DEVELOPMENT OF MULTITAAG: AN AUTOMATED GENOTYPING SYSTEM FOR *EUCALYPTUS* SPECIES USING MULTIPLEX AMPLIFICATION OF MICROSATELLITE MARKERS

Gaëlle Lehouque[*,1], Rebeca Sanhueza[2], Francisco Melo[1]

[1] TAAG Genetics S.A., Antonio Bellet 77, oficina 703. Providencia. Santiago. Chile
[2] Forestal Mininco S.A., Avda. Alemania 751. Casilla 399. Los Angeles. Chile

Autores para la correspondencia: gaelle@taag-genetics.com

Boletín del CIDEU 6-7: 25-34 (2008)
ISSN 1885-5237

Summary

We developed a high-throughput genotyping system for *Eucalyptus globulus* Labill., *E. nitens* and hybrids based on the simultaneous amplification of nine highly polymorphic microsatellite markers in a multiplex polymerase chain reaction (PCR) using fluorescently dyed primers, followed by the automated fragment detection on a capillary electrophoresis equipment. Additionally, we developed a computer software that performs the allele calling directly from the output electropherogram files, allowing the automated analysis of the collected data. This analysis system, named MultiTAAG, allows reliable genotyping with the advantages of assaying a large number of markers in parallel, making it fast and cost-effective. Combined with the power of the computer software for automated allele assignment, MultiTAAG is less prone to error and easily applicable at a large-scale.

Keywords: Fingerprint, genotyping, *globulus*, *nitens*, hybrid

Resumen

Desarrollo de MultiTAAG: un Sistema de Tipificación Automatizado para Especies de *Eucalyptus* Basado en Amplificación de Múltiples Marcadores Microsatélites.

Se describe el desarrollo de un sistema de tipificación de alto rendimiento para *Eucalyptus globulus*, *E. nitens* e híbridos basado en la amplificación simultánea de nueve marcadores microsatélites altamente polimórficos. La tipificación se lleva a cabo en una reacción en cadena de la polimerasa usando partidores fluorescentes, seguida por la detección automática de los fragmentos en un equipo de electroforesis capilar. Además, se desarrolló un programa computacional capaz de reconocer los alelos directamente a partir de los electroferogramas de salida, permitiendo así un análisis automático de los datos generados. Este sistema de análisis, llamado MultiTAAG, permite realizar una tipificación confiable con las ventajas de ensayar múltiples marcadores en paralelo, haciéndolo rápido y costo-efectivo. Combinado con el poder del programa computacional para la asignación automática de alelos, MultiTAAG presenta menos errores y es fácilmente aplicable en gran escala.

Palabras clave: Huella genética, genotipeo, *globulus*, *nitens*, híbridos

I. Introduction

Since most of the Eucalypt plantations for production rely on vegetative propagation, it is essential for forestry companies to count with accurate identification methods such as those based on DNA genotyping. Apart from its use in clonal propagation quality control, genotyping is also indispensable in genetic improvement programs for germplasm identification, control of external pollen contamination, parentage analysis and marker assisted selection, as well as in population genetic studies to estimate population variability, maintain diversity or infer relationship between populations.

Microsatellite markers are powerful tools commonly used for genotyping in forestry. Nevertheless, their use is still expensive and time-consuming because it relies on laborious gel-based techniques or because it generates large data files of difficult interpretation when based on fluorescent detection.

Here we describe the development of a high-throughput genotyping system, *Eucalyptus globulus* MultiTAAG, which overcomes the inconveniences mentioned above. The system is based on the simultaneous amplification of nine highly polymorphic microsatellite markers in a multiplex PCR using fluorescently dyed primers, followed by the automated detection of multiple fragments on a capillary sequencer. Finally, we developed a computer software that performs the allele calling directly from the output electropherogram files, allowing a fully automated analysis of the collected data.

Microsatellites were selected based on their transferability between *Eucalyptus* species. Their polymorphism was tested in eleven different *E. globulus* races and provenances: they display an average of 12 alleles per *locus* in 55 samples.

For the combination of the nine *loci*, the exclusion power in a sample from Forestal Mininco *Eucalyptus* breeding population (a forestry company in Chile), was superior to 99,85%. Preliminary studies show that the selected multiplex primers also amplify in other *Eucalyptus* species and their hybrids.

