Sequence tagged site markers to draft the genomic structure of the banana chloroplast

FC Baurens
JL Noyer
C Lanaud
PJL Lagoda
Cirad-Gerdat
Agetrop Laboratory
BP 5035
34032 Montpellier cedex 01
France

Sequence tagged site markers to draft the genomic structure of the banana chloroplast.

ABSTRACT

INTRODUCTION. Socio-economically, plantain and banana improvement to obtain resistant varieties is the most important topic for breeders. Knowledge of genitor genetic characteristics is required to manage rationally molecular marker assisted selection. Mapping the banana genome is mandatory to molecular breeding strategies. In higher plants, genetic information is located in chloroplast, mitochondria and nuclear genomes. Due to its relatively low complexity, the chloroplast genome is readily used in molecular biology. Concerning banana, RFLP reveal several chloroplastic groups in wild and cultivated types, but previous studies did not allow for fine structural information on the chloroplast chromosome.The present paper highlights the general structure of a *Musa* chloroplast. **MATERIALS AND METHODS.** The structure of the banana chloroplast genome has been explored by choosing 8 homologous sequence tagged site (STS) probes and 5 'universal' primer pairs. **RESULTS AND DISCUSSION.** Evidence of the presence of two regions, inverted repeats, added to sequence homologies enables the banana chloroplast chromosome size to be estimated at about 150 kb. New data suggest that the banana chloroplast, although close, is not strongly related to grasses. Analysis of 5' non-coding sequences of chloroplastic genes and codon usage indicates that differential transcription may occur in banana and rice. **CONCLUSION.** Overall, a collection of 11 RFLP banana probes uniformly covering the chloroplast genome (14 loci) are available.

KEYWORDS
Musa, DNA, chloroplasts, genetic maps, genetic markers.

Utilisation de sites étiquetés par leur séquence pour esquisser la structure génomique du chloroplaste de banane.

RÉSUMÉ

INTRODUCTION. Socioéconomiquement parlant, l'obtention de résistances est l'objectif principal des sélectionneurs de plantains et bananes. Les caractéristiques génétiques de géniteurs doivent être connues pour gérer rationnellement la sélection assistée par des marqueurs moléculaires. Cartographier le génome de la banane est obligatoire pour définir des stratégies de sélection moléculaire. Chez les plantes supérieures, l'information génétique est localisée dans les génomes chloroplastiques, mitochondriaux et nucléaires. Le génome chloroplastique (Gc), le moins complexe, est couramment utilisé en biologie moléculaire. Pour la banane, l'utilisation des techniques de RFLP a révélé plusieurs groupes chloroplastiques dans les type sauvages et cultivés, mais les études antérieures ont donné des informations insuffisantes sur la structure du chromosome chloroplastique. Ce document expose les principaux éléments connus de la structure générale du chloroplaste des *Musa*. **MATÉRIEL ET MÉTHODES.** La structure du Gc de la banane a été exploré à partir de huit sondes correspondant à des sites étiquetés par leur séquence (STS) homologues et de cinq paires d'amorces « universelles ». **RÉSULTATS ET DISCUSSION.** Du fait de deux régions, répétitions inversées (RI), détectées et des homologies de séquence, la taille du chromosome chloroplastique de la banane serait de 150 kb. De nouvelles données suggèrent que le chloroplaste de la banane, bien que proche, n'est pas fortement apparenté aux plantes herbacées. L'analyse de séquences non codantes en 5' de gènes chloroplastiques et l'usage de codons indiquent qu'une transcription différenciée a pu subvenir chez la banane et le riz. **CONCLUSION.** Globalement, une collection de 11 sondes RFLP de banane, couvrant uniformément le génome chloroplastique (14 loci), est utilisable.

MOTS CLÉS
Musa, ADN, chloroplaste, carte génétique, marqueur génétique.

Received 28 January 1997
Accepted 31 July 1997

Fruits, 1997, vol 52, p 247-259

Resumen Español, p 259

introduction

Modern crop improvement is based on molecular marker assisted selection and introgression of agronomic traits of interest, such as pest resistance or quality. Many crops are being investigated on the molecular level using ever improved marker systems. Tropical crops such as *Musa* are seldom, if ever, included in international genome analysis initiatives which would help unravel the mysteries of their sometimes complex genetic structures. However, these crops are of paramount importance to the world for socio-economical and ecological reasons.

