deficit level being higher than in the parental generation (27.7%). Null alleles affected mainly those markers showing the lowest levels of H.

the H deficit found in some markers in *base*. Further research would need an extended pedigree. Naively, the presence of null alleles would be expected to result in a loss of D, given the lack of information provided by a null allele.

Therefore, the presence of a high percentage of null alleles could be one of the main reasons of

Therefore, the presence of a high percentage of null alleles could be one of the main reasons of

However, simulation results showed that the presence of null alleles increased the segregation of distinct genotypes when mating happened between carriers of one null allele, which in turn increased the probability of discrimination between the resulting offspring (i.e. D). MSIM's comprised highly polymorphic markers, and therefore, were unaffected by the presence of null alleles.

REFERENCES

Brondani, R.P.V., Brondani, C., Tarchini, R. and Grattapaglia, D. (1998). *Theor Appl Genet* 97: 816 – 827.

Brondani, R.P.V., Brondani, C., and Grattapaglia, D. (2002). *Mol Genet Genomics* 267: 338-347.

Jones, R.C., Steane, D.A., Potts, B.M. and Vaillancourt, R. (2002). *Can J For Res* 32: 59-66.

Steane, D.A., Vaillancourt, R.E., Russell, J., Powell, W., Marshall, D. and Potts, B.M. (2001). *Silvae Genetica* 50: 89-91.