This system allows reliable genotyping with the advantage of assaying a large number of markers in parallel, thus making it fast and cost-effective. Combined with the power of the computer software for automated allele assignment, our system is less prone to error and easily applicable at a large-scale.

II. Materials and methods

Plant Material and DNA Extraction

Simple sequence repeat (SSR) initial polymorphism screening was performed on material from eleven different *Eucalyptus globulus* provenances and seed sources: Salinas (Colombia), Otways, Moogara, Flinders Island, Mount Dromedary, Jeeralangs, Seymour, Bruny Island and Wilson Promontory (Australia); Huelva (España) and Colcura (Chile). Five individuals from each provenance were randomly chosen from a germplasm collection of plantlets grown from seeds.

To validate the discriminatory power of the resulting genetic fingerprint, 40 samples of *Eucalyptus globulus* from the breeding population of Forestal Mininco, including clones of the same individuals, were blind-genotyped.

To test cross-amplification in other species, 16 samples of *Eucalyptus nitens* x *Eucalyptus globulus* hybrids and 8 samples of *Eucalyptus nitens* were analyzed.

Genomic DNA extraction from fresh and dry leaves was carried out according to Stange *et al.* (1998) with minor modifications. Briefly, approximately 50 mg of leaves were ground in the presence of pre-warmed CTAB extraction buffer.

This buffer contains NaCl to eliminate polysaccharides, as well as PVP40 and antioxidants to remove polyphenols. The homogenates were incubated at 65ºC for 30 minutes. After centrifugation at 13.000 rpm for 10 minutes, samples were extracted twice with chloroform. Aqueous phases were precipited with cold isopropanol for one hour at -20ºC. Samples were centrifuged at 13.000 rpm for 10 minutes and pellets were washed twice with 70% cold ethanol. After air drying, pellets were resuspended in 30 to 70 ul of nuclease-free water. DNA concentration was measured with the Qubit system using Quant-iT kit (Invitrogen, Carlsbad, CA, USA).

Polymorphism screening and estimation of minimum and maximum alleles

Eleven *loci* were selected based on their transferability between *Eucalyptus* species and on the length of the sequence available in the NCBI database: Emcrc1, Emcrc2, Emcrc3, Emcrc5, Emcrc6, Emcrc8, Emcrc9, Emcrc10, Emcrc11, Emcrc12 (Bundock *et al.*, 2000; Steane *et al.*, 2001) and Embra63 (Brondani *et al.*, 2002).

To validate the polymorphic nature of each SSR on our *Eucalyptus* samples, primer pairs were designed from the flanking regions of each *locus* using our own software and monoplex PCRs were performed. Minimum and maximum allele sizes and repeats for each SSR were aproximately estimated by comparison to a 100 bp ladder after separation of PCR products on denaturing polyacrylamide gels.

MultiTAAG primers design

Primer pairs capable of amplifying in a multiplex reaction all selected *loci* were designed using a computer cluster of 16 Pentium IV processors running LINUX operating system and based on a proprietary software written in ANSI C language. The computer software receives a list of template DNA sequences in FASTA format as the input, along with a series of constraints that must be fulfilled to produce a successful set of oligonucleotide primers for multiplex PCR. The best set of primers is produced by a genetic-algorithm-based optimization. The following primer constraints were considered: length range (18- to 30-mers), melting temperature range (55 to 65°C), energy of intramolecular interactions (4 kcal/mol), total number of contiguous intermolecular interactions (6 bp), total number of contiguous intermolecular interactions at the 3’ end (5 bp), total number of noncontiguous intermolecular interactions (9 bp), maximum melting temperature difference (3°C), and specificity or maximum number of contiguous 3’ interactions with other regions in the template sequences (8 bp).

One primer of each pair was labeled at the 5’ end with one of the following fluorochromes: 6-FAM, HEX or NED. As a maximum of three different channels are available for the detection of the labeled primers, the *loci* selected for multiplex amplification were distributed into three groups according to the allelic size range variation. An additional constraint was that the primers for the *loci* included in the same group must amplify products with non-overlapping size ranges.

The fluorochrome-labeled primers were ordered from Sigma Genosys (Woodlands, TX, USA) and Genesys S.A. (Santiago, Chile).

