Biochemical and molecular methods for characterizing coconut diversity

P Lebrun[1], A Berger[2], T Hodgkin[3] and L Baudouin[4]

[1]Molecular Biologist, [2]Molecular Biology Technician and [4]Geneticist, Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Montpellier, Cedex 5, France

[3]Principal Scientist, International Plant Genetic Resources Institute (IPGRI), Rome, Italy

Introduction

The various coconut palm collections worldwide are veritable genepools from which geneticists can tap to create or improve existing varieties. However, if a collection is to be used, it needs to be correctly labelled and cultivars in collections need to be precisely described, in terms of their morphological and genetic traits. As the latter are independent of the environment, they amount to a reliable cultivar identity card. Once precise labelling has been completed, the collection can be reduced by discarding duplicates and by limiting the number of representatives in the case of cultivars with low polymorphism. Such a reduction in numbers leads to considerable savings in conservation costs. Coconut palm collections cover large areas of land and thereby entail substantial management, conservation and renewal costs. Once cultivars have been correctly identified, it becomes possible to compare them on a worldwide scale. In this way, it can be seen that specific types such as the Malayan Yellow Dwarf designation may conceal different cultivars or off-types.

In a disease control context, Lethal Yellowing among others, it is paramount to identify with certainty those cultivars that display tolerance, or even resistance. It is just as important to ensure that any commercially produced hybrid corresponds to what it was claimed to be. This implies that the identity of the parents should be given special attention.

Different genetic markers

Different types of traits have been used to characterize genetic diversity in coconut populations: firstly, using traits which only require observation of the phenotype. Phenotypic observations have the great advantage of being directly related to agronomic traits. However, they are largely subjected to selection and to environmental conditions. A variety of morphologic traits was used by N'Cho *et al.* (1993) to describe the diversity of 17 Tall cultivars from the whole coconut cultivation area, and fruit traits were used by Ashburner *et al*. (1997) to describe the structuring of 29 cultivars from the South Pacific (28 Talls and 1 Dwarf) into genetic

groups. In an initial study involving 17 morphological fruit traits, Zizumbo-Villarreal and Piñero (1998) separated Mexican coconut palms into three groups: two groups of Pacific Talls and one group of Atlantic Talls. In a more recent study, 19 traits were used by Zizumbo-Villareal and Colunga-García Marín (2001) to characterize 18 populations of coconut palms, with or without the presence of Lethal Yellowing. Unfortunately, these analyses did not make it possible to compare the same cultivar at different sites or over different time spans. Apart from fruit traits, most other morphological characteristics reveal mainly the dichotomy between Tall and Dwarf cultivars. It was therefore necessary to develop new tools for characterizing cultivars, independently of their growth habit and associated traits.

Biochemical markers

Biochemical markers, such as isozymes (Benoit and Ghesquière 1984; Cardeña *et al.* 1998) or polyphenols (Jay *et al.* 1988), were first used at the beginning of the 1980s to describe the diversity of coconut collections. Unlike with morphological traits, biochemical markers do not require measuring different characters from different organs in a full-sized palm. It is enough to take an organ sample (leaflet, root, etc.) to reveal the biochemical identity of the palm.

Isozymes

Isozymes, which are different forms of the same enzyme, are proteins produced by RNA translation. They are therefore genetic markers, but given their low polymorphism in coconut, isozymes proved to provide very little information for diversity studies. Nonetheless, isozymes were used by Fernando and Gajanayake (1997) on six cultivars from Sri Lanka. Of the six enzyme systems tested on leaf extracts, only two monomeric esterase loci and one dimeric peroxidase locus proved to be polymorphic, with 2, 3 and 2 alleles, respectively. These results tally with those found by Hartana *et al.* (1993), who tested six systems, three of which were polymorphic. No genetic determinism of the bands was suggested. The most important study was carried out with isozymes by Benoit and Ghesquière (1984), who screened 31 enzyme systems, 12 of which displayed little or no activity, 10 were non-polymorphic and 9 were legible and polymorphic. Although the genetic determinism of the nine systems was not entirely elucidated, only two alleles per locus were found. Cardeña *et al.* (1998) were no luckier, since they discovered two enzyme systems (peroxidase and endopeptidase) with two alleles each from testing four cultivars and two hybrids. Nonetheless, isozymes can be used for one-off studies, such as differentiating between the Rennell Island Tall (RIT) and the West African Tall (WAT) (Cardeña *et al.* 1998).

Polyphenols

Leaf polyphenols are substances involved in plant defence reactions to aggression. They form a family of molecules of controlled chemical formula and abundance. The different types of polyphenols can be identified by HPLC (High Performance Liquid Chromatography). These markers were used by Jay *et al*. (1988) to sketch out an initial picture of coconut genetic diversity. However, later studies revealed that the polyphenol profiles had low repeatability, probably due to a major environmental effect. It was therefore difficult or even impossible to compare collections with each other, or to compare the same collection over several years. At the beginning of the 1990s, new types of markers came into being, which were much nearer to the genetic basis than chromosomes were. These markers, which were directly linked to the genome, offered the advantage of being independent of the environment. They were molecular markers.

Molecular markers

Different types of molecular markers were used at the outset to describe the diversity of coconut collections, and more recently to identify cultivars or individuals. What these markers had in common was to be directly linked to the genome, hence, in theory, to be independent of the environment, the age of the plants or their phytosanitary condition. Each genetic marker corresponded to a DNA sequence located at a precise spot of the genome (or locus), whose polymorphism between individuals was shown using different tools, which generally included enzymatic digestion of DNA, specific or random multiplication of selected sequences [polymerase chain reaction (PCR) amplification]. This polymorphism corresponded to variants (or alleles) of the sequence being studied. It appeared in the form of bands on electrophoresis gel after developing with a radioactive or non-radioactive stain (e.g. silver nitrate or fluorochrome). Depending on the case, a single band was associated with a locus and polymorphism corresponded to the existence or absence of the band ('dominant' markers). In other cases, each allele was associated with a different band ('codominant' markers), making it possible to differentiate between a homozygous individual and a heterozygous individual at the locus in question, which was not the case with dominant markers. Different types of markers were differentiated by the type of sequences considered, their position in the genome (nuclear or cytoplasmic genome, encoding regions or not), the detection technique, their specificity and reliability, and the mutation rate, which determined the evolution rate and degree of polymorphism observed. Lastly, the cost of analyses was an important choice factor.

PCR is a technique for the amplification of a DNA segment between two known sequence regions. Using specific conditions and procedures

(Erlich 1989), selected sequences are multiplied many times so that variation in them can be detected and described. PCR amplification currently takes between two and three hours, depending on the thermocycler used and the hybridization temperature adopted.

