ID Checking by Microsatellite Type Markers (SSR) During the oil Palm Variety Selection and Production Processes

T. Durand-Gasselin (3), N. Billotte (2), V. Pomies (1), G. Mastin (1), F. Potier (1), P. Amblard (3), A. Flori (1) and B. Cochard (1).1

INTRODUCTION

Cirad and PalmElit finger-printing project was launched in 2007 as part of the quality approach applied to our selection and seed production processes. The purpose of this communication is to provide a rapid overview of the results obtained on the method.

This work take advantage of different techniques and know how developed by Cirad team. The molecular markers used are microsatellite type markers (SSR) SRS developed by N. Billotte when he developed the first genetic map of oil Palm (Billotte et al., 2002). It also takes advantage of studies on genetic diversity conducted by B. Cochard (Cochard, 2009).

As such a technique will address huge number of genotypes, technical choices as well as statistical methods were helpful in order to stringent the efficiency of the proposed process.

We will give an overview of the different aspect of the whole process and give some examples to illustrate our presentation.

CHOICE OF TYPE OF MARKERS

For the time being, there has been *no* hesitation on the type of markers to be used, it appeared obvious that microsatellites had all the necessary qualities: already developed by N. Billotte, polymorphic enough, co-dominant, easily readable, portable and reasonably priced.

MAIN LINES OF DNA PREPARATION, PCR AND MIGRATION PROTOCOL

Leaf sample are collected on young leaves, freeze-dried and grinded. The sample can then be stored at -20°C. For DNA extraction, NucleoSpin extraction kit, Macherey-Nagel, was used and in most cases, 1 column (20 mg of ground material) per sample with the NucleoSpin 96 Plant II kit is enough for a legitimacy test. An automated extraction protocol can be developed. DNA quantification is done on the Fluoroscan with Hoechst (BisBenzimide, CMR) in accordance with the quantification workshop protocol and concentration is adjusted at 25 ng/µl, then at 5 ng/µl in working solutions.

Initial work was conducted using simplex microsatellite PCR, and then a multiplex was constituted based on the panel of 12 SSR proposed for the routine work.

---

[1] Cirad, UPR 28, genetic of oil Palm, Avenue agropolis, 34 000 Montpellier.
[2] Cirad, UMR DAP, Avenue agropolis, 34 000 Montpellier.
[3] PalmElit SAS, Bat 14, Parc Agropolis, 2214 Bd de la Lironde, 34980 Montferrier sur Lez.

PCR products were diluted for migration on Li-Cor ®.

CHOICE OF SSR MARKERS

More than 400 microsatellites was initially available on the oil palm: we had to choose the most reliable combination of markers. We had as a practical objective to select 12 or 16 SSR markers. The choice of SSR markers to be used was made in three stages:

1. Initial practical sorting. A first choice of 29 microsatellites was made from among the most polymorphic and the best distributed on the genome. As the project proceeded, those which regularly raised problems (poor amplification, difficult to read, etc.) were eliminated or replaced. In the end, out of 29 markers tested in all, 19 microsatellites were adopted and "used on" 421 genotypes.

2. Then, "statistical" sorting intended to find a panel of 12 markers easy to use on a routine basis was carried out. It was proposed by Albert Flori, who used a method making it possible to seek among the (approximately) 50,000 combinations of 12 markers and find those which made a result as discriminatory as the initial combination of 19. There were around 1,300. Seven markers are present in more than 1,000 of these combinations and therefore play a decisive role, whereas three markers are less frequent (fewer than 500 cases). By insisting on keeping the former and excluding the latter, there barely remain more than 100 combinations. The final sorting operation sought the most robust among them by monitoring what happens if one of the 12 markers is not amplified. This approach led to a unique combination of 12 markers being proposed.

3. Lastly, this combination was tested in practice: migration time, ease of multiplexing.

MEASUREMENT PRECISION AND PROBABILITIES

For each genotype, this work meant calculating the probability of belonging to the cross hoped for (the one expected).

The initial work for the legitimacy checking project allowed for the characterization of 421 individual using 19 molecular markers. In order to optimize laboratory work, it seemed necessary to examine whether checking could be carried out as reliably with fewer markers.

By reducing the number of markers, the main risk taken is to no longer be in a position to identify illegitimate individuals. The opposite risk of wrongly declaring an individual to be illegitimate can only arise from reading or retranscription errors. It is in theory a much lower risk and must be negligible as an initial approximation.

In order to identify illegitimates correctly, the number of polymorphic markers studied must be sufficient for the possible allele combinations in each cross to be clearly specific to it. Thus, any coincidence between the genotype of the individual and one of the allele combinations of the cross cannot be down to chance.

