Estimativa do parentesco numa população de melhoramento de *Eucalyptus globulus* através de microsatélites nucleares

MARIA MARGARIDA RIBEIRO[1*], Leopoldo Sanchéz [2], Nuno Borralho [4], Cristina M. Marques [3]

1 Unidade Departamental de Silvicultura e Recursos Naturais, Escola Superior Agrária, 6001-909 Castelo Branco. Portugal.
2 INRA Centre d'Orléans, Unité Amélioration, Génétique et Physiologie Forestière, 45166 Olivet, France.
3 RAIZ-Direcção de Investigação Florestal, IBET/ITQB II, Av. República, 2780-157 Oeiras, Portugal.
4 BorralhoIDea Urbanização S. Francisco, 18, 2070-220 Cartaxo, Portugal.

*Email: mribeiro@esa.ipcb.pt

Sumário

É situação comum desconhecer-se o grau de parentesco entre a população na origem da maioria dos programas de melhoramento genético de espécies florestais. Para resolver este problema, desenvolvemos um protocolo de avaliação do parentesco utilizando 125 indivíduos e 16 microssatélites, da população base ou de referência (PR) de *Eucalyptus globulus* do RAIZ. Através da recombinação gamética *in silico* foram simulados 10^5 indivíduos com diferentes graus de parentesco: descendentes de autopolinização, meios-irmãos, irmãos completos e indivíduos não aparentados. Por simulação Monte-Carlo foram calculados o valor médio e a variância associada à média dos diferentes grupos de parentesco, com quatro coeficientes de similaridade genética. Compararam-se as funções densidade dos diferentes grupos de parentesco, obtidas com quatro coeficientes de parentesco, utilizando o valor crítico correspondente à intercepção das funções densidade dos indivíduos não aparentados e dos meios-irmãos. O estimador escolhido foi aplicado à PR. Detectaram-se 4,4% de pares de indivíduos potencialmente aparentados, com um erro de tipo II de 8%. Inferimos também, o parentesco de um conjunto de 24 clones elite e encontrámos 4 pares que são potencialmente aparentados. Futuros cruzamentos entre estes indivíduos deverão ser evitados.

Palavras-chave: Microsatelites, *Eucalyptus globulus*, parentesco, melhoramento

Abstract

Founders in tree genetic improvement populations programs usually lack *pedigree* (degree of coancestry) information. To evaluate the genetic similarity between trees in RAIZ base population we genotyped a sample of 125 *Eucalyptus globulus* individuals with 16 microsatellites (SSR) - the reference population (RP). Simulated individuals (10^5) were obtained through gamete recombination, according to different relatedness groups: selfed, half-sib, full-sib and unrelated individuals. The r-values and sampling variances of self, full-sib, half-sib and unrelated individuals were calculated through Monte-Carlo simulations using four pairwise similarity coefficients. The density functions of the relatedness groups were compared by using as threshold the value corresponding to the interception of the probability distribution curves of the unrelated and the half-sib individuals - the critical value. With the selected relatedness estimator 4.4% of individuals putatively related were

detected in the RP, with a type II error of 8%. Additionally, the relatedness among 24 *E. globulus* elite individuals was verified, and four pairs of elite individuals were considered to be putatively related. Future crosses amongst these trees should be avoided.

Key words: Microsatellites, *Eucalyptus globulus*, relatedness, improvement

Introdução

Portugal tem condições únicas, na Europa, para a produção de rolaria de eucalipto destinada a pasta de papel, dadas as condições edafo-climáticas excelentes, especialmente para a espécie *Eucalyptus globulus*. Além disso, Portugal tem vindo a liderar o melhoramento genético desta espécie (Borralho *et al.*, 2007), considerada a melhor para a produção de papel de qualidade. A população de melhoramento genético base deve reflectir a maior diversidade genética possível e deve haver cuidado para que o consecutivo melhoramento da espécie não leve a uma redução excessiva da variabilidade genética, por excesso de consanguinidade entre as árvores seleccionadas. É por isso importante monitorizar o nível de diversidade genética nos diferentes passos do ciclo de melhoramento, com destaque para restrições na realização de cruzamentos controlados entre árvores aparentadas, para isso, o parentesco entre os genitores seleccionados deve ser conhecido ou investigado (Ballou e Lacy, 1995).