MultiTAAG amplification, automated reading and allele calling

MultiTAAG PCR amplifications were carried out in an Applied Biosystems thermal cycler, model 2720 (Foster City, CA, USA). Between 15 and 20 ng of genomic DNA were amplified in a total volume of 15 ul of nuclease free water with 1 unit of Platinum® Taq polymerase

(Invitrogen, Carlsbad, CA, USA), 2,4 mM MgCl2, 0,35 uM each dNTP, 20 mM Tris-HCl (pH8) and 0.1-0.3 uM of each primer.

The amplification conditions consisted of an initial denaturation step (95°C for 3 min), and 35 cycles of 94ºC for 45 sec, 56ºC for 45 sec and 72ºC for 55 sec, followed by a final extension at 72ºC for 10 minutes.To check PCR yields, amplified samples were run on 2% agarose gels stained with ethidium bromide prior to fluorescent reading on an ABI PRISM® 3100 Genetic Analyzer (Applied Biosystems) using matrix D suitable for 6-FAM, HEX, NED and ROX dyes and GeneScanTM 500 ROXTM size standard.

We developed a computer software that performs the allele calling directly from the output electropherogram files of the ABI PRISM®, thus allowing the automated analysis of the collected data. This computer software automatically produces the genetic fingerprints, which are constituted by a pair of alleles for each SSR measured.

Genetic Distances Among Individuals

In order to determine the genetic distance between two individuals, the genetic fingerprints were compared in a pairwise analysis (all-against-all). Considering the heterozygous nature of the genotype data, the shared allele distance estimator was used (Jin and Chakraborty, 1993: Bowcock *et al.*, 1994), which is defined as D_{sa}:

$$D_{sa}^{i,j} = 1 - \left(\frac{\sum_{l=1}^{L} S_l^{i,j}}{2L} \right)$$

where $S_l^{i,j}$ corresponds to the number of shared alleles between individuals i and j at *locus* l; and L corresponds to the total number of *loci* that were compared between the two individuals. The resulting shared allele distance table that contained all parwise comparisons among individuals was then used with the nearest neighbour clustering method and a dendrogram built to graphically represent and analyze the results.

III. Results

Microsatellite loci polymorphism in Eucalyptus globulus

To determine the polymorphic level of each marker, as well as to estimate the minimum and maximum number of repeats at each *locus*, PCR amplification of the eleven selected SSRs was performed across a broad panel of *Eucalyptus globulus* races and provenances.

All eleven primer pairs allowed the specific amplification of products. All *loci* were found to be polymorphic in the 5 individuals from each race and provenance. They displayed between 9 and 14 alleles (Table 1), with an average of 11-12 distinct alleles per *locus* in the 55 samples. Depending on the microsatellite considered, between 38 and 73% of the individuals was heterozygous (data not shown).

Based on these results, the eleven *loci* were considered for the further development of an informative genetic fingerprint for *Eucalyptus*.

Multiplex design

New oligonucleotide primer pairs were then designed for the simultaneous amplification of the eleven selected *loci* by a computer based-optimization protocol (see Materials and Methods). As a result of the calculations, it was only possible to include 9 out of the 11 markers in the multiplex

Table 1: Characteristics of the microsatellite markers tested for the development of the MultiTAAG system

Locus	Accession number in GenBank	Repeat motif	Minimun number of repeats	Maximum number of repeats	Number of observed alleles
Emcrc 1	AJ401136	CA	1	26	12
Emcrc 2	AJ401137	CA	8	33	14
Emcrc 3	AJ401138	TG	2	29	9
Emcrc 5	AJ401140	(CT) (CA)	14	43	11
Emcrc 6	AJ401141	(TC) (AC)	25	57	14
Emcrc 8	AJ401143	(CT) (CA)	17	48	11
Emcrc 9	AJ401144	TG	6	29	11
Emcrc 10	AJ401145	(GT) (GA)	3	36	12
Emcrc 11	AJ401146	(TC) (AC)	19	40	13
Emcrc 12	AJ401147	(CT) (CA)	11	35	12
Embra63	G74884	TC	1	25	11

reaction. The final selected *loci* to be included in the genetic fingerprint were Emcrc1, Emcrc2, Emcrc3, Emcrc5, Emcrc8, Emcrc9, Emcrc10, Emcrc11 and Emcrc12. The expected amplification size range for each SSR system in this multiplex PCR reaction is shown in Figure 1.