The method proposed to determine whether a given set of markers is enough to detect illegitimates involves the following stages:

1) Calculation of the probability of finding the genotype of a given individual assuming its two parents to be known.
2) Determination of a probability limit below which it is considered that the individual cannot come from the assumed cross.
3) Evaluation of the proportion of truly illegitimate individuals that are not detected by applying the test in stage 2).

The proportion of undetected illegitimates makes it possible to compare the different sets of markers to ultimately choose the most efficient one.

Calculating the probability of observing the genotype of an individual

Depending on whether the parents are heterozygous or homozygous and depending on the number of alleles they have in common, it is quite easy in theory to determine what the possible genotypes are at a given locus and therefore what the theoretical probability is that an individual received the observed genotype.

However, it is seen in cases where an allele observed in an individual at a given locus does not correspond to any of the parental alleles, that it is more likely to see a small difference in the number of base pairs between the allele of the individual and the alleles of the parents than a large difference (Figure 1). This suggests that determining the size of DNA fragments by the electrophoresis technique is subject to slight measurement uncertainty.

Figure 1: Frequency of differences between observed allele and parental alleles

For all the analyses carried out up to now (385 phenotype (palms) for which the parents were also analysed or reconstituted studied on 19 loci, i.e. 7,315 readings), virtually perfect agreement is found (to within 2 base pairs) in 6,867 cases between the genotype of the progeny and that of its assumed parents. In 143 cases, i.e. 2% of cases, the agreement is not absolutely perfect. It can therefore be said, as a first approximation, that some reading errors of 2 base pairs (or more) occur in 2% of cases.

If it is taken that reading uncertainty follows a normal null average law, it is possible to calculate as a function of its standard deviation the theoretical probability of the measurement error being at least 2 base pairs. In figure 2, it can be seen that 2% error rates occur for a standard deviation of 1. (Figure 2)

Given the measurement uncertainties, the observed genotypes may differ slightly from the true genotypes. What procedures can be used to calculate the probability of observing the genotype G_{obs} for an individual for which the true genotype is assumed to be G_{true}. This probability depends on the distance between G_{obs} and G_{true} and is easily calculated. Where a given cross and locus are involved, the true genotype of all the members of the family is one of the four genotypes (possibly combined) that are theoretically possible in that cross. Those genotypes are equally likely. Consequently, the probability of observing G_{obs} for an individual assumed to belong to some cross or other is calculated as the mean of the four probabilities of observing G_{obs} assuming that the true genotype is each of the 4 possible genotypes in the cross.

For a combination of loci, the probability of the individual's genotype is the product of the probabilities for each of the loci.

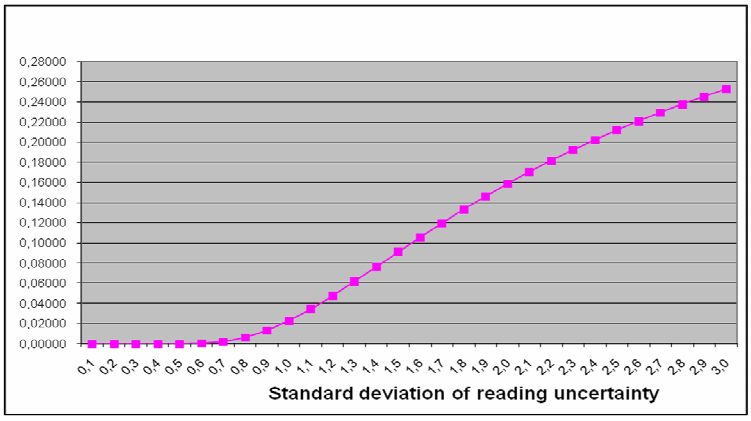

Figure 2: Probability of 2 base pairs error

Apart from the measurement uncertainties, it is accepted that in the case of one reading out of 10,000, a result handling or retranscription error makes it such that the probability calculated at a given locus for an individual is totally wrong. In particular, even if the probability of observing the genotype is nil for the proposed cross, it is considered that recording errors may have led to the observed result. The probability value adopted in the product is therefore slightly modified to take into account this flexibility: for an illegitimate individual, the probability calculated in this way is close to 0. For a legitimate individual it is quite far from 0.

Determination of a probability limit for legitimacy/illegitimacy

When examining the distribution of probabilities for legitimate individuals, it is found that it is reasonable to consider as illegitimate those individuals for which the logarithm of that probability is under a limit equal to Q1 – 1.5 x (Q3-Q1), where Q1 and Q3 are the first and

third quartiles respectively of the distribution of the logarithm of the probabilities for the set of individuals.

By applying this method to the 385 individuals analysed, for which the parents were also analysed or reconstituted, a threshold of 33 is found, which identifies 45 illegitimate individuals belonging to 11 families.