Na população portuguesa de *E. globulus*, como aliás na maioria dos casos, não existe informação detalhada sobre o *pedigree* dos seus fundadores ou população genética de base. O *pedigree* das árvores ou o seu respectivo grau de parentesco pode ser determinado através de marcadores moleculares, em particular microsatélites. Estas estimativas são importantes para inferir o grau de parentesco entre material elite de *pedigree* desconhecido e ajudar no desenho de cruzamentos controlados. Diferentes estimadores baseados em marcadores moleculares têm sido propostos para estimar o parentesco na ausência de *pedigree* conhecido e têm sido utilizados em diferentes áreas de investigação (revistos por Blouin, 2003 e Thomas, 2005). Os estudos publicados até agora, geralmente concordam que não existe um coeficiente que seja universalmente superior aos outros e que o seu comportamento depende do grau de parentesco que se pretende estimar, da capacidade informativa do marcador (número de loci e número e frequência de alelos por locus) e da amostra utilizada para estimar as frequências alélicas (Csillery *et al.*, 2006; Van de Casteele *et al.*, 2001; Wang, 2002). Neste estudo foram utilizados quatro coeficientes de parentesco muito comuns: Ritland (1996) (R), Queller e Goodnight (1989) (Q), Lynch e Ritland (1999) (LR) e Li (1993) (L).

Para seleccionar o melhor estimador recorremos a vários critérios, usando simulações Monte-Carlo: 1) maior precisão, ou seja, intervalos de confiança mais pequenos relativamente à média, 2) uma *soma-p* mais pequena, definida como a soma de todos os valores *p* (probabilidade associada ao teste de t) de todas as comparações entre estimativas de grupos de parentesco, partindo da hipótese nula que a média do grupo de parentesco *x* iguala a média do grupo de parentesco *y* ($x \neq y$) e 3) menor área de sobreposição entre cada duas funções densidade de parentesco adjacentes.

Neste estudo foram usados 16 marcadores microsatélites para genotipar 125 indivíduos da população base de melhoramento de *E. globulus*. Estes indivíduos não têm origem (raça nativa) conhecida, tendo sido seleccionados originalmente em plantações exóticas em Portugal, geograficamente distantes e, também, em povoamentos naturais na Austrália. Esta informação será muito relevante para uma boa gestão das populações de melhoramento e especialmente para monitorizar o parentesco do material comercializado. Foram objectivos

deste artigo: i) obter estimativas de parâmetros genéticos para o conjunto de marcadores microsatélites utilizados, incluindo o seu poder discriminante (D), ii) seleccionar o estimador de parentesco com melhor comportamento e iii) aplicar o estimador escolhido à população de referência e a um conjunto de 24 clones elite usados para comercialização pelo RAIZ.

Material e métodos

A população de referência (PR) incluiu 125 árvores sem parentesco conhecido, mas supostamente não relacionadas, representativas da população fundadora do programa de melhoramento do RAIZ (Instituto de Investigação da Floresta e Papel). A PR e um conjunto de 24 clones elite foram genotipados com 16 microsatélites, após a extracção do ADN total de acordo com o método descrito por Marques (1998). Os 16 microsatélites utilizados, foram seleccionados com base no número de alelos e efectivo número de alelos (tabela 1) e foram caracterizados por Brondani (1998), Steane (2001) e Brondani (2002). Considerámos que todos os marcadores utilizados neste estudo tinham segregação independente, porque os testes de desequilíbrio de ligamento não foram significativos, após a correcção de Bonferroni.

Na análise de parentesco, foram utilizados os coeficientes de parentesco de Ritland (1996) (R), de Queller e Goodnight (1989) (Q), de Lynch e Ritland (1999) (LR) e de Li (1993) (L). Usando um programa construído para efectuar a análise dos dados (*Zeta*, que pode ser obtido junto do co-autor LS: Leopoldo.Sanchez@orleans.inra.fr), foram simulados 10^5 indivíduos com diferentes graus de parentesco: descendentes de autopolinização (DA), meios-irmãos (MI), irmãos completos (IC) e indivíduos não aparentados (NR) a partir da PR, através da recombinação gamética *in silico*. Foram calculados os valores médios e a variância associada à média dos diferentes grupos de parentesco com os quatro coeficientes de similaridade genética, utilizando simulações Monte-Carlo e determinados os intervalos de confiança com base no erro de amostragem.