Restriction Fragment Length Polymorphism (RFLP)

Characteristics. RFLP markers are codominant markers that are more or less specific depending on the probe used (cDNA, genomic DNA, etc). They are difficult and quite expensive to use. They require the extraction of a large quantity of good quality DNA and operations are lengthy. It takes around two weeks after DNA extraction to read the bands.

Starting in 1995, various diversity studies were undertaken at CIRAD (Centre de Coopération Internationale en Recherche Agronomique pour le Développement) using cDNA nuclear probes or cytoplasmic probes (the latter being quite well conserved from one plant to the other (Lebrun *et al.* 1998; Lebrun *et al.* 1999). Cytoplasmic genomes are usually much less polymorphic than the nuclear genome. However, they can be a source of information for establishing phylogenies, or retracing the domestication routes of a plant from its region of origin (Lebrun *et al.* 1999).

Nine cytoplasmic probes (Mitochondrial: Apocytochrome b, sub-units alpha and 6 of ATPase, cytochrome oxidase (cox) sub-units 1, 2 and 3, Chloroplast: Cp IR, Cp sal6 and Cytochrome F) where hybridized with coconut DNAs digested by nine enzymes (*Hind*III, *Eco*RV, *Eco*RI, *Ss*tI, *Bam*HI, *Bgl*II, *Dra*I, *Hae*III, *Hpa*I). Probe Cox1 displayed clear polymorphism, which was, moreover, identical for the two digestion operations: *Ss*tI and *Bgl*II. Cox 2, which was difficult to interpret, could show polymorphism with *Hae*III digestion. All the other enzyme-probe pairs were monomorphic. Consequently, little mitochondrial polymorphism was found, and no chloroplast polymorphism. Perera (2002) has confirmed these results in a study on the chloroplast genome.

The nuclear genome proves to be much more polymorphic. For instance, during studies conducted in 1995 and 1997, two batches of different probes were used to explore coconut palm diversity. In the first study (Lebrun *et al.* 1998), nine rice cDNA probes led to the discovery of 40 polymorphic bands, whilst in the second study (Lebrun *et al.* 1998), 20 rice, oil palm, maize and coconut cDNA probes revealed 60 polymorphic bands.

Procedure. Plant DNA was digested by a restriction enzyme that cut the DNA at a particular site. A large number of fragments were obtained in this way. These fragments were then separated according to their size by migration in agar gel under the influence of an electric field. After

migration, the DNA was transferred to nylon membrane, then exposed to a probe (small DNA primer specific to the locus studied), which was either radioactive-labelled or non-radioactive-labelled (antibody-antigen complex, e.g. Digoxygenine).

Irrespective of the number of probes used, or their origin, the results as regards diversity structuring were comparable. However, this technique is laborious to use. In addition, with the discovery of the PCR technique, new types of markers came into being.

Randomly Amplified Polymorphic DNA (RAPD)

Characteristics. These non-specific (the DNA revealed may belong to another species, such as a fungus) codominant markers, once amplified by PCR, are quite simple to use and only require a small amount of DNA. They are obtained very quickly. Random primers are marketed in kit form. After extraction, it takes two days to display the bands.

Their repeatability remains doubtful and it seems quite risky to use them for comparisons between laboratories, or over different time spans.

Their ease of use and low cost have made them very attractive, but they are difficult to read and their low reproducibility makes them more useful for genetic mapping rather than diversity studies. Nevertheless, they can be used for one-off studies, such as analyzing a few cultivars from the South Pacific (Ashburner *et al.* 1997).

Procedure. DNA was amplified by PCR using short random primers. The amplification products were separated by migration in agar gel then displayed using a DNA intercalator, Ethidium bromide.

Amplified Fragment Length Polymorphism (AFLP)

Characteristics. These are dominant PCR markers, which are quite easy to use but sometimes difficult to interpret. A large number of markers can be developed on the same gel, which often makes for difficult reading, hence their limited use in diversity studies. Moreover, on coconut, each combination tested has few polymorphic bands. These markers are primarily used to saturate genetic maps. It takes around one week after extraction to display the bands (Perera *et al.* 1998; Teulat *et al.* 2000; Lebrun *et al.* 2001).

Procedure. Stage 1: Plant DNA was digested by two enzymes, one rarely cutting (recognition site of 6 bases, e.g. *Eco*RI), the other cutting more frequently (recognition site of 4 bases, e.g. *Mse*I).

Stage 2: Double stranded adaptors were ligated to the ends of the DNA fragments to serve as templates for amplification. Thus, the sequence of

adaptors, followed by the adjacent restriction site, served as the ligation site for the restricted fragment amplification primers.

Stage 3: In order to reduce the number of fragments amplified, the amplification primers were elongated in 3′ (restriction site side) by one or two bases. Only the restricted fragments possessing the selective base(s) just after the restriction site were amplified. These fragments were separated by migration in acrylamide gel under the influence of an electric field. Visualization was either radioactive or non-radioactive (fluorochrome on automatic sequencers, silver nitrate, etc.).

Inverse Sequence-Tagged Repeat (ISTR)

Characteristics. These are specific, dominant and extremely polymorphic markers. However, most of the polymorphism is primarily found within populations, making them of little use for diversity studies. For instance, it is the only way of observing polymorphism within the Malayan Yellow Dwarf from Tanzania and from the Philippines (Rohde *et al.* 1995). Their development is lengthy and costly. Detection of polymorphism is more random than with microsatellite markers. It takes three days after extraction to visualize the bands.

Procedure. Coconut genomic DNA was amplified by primers specific to regions separating repeated sequences. The PCR products were separated by migration on acrylamide gel, and then detected by radioactivity incorporated during amplification.

Microsatellites or Simple Sequence Repeats (SSR)

Characteristics. These are codominant PCR markers, for which it is easy to identify alleles and loci, thereby enabling their use in population genetics. Their development is expensive, but their routine use is affordable. They are very useful for mapping studies, as they can be used to compare different maps with each other. They are highly polymorphic, highly repeatable, and easily transposable from one laboratory to another or from one year to the next. They are ideal markers for diversity studies (Perera *et al.* 1999; Rivera *et al.* 1999; Perera *et al.* 2000; Perera *et al.* 2001; Meerow *et al.* 2003). They formed the basis for a coconut diversity study and cultivar identification kit developed at CIRAD (Baudouin and Lebrun 2002). It takes around three days after extraction to visualize the bands.