Main dessert banana cultivars are triploid, highly sterile, parthenocarpic and clonally propagated. They are susceptible to several diseases that seriously threaten plantations. Thus, there is an urgent need to find new ways for developing cultivars that are tolerant to pests (eg, nematodes) and/or resistant to diseases such as Panama disease caused by *Fusarium oxysporum* f *cubense*, Sigatoka disease caused by *Mycosphaerella musicola* or Black Leaf Streak disease caused by *Mycosphaerella fijiensis* among others. These threats can be quantified as losses in revenue and food production, as well as limits to palliative chemical strategies (cost, pollution, emergence of novel pesticide resistances). Socio-economically, plantain and banana improvement to obtain resistant varieties is the most important topic for breeders. Knowledge of cultivars or parents genetic characteristics – polyploidy, cultivar sterility, parthenocarpy – is required to manage rationally molecular marker assisted selection.

Finally, biomolecular methodology is needed to analyze the 'environment-plant-pathogen' complex, obtaining essential preliminary data to establish protocols for banana improvement through genetic engineering.

Mapping the banana genome is mandatory to molecular breeding strategies. Restriction fragment length polymorphism (RFLP) marker techniques require a rather complex laboratory infrastructure, incompatible with most plantation sites overseas. For pragmatic reasons, and in order to foster exchange between 'biotechnology centers' in the North and 'biodiversity centers' in the South, the most rational approach to enhance traditional breeding knowledge by modern – ie, molecular – markers is to develop polymerase chain reaction (PCR) technology.

Now the complex nucleus is not the only molecular 'container' of genetic information. In higher plants, genetic information is located in three genomes. Chloroplast and mitochondria genomes are smaller than the nuclear genome. Cytoplasmic organelles possess their own independent genetic information. Due to its relatively low complexity, the chloroplast genome is readily used in molecular biology for plant studies and several higher plant chloroplast genomes have been completely sequenced: tobacco (SHINOZAKI et al, 1986), rice (HIRATZUKA et al, 1989), black pine (WAKASUGI et al, 1994) and maize (MAIER et al, 1995).

The size of higher plant circular chloroplast chromosomes varies between 120 000 bp and 217 000 bp (PALMER et al, 1987). Most have a specific genomic structure: the 16S and 23S ribosomic RNA (rRNA) genes are included in a larger sequence which is duplicated. These regions are called inverted repeats (IR) and are separated by two single copy regions. These regions are the large single copy (LSC) and the small single copy (SSC) region (SUGIURA, 1989).

Concerning banana, polymorphism of the chloroplast genome was successfully used with chloroplastic heterologous probes from other monocots (GAWEL and JARRET, 1991a, 1991b) and homologous probes from banana (CARREEL, 1994) in diversity studies. RFLP reveal several chloroplastic groups in wild type and cultivated bananas but previous studies did not allow for fine structural information on the chloroplast chromosome.

In addition, cytosolic genomes in banana are very interesting because of their differential heritability (FAURÉ et al, 1993a). If maternal transmittance of the chloroplast genome and paternal inheritance of the mitochondrial genome (FAURÉ et al, 1993a) can be generalized (CARREEL, 1994), this may be of importance for rebuilding the cross history of cultivated diploid and polyploid cultivars. Thus, better knowledge of the genomic organisation of the chloroplast

genome is useful for further investigation of cultivars. The present paper gives a highlight into the general structure of a *Musa* chloroplast. One of the objectives of the present investigation was to identify homologous probes for *Musa* choroplast DNA (cpDNA) allowing sequence data acquisition as a source for potentially-useful PCR markers in a molecular breeding strategy. The present paper is also an attempt to harmonize available data in order to construct a sound foundation for further investigations.

materials and methods

plant material and DNA extraction

Plant material was obtained from the French West Indies's (Guadeloupe) collection: *Musa acuminata* = *Musa acuminata* spp *banksii* Madang; *Musa schizocarpa* = *Musa schizocarpa* spp *schizocarpa*; *Musa balbisiana* = *Musa balbisiana* Type IV clone Butuhan; Grande Naine, the cultivated triploid clone Cavendish.Total DNA was extracted by the modified cethylmethyl ammonium bromide (CTAB) method (GAWEL and JARRET, 1991a; FAURÉ et al, 1993a).