Um dos parâmetro calculados para escolher o estimador com o melhor comportamento, foi a ***soma-p*** de cada coeficiente, que foi definida como a soma de todos os valores *p* (probabilidade associada ao teste de t) de todas as comparações entre estimativas de grupos de parentesco, partindo da hipótese nula que a média do grupo de parentesco *x* iguala a média do grupo de parentesco *y* ($x \neq y$) (figura 1).

Foram também efectuadas comparações não paramétricas baseadas na percentagem de sobreposição das distribuições dos valores médios para cada coeficiente de parentesco, através da integração da distribuição dos 10.000 valores simulados, entre todas as funções densidade, isto é, NR-MI, NR-IC, NR-DA, MI-IC, MI-DA e IC-DA. As **áreas de sobreposição** foram calculadas no *R statistical package* (R Development Core Team 2008).

Compararam-se as funções densidade dos diferentes grupos de parentesco obtidas com quatro coeficientes de parentesco, utilizando o valor crítico correspondente à intercepção das funções densidade dos indivíduos não aparentados e dos meios-irmãos.

Os parâmetros genéticos, número de alelos por locus (N_a), número efectivo de alelos (N_e), heterozigocidade esperada (H_e) e heterozigocidade observada (H_o), foram calculados com o referido programa *Zeta*, assim como o poder discriminante (D) de cada marcador, definido como a probabilidade de um marcador conseguir descriminar entre padrões genéticos diferentes num conjunto de comparações para aquele marcador na PR. A matriz de parentesco calculada com os valores do coeficiente LR para todos os pares de indivíduos dos 24 clones elite foi usada para obter o dendrograma UPGMA (figura 2), usando o programa *NTSYSpc*, versão 2.1 (Rohlf, 1993).

Resultados e discussão

A heterozigocidade média referida na literatura para o *E. globulus*, usando marcadores SSR foi similar ao valor obtido no presente estudo (~0.85; tabela 1). O H_o obtido foi geralmente mais baixo (0.66, Steane *et al* (2001) e 0.62, Jones *et al.* (2002)) do que o valor que nós determinámos (0.73), mas o número de loci utilizados nestes estudos foi sempre muito menor. O facto de estarmos a utilizar uma população artificial pôde, também, aumentar H_o. Efectivamente, numa população de melhoramento desta espécie na Austrália (140 indivíduos) Jones *et al.* (2006) calcularam 0.82 e 0.71, respectivamente, para a heterozigocidade esperada e observada, mas o H_o era menor nas populações nativas (0.66) do que na população artificial que ele estudou. Astorga (2004) detectou valores semelhantes em *E. globulus* usando 26 marcadores SSR com árvores seleccionadas e, ensaios de progénie: H_e=0.80 e H_o=0.70. Noutros estudos com microsatélites em *E. grandis* e *E. urophylla*, a heterozigocidade média observada era muito menor do que a esperada (H_o ~ 0.56-0.62 and H_e ~ 0.86-0.82) (Brondani *et al.*, 2002; Brondani *et al.*, 1998).

Tabela 1 Parâmetros de diversidade para os 16 SSR loci na população de referência, ordenados de acordo com D. Número de alelos por locus (N_a), número efectivo de alelos (N_e), heterozigocidade esperada (H_e), heterozigocidade observada (H_o) e poder descriminante observado (D).

	N_a	N_e	H_e	H_o	D
EMBRA23	21	12.8	0.93	0.89	0.991
EMBRA12	19	13	0.93	0.89	0.991
EMCRC8	18	12.8	0.93	0.84	0.987
EMBRA18	21	11.5	0.92	0.90	0.987
EMCRC11	16	8.9	0.89	0.83	0.981
EMBRA6	15	8.8	0.89	0.78	0.976
EMCRC10	18	8.6	0.89	0.65	0.960
EMBRA11	21	9.4	0.90	0.87	0.960
EMBRA2	15	6.2	0.84	0.76	0.959
EMBRA8	14	6.2	0.84	0.76	0.956
EMCRC7	14	4.8	0.79	0.70	0.932
EMBRA20	13	4.7	0.79	0.62	0.929
EMCRC2	15	4.5	0.78	0.62	0.915
EMBRA5	21	5.2	0.82	0.50	0.898
EMCRC5	21	5.5	0.81	0.53	0.898
EMBRA19	6	3.4	0.71	0.54	0.855
Média	16.8	7.9	0.85	0.73	0.948