Procedure. Coconut genomic DNA was amplified by PCR using primers specific to regions containing a microsatellite sequence. This type of sequence was characterized by repetition, many times, of a motif formed by 2 or 3 pairs of bases (e.g. GAGAGAGA…, TCTCTC… or

GTCGTCGTCGTC…). As these primers were placed either side of the microsatellite, they enabled specific amplification. As the number of repeats varied from one individual to the next, differences in PCR product lengths were detected by migration in agar gel (and developed with EB) or on polyacrylamide gel (developed radioactively (P33) or nonradioactively (silver nitrate or fluorochrome on an automatic sequencer).

Single Nucleotide Polymorphism (SNP)

Characteristics. SNP are markers characterized by substitution of a nucleotide. They are codominant. Obtaining them requires a great deal of prior sequencing work (e.g. EST (Expressed Sequence Tags) type data obtained on different genotypes), and validation, but their subsequent use is quite easy (PCR, detection). As 'mass' sequencing is not yet available for coconut, SNP markers have yet to be developed.

Procedure. SNP markers are frequent and well distributed in the genome. Their frequency varies depending on the species and on the regions of the genome: from 1 every 3 kb in man, this figure can fall to 131 pb in the case of an EST of barley cytochrome P 450 (Bundock *et al.* 2003) or 54 pb in the case of an EST encoding 6-phosphogluconate dehydrogenase from sugarcane (Grivet *et al.* 2001). Through their variable positions in the genome (coding or non-coding zones), they offer very strong potential for markers which are useful for labelling the genome, or for studying gene regulation.

They are detected in several ways:

- By sequencing (Bundock *et al.* 2003) (microsequencing or primer elongation);
- By analyzing the polymorphism of single stranded DNA conformation;
- By denaturing gradient gel electrophoresis;
- By mass spectrometry; and
- By DNA microchip hybridization (Schmalzing *et al.* 2000).

In the case of sugarcane, Grivet *et al.* (2003) aligned the sequences of different sugarcane cultivars listed in EST databases containing 230 000 sequences. The strictness of such alignments and the quality of the sequences are of great importance for SNP detection. The next stage was finding a restriction enzyme whose recognition site includes the SNP. This SNP was then validated by merely defining the primers either side of the polymorphic site, digesting the amplification products with the enzyme cutting into the SNP, and proceeding with gel migration of the restrictions.

Choice of markers to be used

Consequently, different types of markers can be used depending on the questions involved and the resources available. A comparative study conducted on 31 individuals representative of worldwide diversity showed that AFLP, microsatellite (Teulat *et al.* 2000) and RFLP markers give the same picture of coconut genetic structure. Although they shed some light on the domestication routes taken by the coconut palm (Lebrun *et al.* 1998), RFLP remained a difficult and laborious technique, which could easily be replaced by using PCR type markers. At the present, microsatellite markers seem to offer many advantages. They are quite simple to use, and enough of them exist to choose from and use the most efficient primers.

The coconut microsatellite kit

In connection with a project funded by the International Coconut Genetic Resources Network (COGENT) of the International Plant Genetic Resources Institute (IPGRI), the European Union (EU), BUROTROP (Bureau for the Development of Research on Tropical Oil Crops) and CIRAD, a kit has been developed for coconut diversity studies and cultivar identification. The purpose of the kit is to evaluate genetic diversity using microsatellite markers based on standardized methods that can be used by any laboratory with a minimum of equipment.

Kit contents – The microsatellites

The coconut microsatellite kit consists of:

- Primer sequences available for use by partners in developing countries;
- A set of coconut reference population data, consisting of allelic frequencies for all the microsatellite loci of the kit. This set is representative of global coconut diversity and serves as a reference for further studies;
- A document listing the procedures to be adopted for analysis of diversity in coconut using the microsatellite kit, including both experimental and data analysis protocols; and
- Software called Geneclass2 for assigning individuals to cross-fertilizing populations.

Out of 83 microsatellite primer pairs screened for their ease of development, reproducibility, legibility and number of alleles, 14 were chosen, four of which could be multiplexed by two. The kit also includes a technical manual for laboratory operations, along with population assignment software.

The 14 primer pairs in the kit (Table 1) have been used to study diversity on 571 individuals, spread over 136 cultivars. That figure is continuing to increase, since new populations are being characterized each month using this kit. The results are entered in the Coconut Genetic Resources Database (CGRD) and are accessible to the members of the COGENT network.

Table 1. The 14 microsatellite kit primer sequences

Locus	Repeat array	Primer sequences (5'- 3')	Size range (bp)	T1* (bp)	T2** ((bp)	Embl Genebank Accession no.
CnCir A3	(TG)15	AATCTAAATCTACGAAAGCA AATAATGTGAAAAAGCAAAG	228-248	228 228	240 240	AJ458309
CnCir A9	(GT)9 (GA)8	AATGTTTGTGTCTTTGTGCGTGTGT TCCTTATTTTTCTTCCCCTTCCTCA	89-115	097 097	089 089	AJ458310
CnCir B6	(GT)4 (CT)2 (GT)10 (GA)11	GAGTGTGTGAGCCAGCAT ATTGTTCACAGTCCTTCCA	196-226	196 204	202 202	AJ458311
CnCir B12	(CA)20 (GA)15	GCTCTTCAGTCTTTCTCAA CTGTATGCCAATTTTTCTA	135-189	163 163	169 169	AJ458312
CnCir C3'	(CA)12 X21 (GC)6 (AC)10 (AG)12	AGAAAGCTGAGAGGGAGATT GTGGGGCATGAAAAGTAAC	174-232	178 206	176 176	AJ458313
CnCir C7	(GT)7 (GA)16	ATAGCATATGGTTTTCCT TGCTCCAGCGTTCATCTA	147-189	165 167	161 161	AJ458314
CnCir C12	(CA)15 (TA)6	ATACCACAGGCTAACAT AACCAGAGACATTTGAA	161-185	167 167	183 183	AJ458315
CnCir E2	(CT)17 (GT)9	TCGCTGATGAATGCTTGCT GGGGCTGAGGGATAAACC	115-177	163 163	135 135	AJ458316
CnCir E10	(CA)8 (GA)11	TTGGGTTCCATTTCTTCTCTCATC GCTCTTTAGGGTTCGCTTTCTTAG	226-246	244 244	238 238	AJ458317
CnCir E12	(CT)6 (CCT)2	TCACGCAAAAGATAAAACC ATGGAGATGGAAAGAAAGG	162-174	174 174	164 164	AJ458318
CnCir F2	(TG)11 (AG)12	GGTCTCCTCTCCCTCCTTATCTA CGACGACCCAAAACTGAACAC	191-215	193 193	205 205	AJ458319
CnCir G11	(GT)9 (GA)9 TA (GA)4	AATATCTCCAAAAATCATCGAAAG TCATCCCACACCCTCCTCT	186-212	204 208	194 194	AJ458320
CnCir H4'	(TC)8 X4 (CA)5 (CGCA)5	TTAGATCTCCTCCCAAAG ATCGAAAGAACAGTCACG	218-236	230 230	230 230	AJ458321
CnCir H7	(CT)16 (CA)13	GAGATGGCATAACACCTA TGCTGAAGCAAAAGAGTA	127-149	133 133	139 139	AJ458322

*T1 = standard 1 = WAT 4

**T2 = standard 2 = MYD

Eventually, as the 14 microsatellites of the kit are available from CIRAD or IPGRI, this base should be enhanced by results from all the countries possessing collections. It will then be possible to compare the genetic profiles of the populations in collections in different countries and conclude on the identity or difference between cultivars with the same name.