RFLP analysis

RFLP analyses were performed according to FAURÉ et al (1993b) except for migration (120 v / 4 h) and DNA quantities which were fixed at 1.5 µg per lane. *Eco*RI, *Hae*III and *Sau*3A1 restriction enzymes were used for RFLP analyses.

library construction and screening

One µg of *Eco*RI and *Sau*3A1 restricted DNA used for RFLP analysis was purified using a Qiaquick® spin column (Qiagen™). DNA was ligated into appropriate dephosphorylated pUC19 (*Eco*RI or *Bam*HI for *Sau*3A1 restriction). Ligations were performed with 100 ng of vector and 1 µg of restricted DNA, in a final volume of 25 µl, using 2 U of T4 DNA ligase (BRL) according to the supplier's specifications. Ligation products were put onto a Qiaquick® spin column for purification and recovered in 50 µl of water before transformation. The DH51 *E coli* strain was transformed using 2 µl aliquots of ligation mix and 100 µl of competent cells (HANAHAN, 1987). Blue/white selection using ß-galactosidase complementation was used on bacteria grown overnight. White colonies were individualized, and plasmids were recovered by plasmid DNA minipreparation (SAMBROOK et al, 1989). Inserts were recovered using the appropriate restriction endonucleases *Eco*RI or *Eco*RI/*Pst*I, with migration on 0.8% agarose gel, and screened by blotting onto a Hybond N+ membrane and hybridized with heterologous *Lotus* sp cpDNA. Positive clones were sequenced using the dideoxy dye terminator method.

sequence comparisons

Sequence data were compared to the EMBL and Genbank (gb) databases content. A preliminary run by the BLAST program (ALTSCHUL et al, 1990) was made using standard parameters and confirmed by the FASTA (PEARSON and LIPMAN, 1988) program in the cytosolic genome subset of gb using standard parameters.

primer design

Primer design was performed using the Oligo™ program.

PCR amplifications

PCR was performed in a volume of 50 µl in the presence of 0.2 M of each primer, 200 mM of dNTP (Pharmacia) in a buffer containing: 67 mM Tris-HCl, pH 8.0, 16.6 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 1U of Biotaq® (Eurobio). Each reaction was overlaid with one drop of mineral oil to prevent variations in reaction volume. PCR was computed on a PTC 100 thermocycler (MJ Research) using the following amplification program: 94 °C for 4 min, 35 x [94 °C for 30 s; anealing temperature T_a for 30 s; 72 °C for 45 s], 72 °C for 4 min (table I).

Half of the reaction volume was loaded onto a 1% agarose gel, separated by electrophoresis and visualized with ethidium bromide.

Table I
Name, sequence and amplification product size of the primer pairs.
AGCH 07 and AGCH 09 were defined on the completely sequenced rice chloroplast (HANAHAN, 1983), AGCH 08 was defined in the pMaCIR 80 008 probe, other primer pairs derive from cpDNA conserved non-coding regions (DEMESURE et al, 1995). The AGCH 07 / AGCH 08 primer pair specifically amplifies the IRA/LSC junction and the AGCH 08 / AGCH 09 primer pair the LSC/IRB junction. Sizes were assayed on 12 banana wild and cultivated clones from three species: *Musa acuminata*, *Musa balbisiana* and *Musa schizocarpa*.
Ta, annealing temperature.

Name	*Sequence*	*Name*	*Sequence*	T_a (°C)	*Size (kb)*
AGCH 07	TAACGCCAACGAATCTCATC	AGCH 08	GGACATAGAATGCCAATCTT	45	0.6
AGCH 08	GGACATAGAATGCCAATCTT	AGCH 09	GGAAACCAGTGAAAACGAAT	45	1.0
trnH	ACGGGAATTGAACCCGCGCA	trnK	CCGAGTAGTTCCGGGTACGA	62	2.3 ± 0.2
trnD	ACCAATTGAACTACAATCCC	trnT	CTACCACTGAGTTAAAAGGG	55	1.5 ± 0.1
psbC	GGTCGTGACCAAGAAACCAC	trnS	GGTTCGAATCCCTCTCTCTC	57	1.6
trnS	GAGAGAGAGGGATTCGAACC	trnM	CATAACCTTGAGGTCACGGG	62	1.0 ± 0.2
trnM	TCCTTTCATACGGCGGCAGT	rbcL	GCTTTAGTCTCTGTTTGTGG	59	3.0