A computer software for the automated analysis and allele calling from the results of the multiplex PCR reaction and the automated production of a genetic fingerprint was also developed. The molecular biology kit along with the subsequent integrated computational analysis to produce the final genetic fingerprint was called the MultiTAAG system.

Validation of the MultiTAAG system

The MultiTAAG system was tested on a set of 40 *Eucalyptus globulus* samples from a Chilean breeding population. For one of the samples (M37-G1), though many efforts to extract good-quality DNA were attempted, no successful amplification was achieved. Of the 39 remaining samples, more than 92.3% displayed a high-quality fingerprint, meaning that at least 16 alleles could be read with a good intensity from at least 8 independent *loci*. Less than 7.7% of the samples showed a medium-quality fingerprint, meaning that either between 12 and 15 alleles could be read, and/or that the intensity of the fluorescent signal was weak.

The obtained fingerprints were compared all-against-all, their shared allele distance calculated and clustered. The resulting dendrogram is shown in Figure 2. A total of 21 unique genotypes were identified, most of which are not closely related since they share less than 50% of their alleles among them.

The results blindly obtained with the MultiTAAG system were then compared with the genotype identification of the samples provided by Mininco company.

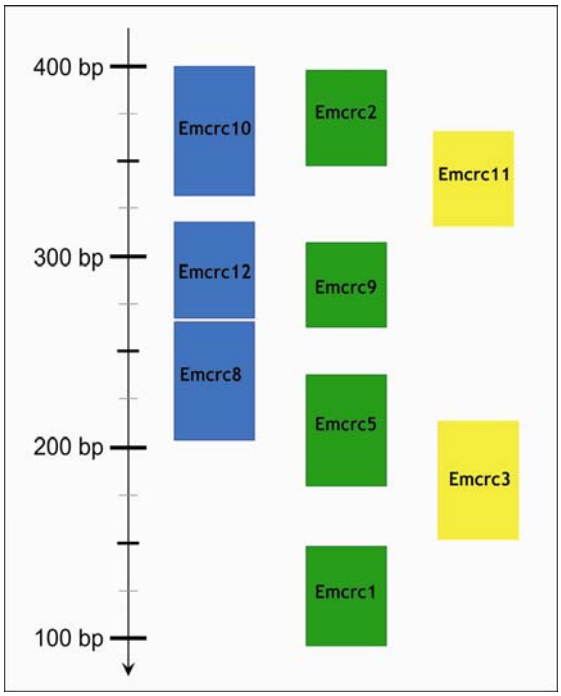

Figure 1: Schematic representation of the MultiTAAG fingerprint. The possible size range of each microsatellite marker in base pairs (bp) is represented by a box. In blue are the *loci* amplified with 6-FAM-labeled primers, in green the ones amplified with HEX-labeled primers, and in yellow the ones amplified with NED-labeled primers.

Four samples (M1-G1, M10-G1, M34-G1 and M36-G1) were found to be genetically different to the clonal group to which they should belong according to their greenhouse identification. However, sample M34-G1 was found to be genetically identical to another genotype, that includes samples M6-G1, M16-G1, M31-G1.

Additionally, sample M28-G1, that was included in the same clonal group of M7-G1, M29-G1, M33-G1, and M38-G1, showed a difference in only one allele at *locus* Emcrc3 with these samples (10/21 against 10/11).

Similarly, samples M15-G1, M27-G1, M30-G1, M32-G1, M33-G1 and M35-G1 were found genetically identical with one another, as according to their identification, but different in one allele at *locus* Emcrc9 (6/11) with samples M6-G1, M16-G1 and M31-G1 of the same genotype (11/11).

The discriminatory power of the genetic fingerprint provided by the MultiTAAG system was calculated from the set of 21 unique and unrelated genotypes (Table 2).

Most of the *loci* were found to be highly polymorphic, with an observed homozygosity minor than 0.40 and a minimum number of 7 observed alleles in this small sample. Based on this set of individuals, the average exclusion power of the complete genetic fingerprint containing the nine *loci* was found to be superior than 99.85%.

Cross-amplification in Eucalyptus species

Preliminary studies show that the selected multiplex primers also amplified in other *Eucalyptus* species and their hybrids. Out of 16 samples of a hybrid between *E.nitens* x *E.globulus*, 15 exhibited a high quality fingerprint (Figure 3) and only one displayed a medium quality fingerprint. In four cases, no allele could be read at *locus* Emcrc5, and in one case no allele could be read at *locus* Emcrc2.