In 2002, COGENT supported the training at CIRAD of 18 coconut researchers from nine countries (one biotechnologist and one curator per country) on the use of the microsatellite kit and its associated software. Subsequently, each country was given a research grant by IPGRI /

COGENT with funding from DFID to use the skills learned to characterize their local conserved varieties with at least one of seven varieties with known tolerance/ resistance to lethal yellowing disease as control. This research is currently ongoing.

Geneclass2 software

The kit is currently accompanied by Geneclass2 software developed with Institut National de la Recherche Agronomique (INRA). The need for this software comes from the cross-fertilizing nature of Tall coconuts. Unlike clones or self-fertilizing varieties, a Tall cultivar comprises a set of genotypes, each of which is different. It can only be identified with molecular markers through the frequencies of the different alleles at the loci tested. In addition, the number of individuals observed is always limited, meaning that these frequencies are only known with a degree of uncertainty. Identifying the population of origin of one or more individuals therefore means resorting to probability calculations.

The method adopted is a Bayesian method (Baudouin and Lebrun 2001). It requires establishing a set of samples representative of the main known cultivars, which are then compared to 'candidate' samples to be identified. Geneclass2 software can then assign a probability to each proposal of the type 'the candidate comes from cultivar x'. This probability is a ratio of the 'score' of cultivar x to the sum of the scores of all the reference populations. The score is the probability of obtaining the candidate, given the reference sample. It is calculated taking into account the uncertainty of allelic frequencies.

This probability is calculated by considering the hypothesis that the candidate actually belongs to one of the reference cultivars, which is not necessarily true. Another test therefore has to be carried out and the Geneclass2 'exclusion' method makes it possible to calculate the probability that the candidate belongs to a reference cultivar. For this calculation, it is necessary to simulate a random sample of the genotypes of that population. The probability of belonging to the cultivar amounts to the percentage of simulated genotypes that obtain a lower score than the candidate.

Geneclass2 software performance

The efficiency of the Bayesian procedure was tested by attempting to determine the population of origin of the reference samples. The samples were representative of the main coconut cultivars. Some cultivars were represented by several populations that could differ slightly, and on the average, five individuals were sampled. Each individual was drawn from the reference database before seeking its origin. This precaution was taken

in order to make the test more realistic: in actual population assignment, the tested individual and the reference samples form distinct sets. The individual assignment tests were approximately 50% successful (the population identified was indeed the population of origin). When the scores of the individuals in the same sample were cumulated, precision was substantially improved – the cultivar of origin was correctly identified in 72% of cases[1] (Table 2).

Table 2. Result of assignment test with Geneclass2

Group	Same population	Same cultivar	Same area (could be the same cultivar)	Other (within the same group)	Total
Pacific					
Dwarf	19	0	0	0	19
Panama	2	0	0	0	2
Southeast Asia	11	0	6	0	17
Micronesia, Polynesia	6	4	1	0	11
Melanesia	20	6	12	10	48
Indo-Atlantic	18	0	1	2	21
Total	**76**	**10**	**20**	**12**	**118**
%	**64**	**8**	**17**	**10**	

In 17% of cases, the cultivar identified had another name but came from a neighbouring region. It was highly likely that some biologically identical cultivars were given different names, either because they were described by different people, or because the corresponding populations had different morphometric characteristics due to environmental or age differences. It was therefore reasonable to assume that a proportion of the 17% actually corresponded to correctly classified cultivars.

In 10% of cases, the samples were attributed to cultivars in the same group, but from another region. In a small number of cases, assignment was ambiguous. This may have involved cultivars for which genetic differences were too slight to be picked up with the number of individuals used. The possibility of germplasm exchange over long distances could not be ruled out either.

The method gave different results depending on the groups of populations. For instance, each Dwarf cultivar was correctly identified. These cultivars are self-fertilizing and extremely homogeneous, making it easier to distinguish between them. On the other hand, most of the cultivars that were imprecisely identified came from Melanesia. In fact, more populations were sampled from this region than from any other. As a result, some populations were very similar to each other and thus difficult to distinguish with only five individuals per population.

[1] Agrees with the table: members of the same populations belong to the same cultivar Same population + same cultivar = 64% + 8% = 72%

When it was used to distinguish between the main coconut cultivars, the Bayesian method used in Geneclass2 proved to be efficient in identifying coconut cultivars, even with a small sample size. This exercise became more difficult only when closely related populations were involved. In fact, the efficiency of the method depended on three factors: the sample size (both in the reference samples and in the candidates), the genetic diversity of the studied populations and the genetic divergence between populations. For this reason, larger samples are needed to study fine genetic structure on a regional scale than on a global scale. One of the important characteristics of the method is that with time, more samples can be included in the reference set, making it possible to improve its discriminating power.

Use for classification purposes

Groups of populations were compiled based on geographic origin and genetic structure. Each time when it was doubtful that a population belonged to a group, confirmation was obtained by checking that its classification remained the same when its data were excluded from the reference database. In that way, only populations MXPT2 (Mexican Pacific Tall), Colima and GGZ (one of the Gazelle Peninsula Tall samples) remained unclassified. The former case may reflect the fact that coconuts were introduced into Mexico from different countries (Zizumbo Villarreal 1996). The Colima population could result from the intercrossing between populations of Melanesian and Southeast Asian origin. On the other hand, the GGZ sample appeared to be an illegitimate accession of the Gazelle Peninsula Tall. The proposed classification is given in Annex 1.

Results

Studies using samples representing the worldwide coconut diversity were carried out using RFLP (Lebrun *et al.* 1998; Lebrun *et al.* 1999), microsatellite (Baudouin and Lebrun 2002), or AFLP (Teulat *et al.* 2000) markers. All these studies detected two major cultivar groups. The first was the Pacific group that was both the most polymorphic and the geographically most extensive. It spreads from Southeast Asia to the east coast of Latin America. It includes four sub-groups with blurred boundaries: Southeast Asia, Melanesia, Micronesia and Polynesia. The coconut palms from Panama and Peru form a fifth sub-group, related to the precedents, but clearly distinct. All Dwarfs, irrespective of their geographical origin, form a genetically uniform group that belongs to the Pacific group, sub-group Southeast Asia. The Niu Leka Dwarf is a notable exception as this cross-fertilizing Dwarf originated from the South Pacific.