By recalculating the probabilities for each individual by assuming its parents to be all the pairs of parents achievable by crossing two analysed parents and no longer only the parents from whom they are assumed to come, it happens that values are found which are over the illegitimacy limit and over the probability obtained for the original cross. In the case where the individual is illegitimate, that suggests an alternative cross to which the individual belongs in all likelihood. In the opposite case, where the individual is legitimate, this alternative probability is never different enough from the original probability to justify casting doubt on the legitimacy of the individual.

Proportion of truly illegitimate individuals that are not detected as illegitimate

By artificially simulating illegitimate individuals, which are obtained by choosing at random certain individuals identified as being legitimate but assigning to them a false cross also drawn at random (avoiding the true cross), an approximation is obtained of the diversity of situations of all the illegitimate individuals.

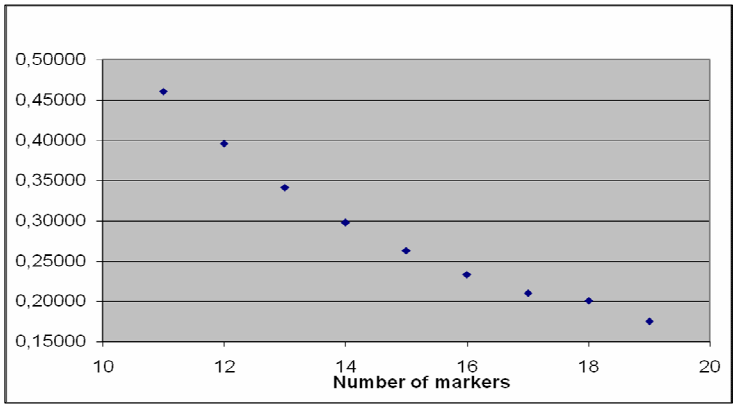

Figure 3: Average error percentage depending on the number of markers.

A base of 4,000 illegitimate pseudo-individuals, each corresponding to the drawing of an individual and a false cross, has been constructed using this procedure. By applying the previously described decision-making rule to these pseudo-individuals, it happens for a certain proportion of them, that they are not seen as illegitimate. That proportion of errors has been determined for all the possible combinations of markers comprising at least 11 markers. The mean error percentage depending on the number of markers of the combinations is shown in figure 3.

It can be seen that the fewer loci there are, the more errors are made on average. As the Li-cor analyser can reveal 4 loci at the same time, it is worth fixing a number of loci to be

kept that is a multiple of 4. It is therefore worth examining the results obtained for 12 markers.

There are around 50,000 possible choices of 12 loci from among the initial 19 loci. For those 12-locus combinations, the estimated error percentage ranges from 0.1% to 1.3%. 1,347 combinations obtain an error percentage under 0.2%, which is the error percentage for the initial combination of 19 loci (Figure 4).

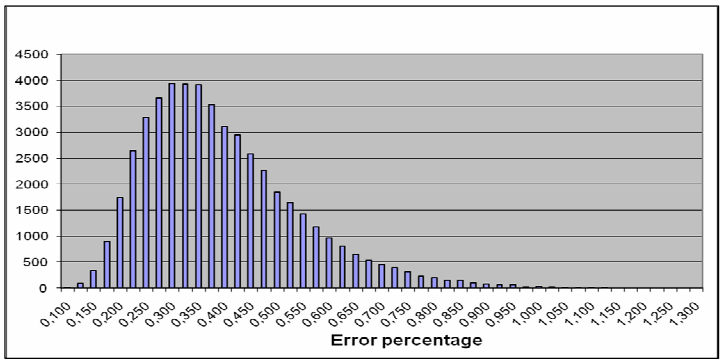

Figure 4: Distribution of combinations with 12 loci depending on error rate

These estimated error percentages for each combination necessarily depend on the set of 4,000 data simulated but also the 385 observed data that served as a basis for the simulation. It even seems that this influence of the initial set of data is unfortunately quite large since a preliminary analysis carried out before correction of a few errors due to the coding of the missing data indicated that the best combination was a combination that did not figure among the 1,347 combinations considered to be efficient with the current corrected dataset. This partly comes from the fact that, in the end, the error rates are rather low and fairly grouped for a majority of combinations. Indeed, ¾ of the combinations have error percentages under 0.5%. That is a relatively small range compared to the uncertainty of estimating the error rates, which is at least equal to the uncertainty arising from the sampling of the 4,000 data. For such percentage levels (0.5%), the uncertainty of sampling alone leads to a confidence interval of ± 0.2%.

It is found that some markers are frequently involved in the best combinations: they are each involved in more than 1,000 combinations out of the 1,347 12-locus combinations as efficient as the initial combination. Conversely, some markers are found in fewer than 500 of the best 1,347 combinations. The markers frequently associated with the best combinations therefore each seem to bring individually an ability to distinguish between crosses. Conversely, the less represented markers seem to provide less information. It therefore seems more reasonable to seek the combination to be adopted from among the 106 combinations that jointly possess all the favourable markers and which, conversely, do not include any of the least informative markers.