codon preference analysis

Putative coding sequences identified by sequence comparison were translated using the standard genetic code and aligned against equivalent rice genes. Strictly homologous amino acid positions were counted and DNA correspondance was determined. Distribution of permitted nucleotides was assayed using the χ^2 test at $P = 0.001$ against equiprobable distributions of purines vs pyrimidines and (A / U) vs (C / G). Preference on the first base of the arginine (Arg), the leucine (Leu) and the serine (Ser) codon was separately determined. Methionine, tryptophan and stop codons were excluded from the study.

phylogenetic tree building

Data from the large subunit of rubisco (RBCL) genes were treated using the Clustal algorithm (THOMPSON et al, 1994) for minimal tree length with the exhaustive search option and default parameters.

results and discussion

homology and genomic structure

In order to determine the homology of the banana chloroplast genome with other plants, *Lotus* sp chloroplastic total DNA was hybridized on restricted total banana DNA (figure 1). A strong hybridization signal was observed showing that most chloroplastic sequences are highly conserved between *Musa* and *Lotus*. As the *Lotus* species is phylogenetically very distant from banana, the global sequence of the chloroplast genome seems to be conserved within plants. Analysis of the hybridisation patterns shows that band intensities can be classified into two groups: weak signals and strong signals. Weak signals can be assigned to single copy chloroplastic sequences. Strong signals could be a result of duplicated quantities of specific size fragments. Due to the limitations of separation by agarose gel electrophoresis, co-migration of fragments exhibiting slightly different sizes at the same place could also produce a strong hybridization signal. It has been shown that several plant species have a chloroplastic genome organization carrying two highly-conserved inverted repeats (SUGIURA, 1989). Most of the strong bands in figure 1 could be due to restriction fragments from chloroplast IR. Based on the relative intensities of hybrization signals, the presence of two IRs in the banana chloroplast genome could not be completely evidenced.

For further investigations of the banana chloroplast genome, probes from banana libraries were identified and completely sequenced. Sequences are available in the EMBL data bank under the accession numbers listed in table II. The pMaCIR 81 008 probe

(table II) was partially sequenced and sequence comparison shows that it is related to the chloroplast IRs in rice, maize and tobacco. Sequence comparison of the partially sequenced pMaCIR 81 008 probe shows that the same genomic organization occurs in rice (EMBL an X15 901) and maize (EMBL an X86 563) and is slightly degenerated when compared with the tobacco (EMBL an Z00 044) chloroplast (data not shown). In rice, and maize, the homologous segments are respectively 2 304 base pairs (bp) and 2 307 bp long. This is consistent with the length of the banana probe, which is approximatively 2 400 bp. Comparison with the rice, chloroplast chromosome shows that the part of the probe between nucleotide positions 41 and 219 is strongly homologous to a fragment located between positions 55 806 and 55 966 of the chloroplast genome of rice, near the *rbcL* gene (figure 2). As demonstrated by KATAYAMA and OGIHARA (1993), a partial copy of the region of the *rpl23* gene, usually located inside the IR and homologous to the pMaCIR 81 008 probe, is duplicated in rice and other grasses near the *rbcL* gene. The organization of the pMaCIR 81 008 probe suggests that, in banana, no rearrangements occur in the *rpl23* gene located in the IRs, as the size of the probe is globally conserved throughout plant species and high homology is found at the sequence level. In order to determine whether translocation of a fragment occurs in the banana chloroplast genome, the pMaCIR 81 008 probe was hybridized onto total restricted banana DNA. As this probe reveals only one or two bands, depending of the restriction enzyme used (figure 3), the absence of translocation of part of the *rpl23* gene near the *rbcL* gene or in other locations is demonstrated. Similar results have previously been obtained in most of the monocots, except grasses.