The second was the Indo-Atlantic group that originated from the Indian subcontinent, from where it was subsequently transported to West Africa and the Atlantic coast of Latin America. East Africa is also populated with coconut palms of the Indo-Atlantic type, though these received an input from cultivars of Southeast Asian origin, resulting from Austronesian migrations to Madagascar.

For the most part, these results are confirmed by other studies; the distinction between the Indo-Atlantic and Pacific groups has been observed with other markers such as ISTR (Rohde *et al.* 1995; Fernando *et al.* 1997) or in a partial study carried out with RAPD markers (Wadt *et al.* 1999). The same applies for the AFLP gathered by Perera *et al.* (1998) where the main two groups were formed by local cultivars and Dwarfs of the Pacific group, along with the "auranthiaca" (King coconut and Rathran Thembili) genotypes, which actually came from ancient hybridization between the Dwarfs and local genotypes. In the study by Rivera *et al.* (1999), microsatellites showed that Dwarfs formed a uniform group within the local Talls and clearly stood out from the Pacific coconuts (RIT or Rennell Island Tall and PYT or Polynesian Tall). Minor differences exist with the RAPD study by Ashburner *et al.* (1997), which focused on the South Pacific and proposed two groups, North and South. Lastly, the classification of Florida populations by Meerow *et al.* (2003) was the only one to group the Jamaica Tall with the Panama Tall. This casts doubt on the legitimacy of the planting material used. Finally, these two groups are compatible with the results of phenotypic and polyphenol observations (Lebrun *et al*. 1999).

The contribution of microsatellites to a better understanding of coconut genetic diversity can be illustrated by comparing the resulting classification with the theory proposed by Harries (1978). According to him, all coconut cultivars were derived from two types: 1. the '*Niu Vai*' (also called 'domesticated') type, which has rounded nuts with high water content, early germination and an erect stipe and 2. the '*Niu Kafa*' (considered as 'wild' type), which have more or less triangular nuts, thick husk, a slow rate of germination and a more slender and curved stem. Many currently available coconut cultivars are intermediate between those two types. Following this typology, almost all the cultivars that tend towards the *Niu Kafa* type belong to the Indo-Atlantic group and almost all those that were considered by Harries as *Niu Vai* or introgressed belong to the Pacific group.

While concordance is the general rule, differences between these classifications are not difficult to explain: while the reasoning underlying the morphological classification mainly involves natural and artificial selection forces, the molecular classification focuses on the genetic relationships between cultivars and on their regions of origin. Due to

cross pollination, Tall coconut cultivars that have grown in the same region for some time are likely to have a more or less similar genetic structure. As a result, a clear geographic pattern of variation is expected and, indeed, observed with Tall coconuts. The situation is different with Dwarfs, because they have a strong tendency to self-pollinate and tend to conserve their genetic structure irrespective of the place where they are planted. This structure therefore reflects the region where it appeared (i.e. Southeast Asia), rather than the one in which it is found, even if it has been there for a long time.

An example of discrepancy between the two classifications is the Kappadam from India. This is clearly of the *Niu Vai* type according to Harries (1978). However, microsatellites indicate that it is an Indo-Atlantic cultivar, as could be inferred from its place of origin. However, the presence of a few microsatellite markers from Southeast Asia makes it possible to understand the origin of this cultivar. Its ancestors were probably imported from Southeast Asia and strict selection was necessary to enable it to conserve its distinct *Niu Vai* traits over generations, while a large proportion of its genes have been replaced by those of the local populations, which are of the *Niu Kafa* type.

The main arguments of the theory proposed by Harries remain valid, though they do need to be moderated and complemented using molecular marker information. Considering the high diversity found in this region, the hypothesis of a centre of diversification located in the vast archipelago situated between Southeast Asia and PNG (Harries 2002) seems to be confirmed. Before its domestication, the coconut palm probably had characteristics similar to those of the Niu Kafa type, which benefited from a definite advantage for its dissemination over large distances by ocean currents. However, this selective factor was probably less important in the centre of diversification, where the distances to be covered were shorter than in the periphery. It is likely that this facilitated the appearance of the *Niu Vai* type, under the effect of domestication in Southeast Asia and/or Melanesia. In other respects, the *Niu Kafa* type predominates on the Indian subcontinent even after a long period of cultivation. Either Indian farmers had different breeding objectives from those in Southeast Asia (notably fibre production), or, despite their efforts, the populations available to them were too homogeneous to develop significantly towards the *Niu Vai* type. The counter-example given by the Kappadam suggests that the second hypothesis explains at least partly the Indian coconut palm characteristics. Moreover, this reduced diversity in the Indian sub-continent is confirmed by microsatellite markers.

There are cultivars with large and clearly *Niu Kafa* fruits in the South Pacific. The absence of obvious similarities between their microsatellite

profiles and those found in the Indian subcontinent suggest that they evolved independently. During their expansion from Southeast Asia, the Polynesian ancestors might have 'rediscovered' the *Niu Kafa* type, which was very different from the *Niu Vai* types with which they had been familiar. This type was apparently maintained (and maybe accentuated by selection) due to its advantages for the production of fibres, which were valuable for navigation.

Case study: Panama Talls

The Panama Tall is a particularly important cultivar due to its place in the Pacific group and the role it plays in genetic control of Lethal Yellowing. Historical data show that it existed on the Pacific coast before the Spanish arrived (Zizumbo-Villareal and Quero 1998) and, along with the Peru Tall, it forms a group with a narrow genetic base. Five different origins of Panama Tall were compared. These were: (1) Nine PNT Aguadulce individuals (called Aguadulce IC); (2) Twenty Monagre individuals (called Monagre IC) conserved at the Marc Delorme Station in Ivory Coast; (3) Ten PNT individuals from Jamaica of 'Bowden' origin (called Bowden Jamaica); (4) Twelve PNT Aguadulce individuals from Nicaragua, of which seven displayed a typical phenotype (called Typical Nicaragua Aguadulce) and five of which were more or less atypical (called Offtype Nicaragua Aguadulce); and (5) Fifteen PNT individuals from Costa Rica sampled in Nicaragua (called Costa Rica). Four Peru Tall individuals (called Peru) were added. These 70 palms were analyzed with the 14 microsatellites in the Kit. Profiles were obtained by characterizing each individual for the 14 systems. The profiles were coded according to the number of alleles and then entered on a spreadsheet. All the samples were statistically analyzed as being individuals not attributing a priori to populations.