In *E.nitens* samples, between 5 and 7 markers could be amplified, also generally excluding *loci* Emcrc5 and Emcrc2, similarly as it was observed for the hybrids.

IV. Discussion

In this work, aided by an optimal primer design scheme and the careful set-up of relative primers concentrations and amplification conditions, we were able to multiplex the amplification of a total of nine microsatellite *loci* in a single reaction. The high discrimination power of the

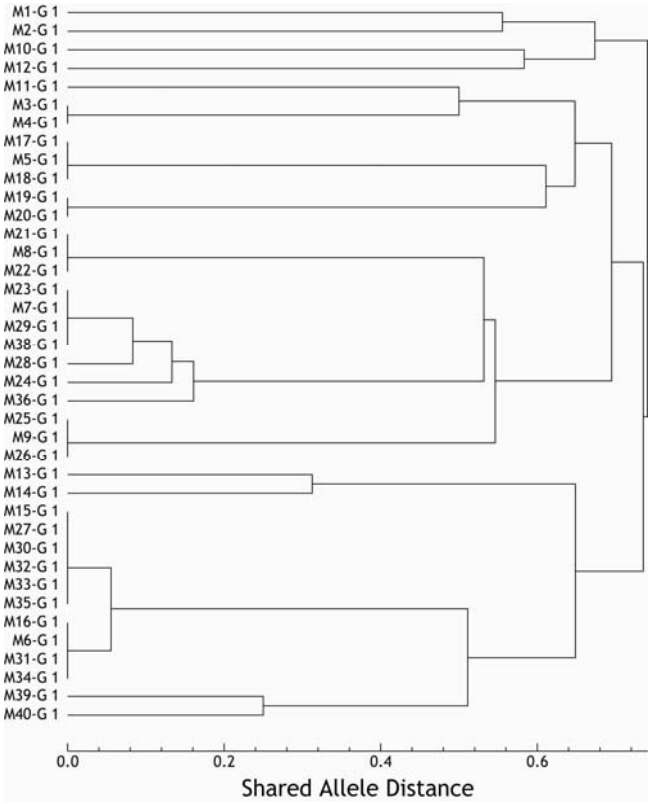

Figure 2: Genetic distance among samples. Dendrogram resulting from microsatellite-based genetic distance of nine SSR of *Eucalyptus globulus* in 40 samples from a Chilean breeding population (M1-G1 to M40-G1), estimated by shared allele distance in pairwise comparisons. The dendrogram was produced using nearest neighbour clustering method. The scale represents the proportion of alleles differing between two samples.

Table 2: Variability of the nine microsatellite *loci* included in the MultiTAAG fingerprint

Locus	Number of possible alleles	Number of observed alleles	Number of homozygous individuals	Number of individuals considered	Observed homozigosity	Average exclusion power
Emcrc1	27	9	6	21	0.2857	0.450705
Emcrc2	26	7	13	20	0.6500	0.086271
Emcrc3	28	8	2	20	0.1000	0.795420
Emcrc5	30	7	11	21	0.5238	0.167503
Emcrc8	32	10	1	20	0.0500	0.898213
Emcrc9	25	8	11	18	0.6111	0.107306
Emcrc10	35	8	5	19	0.2632	0.487527
Emcrc11	22	9	8	20	0.4000	0.290880
Emcrc12	26	7	5	19	0.2632	0.487527

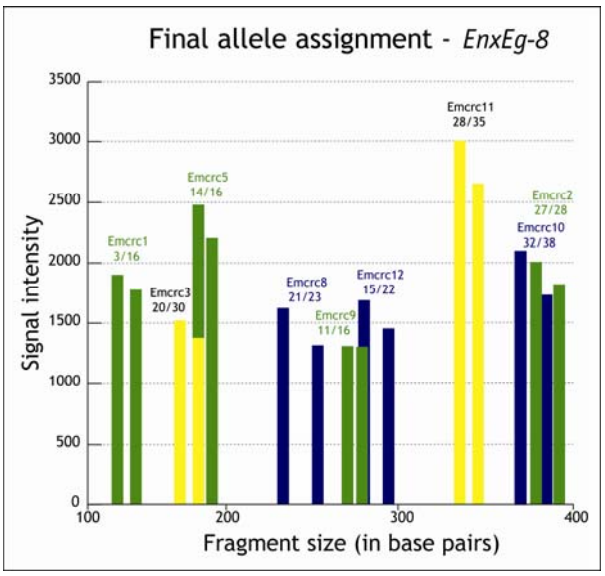

Figure 3: MultiTAAG fingerprint of an *E.nitens* x *E.globulus* sample. The MultiTAAG software graphical output displays the fragment size against the intensity of the fluorescent signal. In blue are represented the PCR products of the *loci* amplified with 6-FAM-labeled primers, in green those amplified with HEX-labeled primers, and in yellow those amplified with NED-labeled primers.