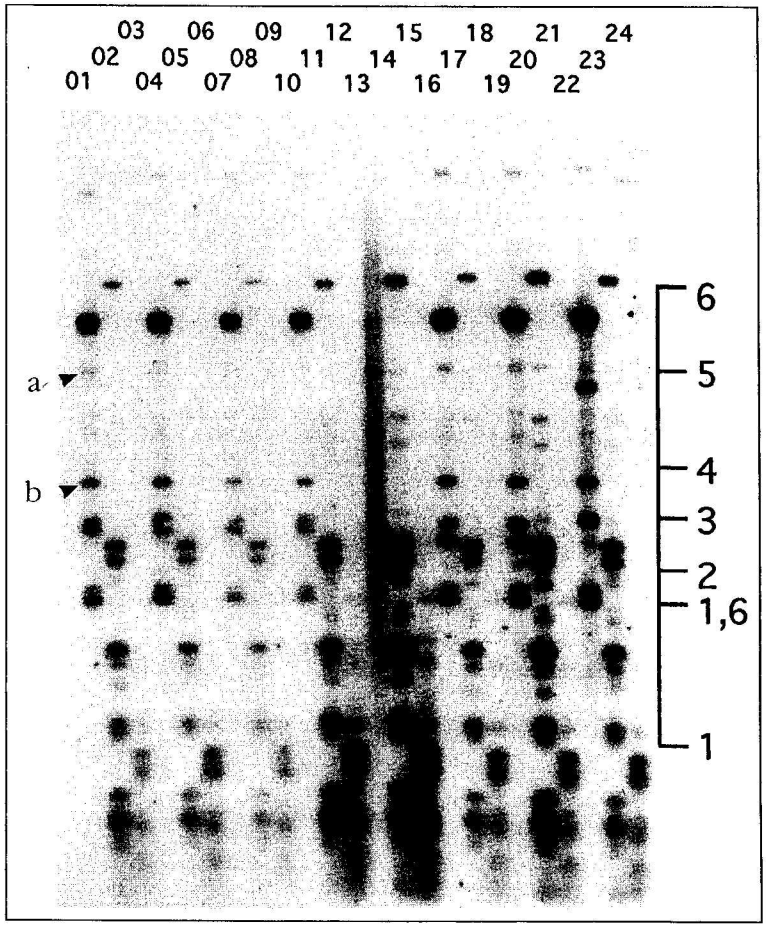

Figure 1
RFLP hybridization of Lotus *sp chloroplast DNA on banana: overnight exposure of total* Lotus *sp chloroplast DNA probe on total endonuclease digested DNA of several representatives of the* Eumusa *section.*

Lanes 1–3: M schizocarpa *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 4–6:* M acuminata *ssp* banksii *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 7–9:* M acuminata *ssp* zebrina *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 10–12:* M acuminata *ssp* malaccensis *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 13–15:* M acuminata *ssp* burmannicoides *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 16–18:* M balbisiana *'Pisang Batu' DNA restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 19–21:* M balbisiana *'Honduras' DNA restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 22–24:* Musa balbisiana *'Butuhan' DNA restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Sizes in kb are indicated on the right, arrowheads indicate exemples of weak hybridisation signal (a) and strong ones (b).*

In order to determine the number of IRs in the banana chloroplast, sequence data were used to design a primer pair able to amplify specifically the junctions between the LSC and the two IRs. Oligonucleotide primer AGCH 08 (table I) was defined in the pMaCIR 81 008 probe which was located in the IR near the junctions. PCR amplifications using AGCH 08 and a primer designed in the pMaCIR 20 016 (table II) banana probe located in the LSC near the junction with IR, failed to detect any products. This was probably due to the size of the amplification region (approximately 5 kb in rice). As no other homologous probe could be located close to the junction, and based on the hypothesis that the junction regions of rice and banana are conserved, two primers

Table II
Characterization of several banana chloroplastic probes.