Principal component analysis (PCA) was used to present the distribution of population diversity in the plane corresponding to the first two components. In this PCA (see Figure 1), the populations were organized into three distinct groups: the central group contained the populations of Monagre IC and Bowden Jamaica origin, along with the Peru Tall. The resemblance between the genetic structure of the latter and that of the Panama Tall was such that these two cultivars could be considered synonymous. It implied a common origin, despite the distance separating them. The group on the left corresponds to the two populations of Aguadulce origin, which were indistinguishable using molecular markers, irrespective of their geographical provenance or their phenotypic appearance. An examination of the alleles distinguishing the Aguadulce origin from the other Panama Talls showed that this origin displayed a

low percentage of genes from 'Alto Atlantico' coconut palms. This doubtless explains its greater phenotypic diversity. The Costa Rica origin forms the third group, which stands out from the central group through a few rare alleles, of unknown origin. The groupings shown here were confirmed using the Geneclass2 assignment procedure.

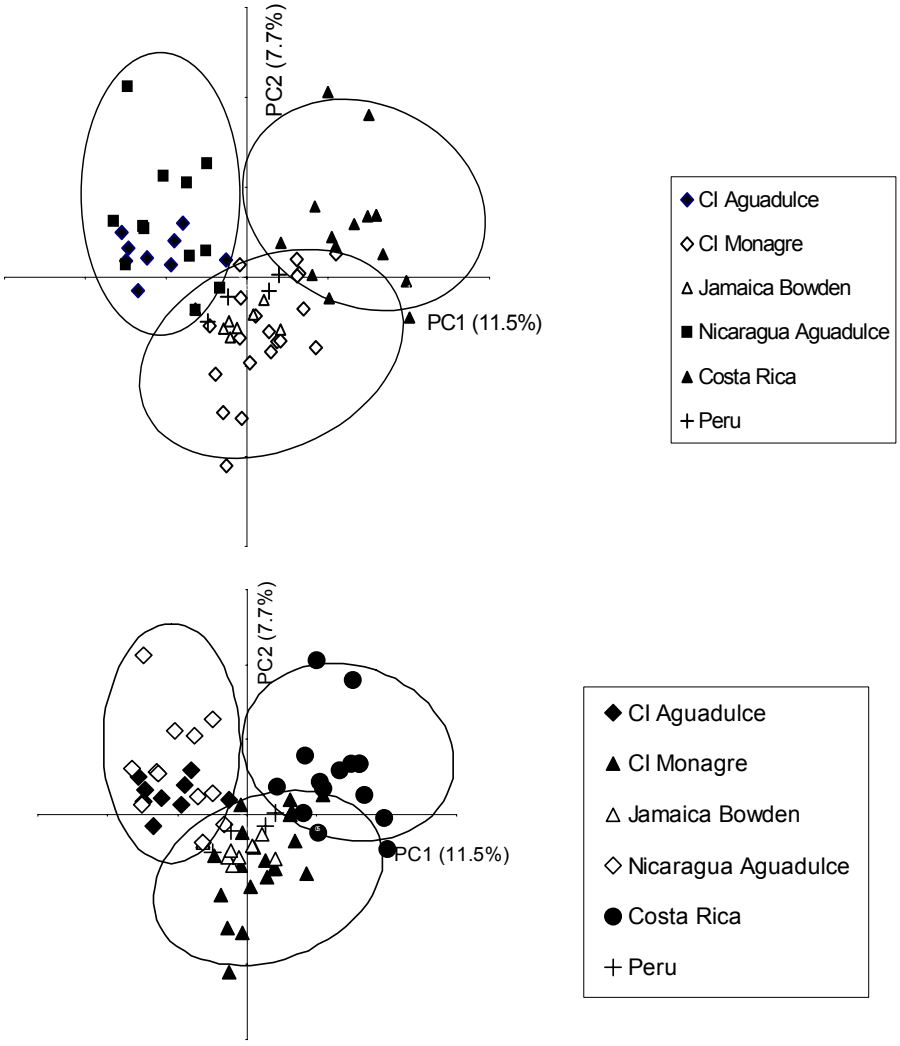

Figure 1. Distribution of population diversity of five different origins of Panama Tall along with Peru Tall

Most modern authors (Child 1974; Harries 2002) agree that the coconut palm does not originate from America, even though it has been present on that continent for a long time. This raises three questions: where did the founding individuals come from? How did they get there? When did they get there?

These three points were examined using Shannon's information theory. The resemblance between the allelic structure of the Panama Tall and that of coconut palms from different regions of the Pacific was evaluated by calculating a parameter called "ambiguity" (the less a marker provides information likely to distinguish between two origins, the greater is the ambiguity). It turned out that, despite its particular genetic structure, the Panama Tall most resembles the coconut palms of Southeast Asia, followed by that of Melanesia, Polynesia and Micronesia.

There are several arguments in favour of a Southeast Asian origin. Firstly, several plants seem to have followed the same route, including the plantain banana and bamboo. Secondly, archaeological remains dating back 2500 years, and revealing several Southeast Asian traits, have been found in Ecuador and Bahía de Caraquez (Estrada and Meggers 1961). Unlike in Panama and Peru, there is no direct evidence of a pre-Columbian existence of coconut palms in Ecuador, but if the coconut palm effectively reached America in Ecuador, it explains the existence of the same cultivar in Peru and Panama.

The Southeast Asian origin of "Bahía Culture" still arouses bitter discussions, but it nonetheless remains the case that the expansion of Austronesian peoples to Polynesia on the one hand and to Madagascar on the other hand, clearly demonstrates that those people in the remote past had the ability to undertake long sea voyages. This provides an answer to the third question raised above; the distance covered (almost 18000 km) is well beyond the possibilities of dissemination by floating on ocean currents.

Conclusion

With the microsatellite kit, molecular markers have become a powerful tool to explore genetic diversity in coconut. Although molecular markers have (in principle) no direct connections with phenotypic variation, they contribute to a better understanding of its distribution, because the distributions of characters were shaped by the evolutionary history of the populations. By combining historical records, morphological and microsatellite variation, it is possible to retrace important features of the history of the crop since its domestication. This makes it possible to understand the relationships among the cultivars found around the world. Above all, it is useful for an efficient use of genetic resources.

Managing genebanks is very expensive. For this reason, it is necessary to avoid any unnecessary duplication. By characterizing populations *in situ*, it is possible to introduce only populations with an original genetic pattern into collections. It is also possible to screen collections in order to spot possible 'synonym' or duplicate accessions. Such accessions do not need to be replicated further.