In order to distinguish between these 106 theoretically satisfactory combinations, the mean of the error rates generated by all the 11-locus sub-combinations obtained by

eliminating one of the initial 12 loci was calculated for each of them. This makes it possible to determine the most robust combinations, i.e. which remain informative even if one of the markers is missing.

Lastly, as several combinations obtain similar results, it seems reasonable to choose a combination whose estimated quality does not depend too much on the hypotheses put forward to estimate the error percentage and, in particular the estimated amplitude of the uncertainties for determining the number of base pairs. Finally a combination of 12 markers, which should theoretically be the most efficient were proposed.

LEGITIMACY ANALYSIS

The first intuitive approach is to read the data and note the impossible cases: when the probability of observing the genotype, knowing the parents, is nil. This approach can be assisted by automated calculation of such cases.

However, this approach is not enough. Indeed, given a sometimes high degree of homozygosity, not all anomalies are detected without more advanced calculation. Let us take an example to illustrate this. PO 3052 D is planted at Pobè (Inrab, Bénin) and has been planted as being derived from cross DA 115 D x DA 3 D. The set of palms planted with it indeed belong to cross DA 115 D x DA 3 D (or at least there is nothing to indicate the opposite). A single marker (MS 5), out of 12, gives us a contrary indication for this parent. If we assume that this information is not available, then all the other cases are possible. However, when taking a closer look, we realize that, on two alleles of the father, each time one is common with the mother and the other is not (case MS 1, 3, 4, 7, 8, 11), then it is the common allele with the mother that would seem to have been chosen. The probability of such an event is (½)^6, which is little (Table 1).

Table 1: Example of an illegitimate palm which may be difficult to detect

		MS 1		MS 2		MS 3		MS 4		MS 5		MS 6	
DA 115 D ♀	**DABOU**	**127**	**135**	**307**	**311**	**201**	**201**	**210**	**212**	**217**	**239**	**154**	**160**
DA 3 D ♂ reconstituted	**DABOU**	**131**	**135**	**307**	**309**	**201**	**207**	**182**	**212**	**213**	**245**	**154**	**154**
PO 3052 D	DA 2631	135	135	309	311	201	201	210	212	239	239	154	160

		MS 7		MS 8		MS 9		MS 10		MS 11		MS 12	
DA 115 D ♀	**DABOU**	**229**	**229**	**360**	**360**	**250**	**228**	**289**	**301**	**122**	**122**	**214**	**216**
DA 3 D ♂ reconstituted	**DABOU**	**229**	**243**	**360**	**366**	**228?**	**238**	**289**	**301**	**122**	**134**	**214**	**216**
PO 3052 D	DA 2631	229	229	360	360	250	228	301	301	122	122	216	216

More generally, different methods are used to calculate the probability of belonging to one cross rather than another. The method ultimately proposed also takes into account as well as possible the measurement uncertainties mentioned in point 5. It will be useful to carry out these calculations systematically and propose a warning each time there exists a more likely cross than the one hoped for.

SEARCH FOR PARENTAGE

In the event that some palms are illegitimate, it is worth trying to find their probable ancestry. Here too there are different methods.

Benoît Cochard proposes several approaches. By making alternative use of software intended for studying the structuring of genetic diversity, the field of the search can be reduced when there are co-localizations. It is in this way that PO 4102 T and PO 4104 T have been drawn closer to origin "YA 3" then family LM 426 T self.

A principal components study of molecular marking data enables quite a similar approach to be taken, as well as a discriminant analysis.

"FaMoz" software could also be used, if the search for parentage can be properly parameterized so that the results are not too wide-ranging.
Lastly, with the same calculation as in point 5, it is possible to obtain the most likely cross(es) if the parents are in the available database, which is not necessarily the case.

CONCLUSION

This project has made good headway. A set of the 12 markers has been proposed on a rational basis and has been adopted and is routinely used. A method which take into account DNA gel reading uncertainty has been developed which allowed the calculation of the probability of observing the genotype of an individual. From that calculation a probability limit for legitimacy/illegitimacy has been proposed. At the end, routine analysis is performed with warnings and proposals to search for parentage according to different methods.
ID checking should be used for two purposes:

- During breeding process to assess that each result, for example in a progeny test is attributed to the right parent.
- During seed production. As male parents can give a huge number of seeds, they should be individually proved to be legitimate. For females parents, at least the families should be proven to be legitimate: the legitimacy of a significant sample of palms (13 is a good number) from the family will be assessed.

This has to become part of quality control as well during research programmes as for seed production.

BIBLIOGRAPHY