Blouin (1996) conclui no seu estudo que 10 loci com H_e = 0.75 poderiam de forma precisa descriminar mais do que 90% dos irmãos completos de indivíduos não relacionados, mas seriam necessários 14 loci para conseguir a mesma descriminação entre irmãos completos e meios-irmãos. Nas circunstâncias do presente estudo as condições estão consideravelmente acima do requerido, pois só um marcador entre os 16 tem H_e <0.75.

As estimativas de parentesco baseadas em marcadores moleculares têm, em geral, um erro de inferência grande (Lynch e Ritland, 1999; Ritland, 1996). De facto, as variâncias obtidas são grandes e intrínsecas ao processo, devido à recombinação quando se formam os gâmetas. Existem factores extrínsecos que afectam os comportamentos absolutos e relativos dos estimadores, como seja a distribuição das frequências alélicas, o número de alelos por locus e a relação real de parentesco. Devido à dependência das diferentes propriedades dos

estimadores (enviesamento e variância) na distribuição das frequências génicas e padrões de parentesco, Van de Casteele (2001) sugere o uso de simulações Monte-Carlo com dados reais, para se poder determinar o coeficiente mais adequado para populações em equilíbrio de Hardy-Weinberg (EHW).

No nosso estudo usámos os genótipos conhecidos da população de referência para estabelecer um "*pool*" genético a partir do qual simulámos gâmetas, através do posicionamento aleatório dos alelos. Estes gâmetas virtuais foram cruzados em seguida de acordo com as diferentes classes de parentesco (NR, MI, IC e DA) e esta sequência de recombinação, segregação e cruzamento, foi repetida de forma a obter as curvas de densidade dos estimadores de parentesco. Com este procedimento não necessitamos de supor que a população está em EHW, porque as condições de recombinação aleatórias irão minimizar o desequilíbrio de ligamento. Este procedimento permitirá inferir resultados em qualquer população de melhoramento, que não estará de certeza em EHW, tal como acontece a nossa.

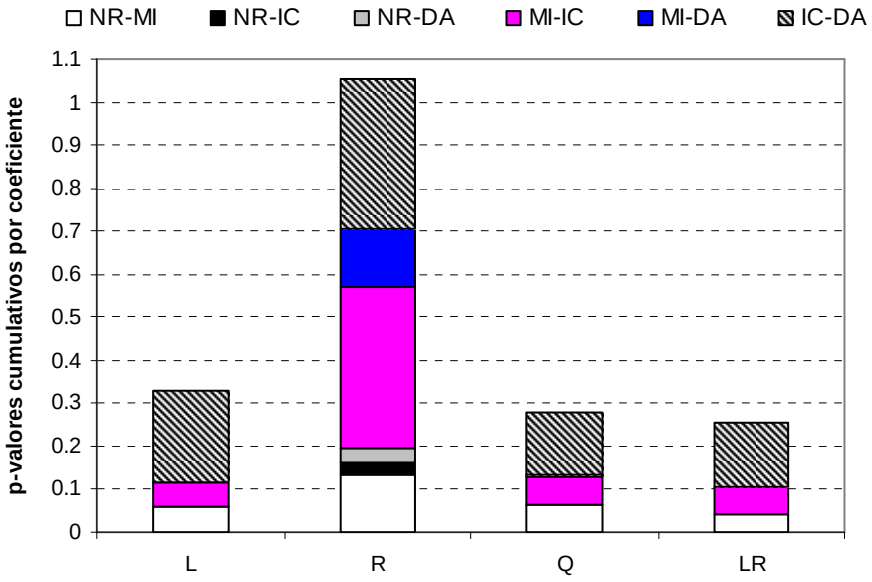

Figura1 Valores cumulativos dos p-valores (*soma-p*) de cada coeficiente de parentesco (significado das abreviaturas na secção de M & M).