A second application in genebanks is to ensure legitimacy of their accessions. Plant breeding activities involve substantial resources, which can only be efficiently used if the planting material is correctly identified. If a significant number of foreign genes are found in an accession, it is worth considering resampling the cultivar. This is particularly true if the characters involved are difficult to assess, such as resistance to diseases. It was shown that some of the Panama Tall accessions display a certain percentage of 'Alto Atlantic' genes, irrespective of their phenotypic appearance. Illegitimacy in the parental populations could explain in part the fact that 'Maypan' hybrids were found to be susceptible to Lethal Yellowing. In any event, illegitimacy in disease resistance trials is likely to lead to false conclusions about resistance level.

Knowledge of genetic relationships between cultivars is also important for breeding, particularly in the framework of a global programme. It provides breeders with a way of exploiting the results obtained elsewhere. By finding the closest relative of already tested cultivars in their own collections, they can reproduce (at least in part) their useful features. One of the reasons why such reproduction may be only partial is the effect of genetic-environment interaction. Conversely, knowing the relationships between cultivars may help in predicting interactions. A good example is given by the Dwarf group, including the Marshall Green, the Kiribati Green, the Raja Brown, the Madang Brown and the Sri Lanka Green Dwarfs. In the last two, nuts are relatively large (for Dwarfs) in the Pacific Ocean, but small in West Africa. Knowledge of the genetic distance between cultivars also helps to choose right parents for hybridization, be it for producing hybrid varieties, or for the introgression of useful traits. In the first case, it was shown in Baudouin (1999) that a substantial degree of heterosis might be obtained by crossing cultivars from different molecular groups. In the second case, crossing distant cultivars will maximize genetic diversity to select from in the second and subsequent generations. Such crosses make it possible to develop the best of marker-assisted selection programmes.

Further studies on genetic diversity will be important for a clearer understanding of genetic diversity at different levels. Participatory research appraisals often lead to the identification of more coconut types than conventional 'random' sampling. The question is whether those

types represent normal variation in the population or different genetic origins. In the efforts to find a suitable planting material policy in the presence of Lethal Yellowing, it will be important to study pathogen diversity in relation to the prevailing coconut varieties. Finally, the legitimacy test provided by Geneclass2 is not applicable for assessing the quality of hybrid seed nuts for the moment, because between-population hybrids are not in Hardy-Weinberg equilibrium. Applying suitable genetic model will require further statistical and software developments. A provisional classification of coconut cultivars is presented in Annex 1.

Acknowledgement

We are grateful to Dr JHA Barker for her helpful discussions about coconut microsatellite kit development, and K. Devakumar for his considerable assistance in SSR testing. IPGRI, COGENT, the CEC and BUROTROP provided the financial support for the kit development.

References

Ashburner, GR, WK Thompson and GM Halloran. 1997. RAPD analysis of South Pacific coconut palm populations. Crop Science 37:992-997.

Ashburner, GR, WK Thompson, GM Halloran and MA Foale. 1997. Fruit component analysis of South Pacific coconut palm populations. Genetic Resources and Crop Evolution 44**:**327-335.

Baudouin, L. 1999. Genetic improvement of coconut palms. Pp. 45-56. *In:* C Oropeza, JL Verdeil, GR Ashburner, R Cardena, JM Santamaria (eds). Current Advances in Coconut Biotechnology. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Baudouin, L and P Lebrun. 2001. An operational Bayesian approach for the identification of sexually reproduced cross-fertilized populations using molecular markers. Acta Horticulturae 546:81-93.

Baudouin, L and P Lebrun. 2002. The development of a microsatellite kit for use with coconuts. IPGRI, Rome**.** 66p.

Benoit, H and M Ghesquière. 1984. Electrophrèse, compte rendu cocotier. IV. Le déterminisme génétique. CIRAD-IRHO, Montpellier, France.

Bundock, P, J Christopher, P Eggler, G Ablett, R Henry and T Holton. 2003. Single nucleotide polymorphisms in cytochrome P450 genes from barley. Theoretical and Applied Genetics 106:676-682.

Cardeña, R, C Oropeza and D Zizumbo. 1998. Leaf proteins as markers useful in the genetic improvement of coconut palms. Euphytica 102:81-86.

Child, R. 1974. Coconuts. London, Longman. 54p.

Erlich, H. 1989. Principles and applications for DNA amplification. Stockton Press, New York, USA.

Estrada, E and BJ Meggers. 1961. A complex of traits of probable transpacific origin on the coast of Ecuador. American Anthropologist 63:913-939.

Fernando, WMU and G Gajanayake. 1997. Patterns of isozyme variations in Coconut (*Cocos nucifera* L.) populations used for breeding improved varieties. Plantations, Recherche et Développement. 4(4): 256-263.

Fernando, WMU, L Perera and RRA Peries. 1997. An overview of breeding research in coconut: The Sri Lankan experience. Outlook on Agriculture 26:191-198.

Grivet, L, JC Glaszmann and P Arruda. 2001. Sequence polymorphism from EST data in sugarcane: A fine analysis of 6-phosphogluconate dehydrogenase genes. Genetics and Molecular Biology 24:161-167.

Grivet, L, JC Glaszmann, M Vincentz, F da Silva and P Arruda. 2003. ESTs as a source for sequence polymorphism discovery in sugarcane: example of the Adh genes. Theoretical and Applied Genetics 106:190-197.

Harries, HC. 1978. The evolution, dissemination and classification of *Cocos nucifer*a L. The Botanical Review 44:266-319.

Harries, HC. 2002. The 'Niu' Indies: Long lost 'home' of the coconut palm. Palms 46:97-100.

Hartana, A, H Novarianto and D Asmono. 1993. Analisis keragaman dan pewarisan pola pita isozim tanaman kelapa. Jurnal Matematika & Sains 1:63-76.

Jay, M, R Bourdeix, F Potier and C Sanlaville. 1988. Premiers résultats de l'étude des polyphénols foliaires du cocotier. Oléagineux 44:151-161.

Lebrun, P, L Grivet and L Baudouin. 1998. The spread and domestication of the coconut palm in the light of RFLP markers. Dissémination et domestication du cocotier à la lumière des marqueurs RFLP. Plantation, Recherche et développement 5:233-245.

Lebrun, P, YP N'Cho, M Seguin, L Grivet and L Baudouin. 1998. Genetic diversity in coconut (*Cocos nucifera* L.) revealed by restriction fragment length polymorphism (RFLP) markers. Euphytica 101:103-108.