MultiTAAG fingerprint makes it suitable for clone identification as well as parentage-studies.

The results of the validation step demonstrate that the MultiTAAG system is reliable for *E.globulus* genotyping, as the clustering based on the fingerprints of the blind genotyped samples was confirmed by comparison with their genotype identification. In only four cases out of 40, the fingerprint of the sample obtained with the MultiTAAG system excluded the sample from its clonal group. This is most probably due to an error of identification of the clone in the greenhouse, as the repetition of the complete process (from extraction to allele calling) gave the same results.

In two cases, samples show a difference for only one allele at one *locus* with the other clones from the same genotype. These differences are likely to be due to experimental artefacts such as Taq polymerase slippage in the first steps of the PCR, or preferential amplification of only one allele in an heterozygous sample, due to the low quality of the DNA (Pompanon *et al.*, 2005). According to the International Society of Forensic Genetics, genetic difference is declared when at least two independent *loci* differ in at least two base pairs. Therefore, the samples exhibiting only one difference in their fingerprint should not be considered as genetically different.

To the best of our knowledge, the multiplexing of more than four primer pairs has not been described before for *Eucalyptus*. This high multiplexing level delivers two main advantages. While reducing the labor required for genotyping per sample, it also makes it more cost-effective as it allows the analysis of 9 *loci* in one single reaction instead of a minimum of three reactions. When used for high-throughput analysis, the cost of the fluorochrome-labeled primers is counter-balanced by the reduction in cost of the other reagents like the Taq DNA polymerase. Combined with the automated sequencer-based reading, the MultiTAAG system dramatically decreases the time of analysis compared with the time- and labor-consuming gel-based techniques. Moreover, it is more resolutive and more accurate than gel-scoring.

Furthermore, the automated computer software based allele calling should considerably lower the error rate, as most of the genotyping errors are known to be caused by human factors (Hoffman and Amos, 2005) and in particular at the step of reading the profiles.

Acknowledgments

This work was supported by Programa Bicentenario de Ciencia y Tecnología IPC50.

We would also like to thank Forestal Mininco for providing the wide collection of *E. globulus* provenances, hybrids samples and other species.

References

Steane, D.A.; Vaillancourt, R.E. ; Russell, J. ; Powell, W. ; Marshall, D.; Potts, B. M. (2001). Development and characterisation of microsatellite *loci* in *Eucalyptus globulus* (Myrtaceae). *Silvae Genetica*, 50; 89-91.

Stange, C.; Prehn, D.; Arce-Johnson, P. (1998). Isolation of Pinus radiate genomic DNA suitable for RAPD analysis. *Plant Molecular Biology Reporter*, 16: 1-8.

Brondani, R.P.V.; Brondani, C.; Grattapaglia, D. (2002). Towards a genus-wide reference linkage map for *Eucalyptus* based exclusively on highly informative microsatellite markers. *Mol. Gen. Genomics* 267: 338-347.

Bundock, P. C.; Hayden M.; Vaillancourt R. E. (2000). Linkage maps of *Eucalyptus globulus* using RAPD and microsatellite markers. *Silvae Genetica,* 49: 223-232.

Bowcock, A. M.; Ruiz-Linares, A.; Tomfohrde, J.; Minc, E.; Kidd, J. R.; Cavalli-Sforza L.L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368: 455-457.

Jin, L.; Chakraborty, R. (1994). Estimation of genetic distance and coefficient of gene diversity from single-probe multi*locus* DNA fingerprinting data. *Mol Biol Evol*, 11: 120-127.

Hoffman, J.I.; Amos, W. (2005). Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Mol. Ecol.* 14(2): 599-612.

Pompanon, F.; Bonin, A.; Bellemain, E.; Taberlet, P. (2005). Genotyping errors: causes, consequences and solutions. Nature reviews, 6: 847-859