O estimador LR parece ser o que melhor comportamento tem para o conjunto de dados, porque para além de preciso e não enviesado, teve a menor percentagem de sobreposição de áreas entre grupos de parentesco (tabela 2), intervalos de confiança mais pequenos (dados não apresentados) e *soma-p* mais pequena (figura 1). Thomas (2005) refere que o estimador de Lynch e Ritland (1999) apresenta propriedades mais interessantes num espectro mais alargado de dados moleculares. Csillery (2006) estudou populações alogâmicas que eram menos relacionados do que meios-irmãos (a maioria dos pares tinha valores de parentesco inferiores a 0,25) e concluíram que o estimador Q tinha menor erro de amostragem para categorias de parentesco mais elevadas e LR comportava-se melhor para relações de parentesco mais baixas. Os mesmos autores referiram que para todas as cinco populações que estudaram, a taxa de erro na classificação de parentesco era menor para o estimador LR. Concluíram, também, que a maior proporção da variância de parentesco real era explicada quando LR era utilizado, reflectindo o facto deste estimador ter uma variância mais pequena quando as categorias de parentesco baixas são mais comuns (NR ou pares pouco relacionado), o que é habitual em populações alogâmicas. No nosso estudo, queríamos

distinguir indivíduos não relacionados de relacionados e precisámos, por isso, de um estimador que fosse mais sensível para indivíduos com baixo grau de parentesco

Os resultados do nosso estudo são confirmados por aqueles apresentados por Ritland e Travis (2004): o estimador LR tem um comportamento superior ao estimador R, no caso de graus mais elevados de parentesco e o último é mais sensível para valores mais baixos de parentesco. Na entanto, o estimador R não pode ser usado, porque as variâncias aumentam com o grau de parentesco esperado e torna-se impossível distinguir indivíduos pertencentes a diferentes grupos de parentesco (dados não apresentados). No estudo de um caso com populações cativas de papagaios, Russello e Amato (2004) concluíram que a medida LR explicava a maior parte da variação na relação de parentesco verdadeira, comparada com o estimador Q. Concluindo, parece que o coeficiente LR tem um melhor comportamento no caso do nosso tipo de população, onde os indivíduos são, geralmente, não relacionados, e o estimador R é o que tem pior comportamento dos quatro estimadores que estudámos.

Tabela 2 Áreas de sobreposição das distribuições de parentesco (%). NR=não relacionados, MI=meios-irmãos, IC=imãos completos e DA= descendentes de autopolinização.

	L	**R**	**Q**	**LR**	**Média**
NR-MI	21.49	15.45	21.87	15.53	18.58
NR-IC	1.32	2.23	1.40	0.67	1.40
NR-DA	0.07	0.31	0.03	0.00	0.10
MI-IC	19.38	44.35	21.78	21.00	26.63
MI-DA	2.48	16.80	1.83	1.56	5.67
IC-DA	38.35	45.50	31.13	30.87	36.46
Média	13.85	20.77	13.01	**11.60**	

Nós usámos o valor do parentesco correspondente à intercepção das distribuições de probabilidade obtidas com as simulações Monte-Carlo, como valor limite – valor crítico. O valor crítico minimiza os erros α e β (o β é a área de sobreposição à esquerda do valor crítico e o α é a área à direita). Dado o nosso interesse em distinguir um par de indivíduos aparentados num certo grau de um par de indivíduos não aparentados, o erro de tipo II torna-se mais importante, isto é, considerar um par não aparentado, quando efectivamente o é. O valor crítico obtido com a intercepção da curva dos indivíduos não relacionados com a curva dos meios-irmãos, calculadas com o estimador LR foi de 0,126 (um valor aliás próximo do valor esperado que é de 0,125), e o valor da área à esquerda foi de 8%. Assim sendo, os pares de indivíduos com valores superiores ao valor crítico podem estar associados a um certo grau de parentesco, pelo menos ao nível dos meios-irmãos (ver figura 2), com uma probabilidade de erro de tipo II de 8%.

O estimador escolhido foi aplicado à PR e detectámos 4,4% de pares de indivíduos potencialmente aparentados, com $\beta = 8\%$. Quatro pares de indivíduos potencialmente relacionados foram encontrados no conjunto dos 24 clones elite com o estimador LR (figura 2). No caso da população amostrada, eles constituem uma pequena proporção (1,4%) de todos os possíveis pares (276) na matriz de parentesco, um resultado promissor do ponto de vista da manutenção da diversidade genética da população de melhoramento.