Lebrun, P, YP N'Cho, R Bourdeix and L Baudouin. 1999. Le cocotier. Pp. 219-239. *In:* JC Glaszmann (ed). Diversité génétique des plantes tropicales cultivées. CIRAD, Montpellier, France.

Lebrun, P, L Grivet and L Baudouin. 1999. Use of RFLP markers to study the diversity of the coconut palm. Pp. 73-89. *In:* C Oropeza, JL Verdeil, GR Ashburner, R Cardena, JM Santamaria (eds). Current advances in coconut biotechnology. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Lebrun, P, L Baudouin, R Bourdeix, JL Konan, JH Barker, C Aldam, A

Herran and E Ritter. 2001. Construction of a linkage map of the Rennell Island Tall coconut type (*Cocos nucifera* L.) and QTL analysis for yield characters. Genome 44:962-70.

Meerow, AW, RJ Wisser, SJ Brown, DN Kuhn, RJ Schnell and TK Broschat. 2003. Analysis of genetic diversity and population structure within Florida coconut (*Cocos nucifera* L.) germplasm using microsatellite DNA, with special emphasis on the Fiji Dwarf cultivar. Theoretical and Applied Genetics 106:715-726.

N'Cho, YP, A Sangaré, R Bourdeix, F Bonnot and L Baudouin. 1993. Evaluation de quelques écotypes de cocotier par une approche biométrique.1. Etude des populations de Grands. Oléagineux 48(3): 121-132.

Perera, L. 2002. Chloroplast DNA variation in coconut is opposite to its nuclear DNA variation. CORD. XVIII: 56-72.

Perera, L, JR Russell, J Provan, JW McNicol and W Powell. 1998. Evaluating genetic relationships between indigenous coconut (*Cocos nucifera* L.) accessions from Sri Lanka by means of AFLP profiling. Theoretical and Applied Genetics 96:545-550.

Perera, L, JR Russell, J Provan and W Powell. 1999. Identification and characterisation of microsatellite loci in coconut (*Cocos nucifera* L.) and the analysis of coconut populations in Sri Lanka. Molecular Ecology 8:344-346.

Perera, L, JR Russell, J Provan and W Powell. 2000. Use of microsatellite DNA markers to investigate the level of genetic diversity and population genetic structure of coconut (*Cocos nucifera* L.). Genome 43:15-21.

Perera, L, JR Russell, J Provan and W Powell. 2001. Levels and distribution of genetic diversity of coconut (*Cocos nucifera* L., var. Typica form typica) from Sri Lanka assessed by microsatellite markers. Euphytica 122:381-389.

Rivera, R, KJ Edwards, JHA Barker, GM Arnold, G Ayad, T Hodgkin and A Karp. 1999. Isolation and characterisation of polymorphic microsatellites in *Cocos nucifera* L. Genome 42:668-675.

Rohde, M, A Kullaya, J Rodriguez and E Ritter. 1995. Genome analysis of *Cocos nucifera* L. by PCR amplification of spacer sequences separating a subset of *copia*-lide *Eco*RI repetitive elements. Journal of Genetics and Breeding 49:179-186.

Schmalzing, D, A Belenky, MA Novotny, L Koutny, O Salas-Solano, S El-Difrawy, A. Adourian and P Matsudaira. 2000. Microchip electrophoresis: A method for high-speed SNP detection. Nucleic Acids Research 28:1-6.

Teulat, B, C Aldam, R Thehin, P Lebrun, JHA Barker, GM Arnold, A

Karp, L Baudouin and F Rognon. 2000. An analysis of genetic diversity in coconut (*Cocos nucifera*) populations from across the geographic range using sequence-tagged microsatellites (SSRs) and AFLPs. Theoretical and Applied Genetics 100:764-771.

Wadt, LHO, NS Sakiyama, MG Pereira, EA Tupinamba, FE Ribeiro and WM Aragao. 1999. RAPD markers in the genetic diversity study of the coconut palm. Pp. 89-97. *In:* C Oropeza, JL Verdeil, GR Ashburner, R Cardena and JM Santamaria (eds). Current advances in coconut biotechnology. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Zizumbo Villarreal, D. 1996. History of coconut (*Cocos nucifera* L.) in Mexico: 1539-1810. Genetic Resources and Crop Evolution 43:505-515.

Zizumbo-Villareal , D and HJ Quero. 1998. Re-evaluation of early observations on coconut in the new world. Economic Botany 52:68-77.

Zizumbo-Villarreal, D and P Colunga-García Marín. 2001. Morpho-physiological variation and phenotypicplasticity in Mexican populations of coconut (*Cocos nucifera* L.). Genetic Resources and Crop Evolution 48: 547-554.

Zizumbo-Villarreal, D and D Piñero. 1998. Pattern of morphological variation and diversity of *Cocos nucifera* (Arecaceae) in Mexico. American Journal of Botany 85:855-865.

Annex 1. A provisional classification of coconut cultivars

This classification is primarily based on molecular data (microsatellites). It also takes into account morphological and geographical criteria.

- There are two groups of order 1. These groups represent the major branches of coconut palm evolution, corresponding to two distinct centres of differentiation.
- In all, there are 10 groups of order 2. They represent the major divisions within the main two groups. Their name refers to the region from where the group is supposedly originated (but not necessarily the place where samples were actually taken). In the Indo-Atlantic group, cultivars are classified according to the rate of introgression by Pacific genes rather than on geographic criteria.
- There are 17 groups of order 3. They represent a more subtle division of the diversity of the species on a regional scale. This level of classification may be amended by more in-depth studies.

The molecular group is described by a code comprising three characters: capital letter for the first level, a digit for the second level and a lower case letter for the third level. (e.g. *A1a* for the southeast Asian Dwarfs, Malayan type, which belong to the Pacific group). The resulting groups are summarized in the following table. Examples of use of this classification can be found using the CGRD software.

A	**Pacific Group**		
	A1	**Southeast Asian Dwarfs**	
		A1a	*Malayan type*
		A1b	*Philippine type*
	A2	**Pacific Dwarfs (and semi-Talls)**	
	A3	**Southeast Asian Tall**	
		A3a	*Continental type*
		A3b	*Indonesian type*
		A3c	*Philippine type*
	A4	**Melanesia**	
		A4a	*North New Guinea type*
		A4b	*South New Guinea type*
		A4c	*Insular PNG type*
		A4d	*Markham Valley type*
		A4e	*Vanuatu Type*
	A5	**Micronesia**	
	A6	**Polynesia**	
	A7	**Panama (and Peru)**	
B	**Indo-Atlantic group**		
	B1	**Introgression absent to very low**	
	B2	**Low introgression rate**	
	B3	**Moderate to high introgression rate**	