Apesar da origem destes progenitores de base ser desconhecida, existiriam à partida oportunidades de ocorrerem parentescos insuspeito entre árvores da "raça" Portuguesa de eucalipto (Borralho *et al.*, 2007). Muitas das plantações onde foram seleccionados tiveram origem em plantações de semente colhida em poucas árvores, possivelmente com uma polinização restrita e com um pequeno número de árvores envolvidas..

O facto de a proporção de indivíduos aparentados na população de *E. globulus* do RAIZ ser baixa é uma boa notícia. Porém, este trabalho mostrou que na ausência de informação sobre o *pedigree*, as estimativas obtidas com base no coeficiente de LR a partir de marcadores moleculares podem ser úteis para se quantificar o grau de consanguinidade entre árvores na população de melhoramento.

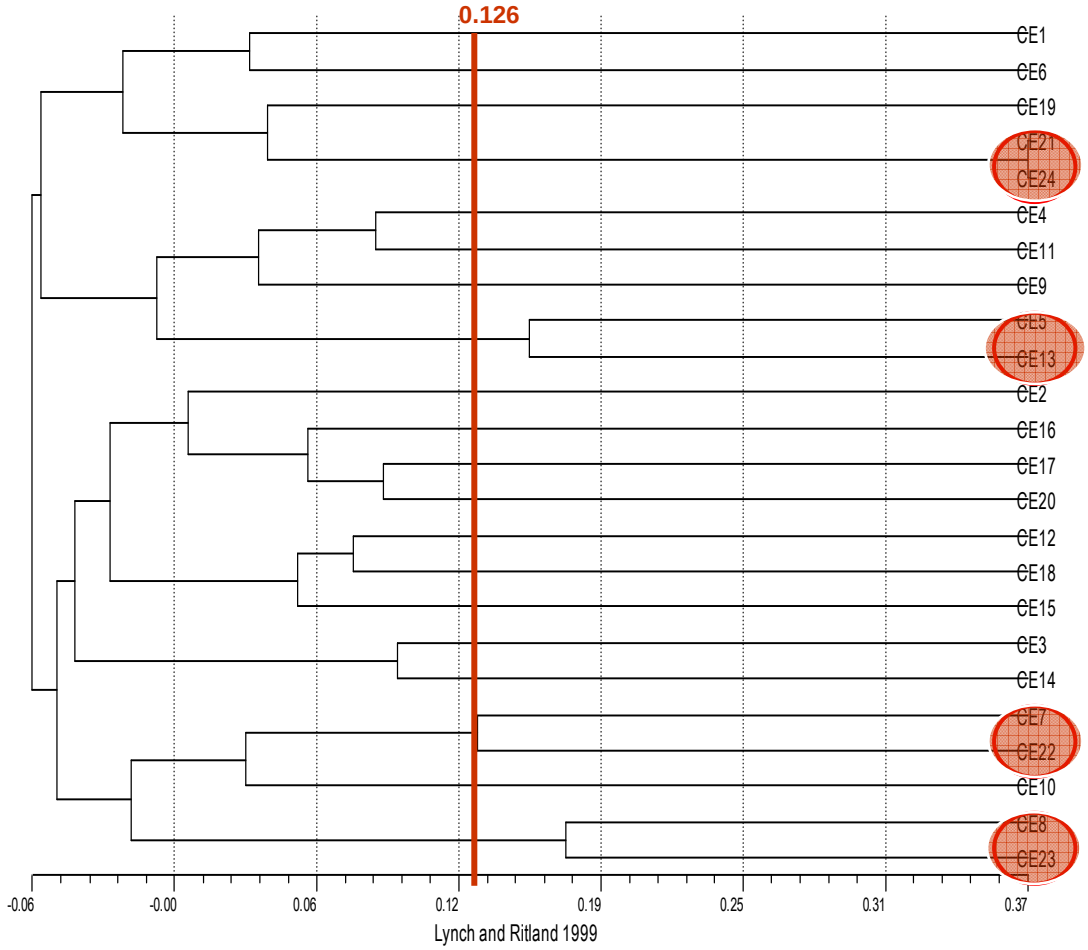

Figura 2 Dendrograma de parentesco dos clones elite (UPGMA) construído com base na matriz de parentesco de Lynch e Ritland (1999).

Agradecimentos

Uma versão mais completa deste artigo foi submetida à revista Trees Genetics & Genomes.

Referências

Astorga, R, Soria, F, Basurco, F, Toval, G (2004) Diversity analysis and genetic structure of *Eucalyptus globulus* Labill. In: Borralho NMG, Pereira JS, Marques C, Coutinho J, Madeira M, Tomé M (Eds.), Eucalyptus in a Changing World. IUFRO. RAIZ, Instituto Investigação da Floresta e Papel, Aveiro, pp. 351-358.

Ballou, JD, Lacy, RC (1995) Identifying genetically important individuals for management of genetic variation in pedigreed populations. In: *Population Management for Survival and Recovery* (ed. Ballou JD), pp. 76-111. Columbia University Press.

Blouin, MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology & Evolution* **18**, 503-511.

Blouin, MS, Parsons, M, Lacaille, V, Lotz, S (1996) Use of microsatellite loci to classify individuals by relatedness. *Molecular Ecology* **5**, 393-401.

Borralho, NMG, Almeida, MH, Potts, BM (2007) O melhoramento do eucalipto em Portugal. In: *O eucaliptal em Portugal* (eds. Alves AM, Pereira JS, Silva JMN), pp. 62-110. ISA Press, Lisbon.

Brondani, R, Brondani, C, Grattapaglia, D (2002) Towards a genus-wide reference linkage map for *Eucalyptus* based exclusively on highly informative microsatellite markers. *Molecular Genetics and Genomics* **267**, 338-347.

Brondani, RPV, Brondani, C, Tarchini, R, Grattapaglia, D (1998) Development, characterization and mapping of microsatellite markers in *Eucalyptus grandis* and *E. urophylla*. *Theoretical and Applied Genetics* **97**, 816-827.

Csillery, K, Johnson, T, Beraldi, D, Clutton-Brock, T, Coltman, D, Hansson, B, Spong, G, Pemberton, JM (2006) Performance of Marker-Based Relatedness Estimators in Natural Populations of Outbred Vertebrates. *Genetics* **173**, 2091-2101.

Jones, RC, Steane, DA, Potts, BM, Vaillancourt, RE (2002) Microsatellite and morphological analysis of *Eucalyptus globulus* populations. *Canadian Journal of Forest Research* **32**.

Jones, TH, Steane, DA, Jones, RC, Pilbeam, D, Vaillancourt, RE, Potts, BM (2006) Effects of domestication on genetic diversity in *Eucalyptus globulus*. *Forest Ecology and Management* **234**, 78-84.

Li, CC, Weeks, DE, Chakravarti, A (1993) Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity* **43**, 45-52.

Lynch, M, Ritland, K (1999) Estimation of pairwise relatedness with molecular markers. *Genetics* **152**, 1753-1766.

Marques, CM, Araujo, JA, Ferreira, JG, Whetten, R, O'Malley, DM, Liu, BH, Sederoff, R (1998) AFLP genetic maps of *Eucalyptus globulus* and *E. tereticornis*. *Theoretical and Applied Genetics* **96**, 727-737.

Queller, DC, Goodnight, KF (1989) Estimating relatedness using genetic markers. *Evolution* **43**, 258-275.

Ritland, K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research* **67**, 175-185.

Ritland, K, Travis, S (2004) Inferences involving individual coefficients of relatedness and inbreeding in natural populations of Abies. *Forest Ecology and Management* **197**, 171-180.

Rohlf, FJ (1993) *NTSYSpc: Numerical Taxonomy and Multivariate Analysis System (V 2.1.)* Exeter Publisher, New York.

Russello, MA, Amato, G (2004) Ex situ population management in the absence of pedigree information. *Molecular Ecology* **13**, 2829-2840.

Steane, DA, Vaillancourt, RE, Russel, J, Powell, W, Marshall, D, Potts, BM (2001) Development of microsatellite loci in *Eucalyptus globulus* (Myrtaceae). *Silvae Genetica* **50**, 89 - 91.

Thomas, SC (2005) The estimation of genetic relationships using molecular markers and their efficiency in estimating heritability in natural populations. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1457-1467.

Van de Casteele, T, Galbusera, P, Matthysen, E (2001) A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology* **10**, 1539–1549.

Wang, J (2002) An estimator for pairwise relatedness using molecular markers. *Genetics* **160**, 1203-1215.