Name	*EMBL an*	*Size (pb)*	*%*	*homology with rice and characteristics*
pMaCIR 10 003	X99492	109	95	*pbsG*, partial coding sequence
pMaCIR 10 008	X99493	190	77	Non-coding intronic region
pMaCIR 10 113	X99494	231	80	Inf A non coding 5' region + partial coding sequence
pMaCIR 20 016	X99495	512	80	*rps8* 5' non coding region + partial coding sequence
pMaCIR 70 036	X99498	204	82	*rpoA*, partial coding sequence
pMaCIR 81 008	X99499 /X99500	2400	85	Inverted repeat
pMaCIR 21 108*	X99496	958	90*	Putative *ORF 1708*, partial coding sequence
pMaCIR 31 315	X99497	994	94	*rbcL*, partial coding sequence

Determination of probe size was based on sequence data except for pMaCIR 81 008 which was only partially sequenced. Percentage of homology and nature of the probe were determined using nucleic-acid sequence alignment data (see *Materials and methods*). The greatest homology percentage was found with rice excepted for the probe pMaCIR 21 108 *, which is absent in rice chloroplast genome (homology was determined on tobacco in this particular case), and for pMaCIR 31 315 which is more homologous to *Phoenix reclinata*. Sequences of all these probes are available in the EMBL database under the given accession numbers (EMBL an).

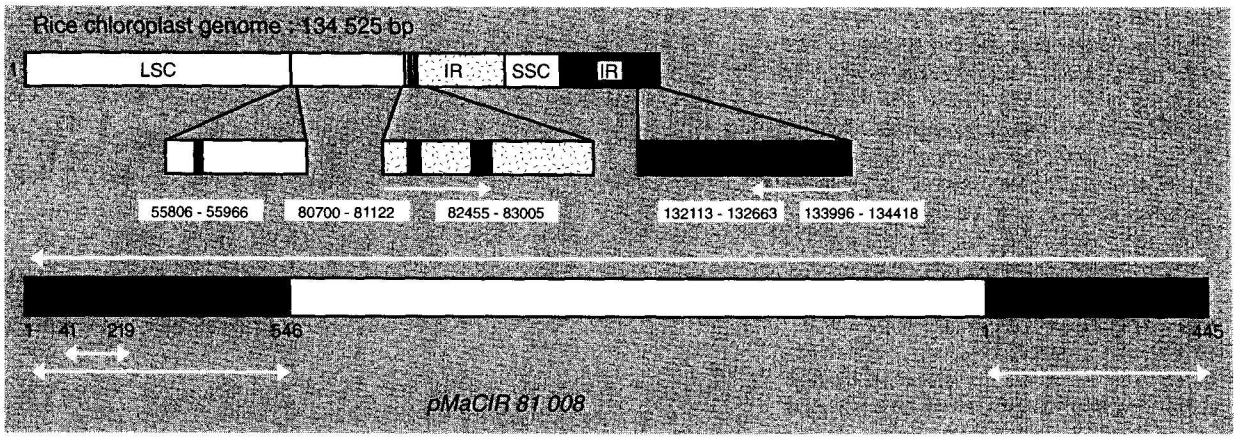

Figure 2
Position of the pMaCIR 81 008 probe on rice linear chloroplast map.
Genomic organization of pMaCIR 81 008 was determined using sequence comparison with total rice chloroplast (see Materials and methods*). Open squares correspond to single copy regions, grey squares to IRs. Arrowheads indicate the sequenced part of the banana probe and the correspondance to rice chloroplast (dark squares). Numbers indicate nucleotide rank.*

(AGCH 07 and AGCH 09; table I) were designed in rice chloroplatic genes located close to both junctions. The expected products for AGCH 07/08 IRA/LSC junction amplification and AGCH 08/09 LSC/IRB junction amplification were 588 bp and 845 bp respectively in rice. PCR amplifications using these primer pairs and total banana DNA are shown in figure 4. The two expected amplification products are present, demonstrating the presence of two IRs in the banana chloroplast.

Considering a chloroplast with two IRs, the hybridization signals of IR fragments are included in the strong band subset of the RFLP pattern. Using two different restriction enzyme patterns (*Eco*RI and *Hae*III), the size of the inverted repeat can be estimated at 26 ± 3 kb. This value corresponds to a minimum size of the IR in banana.

The size of the chloroplast genome in higher plants is strongly related to the size of the IR (FORCIOLI, 1995). Figure 5 shows the linear correlation (r = 0.941) between the size of the IR and the total size of the chloroplast genome. Using the regression coefficient obtained from these data, the size of the banana chloroplast genome can be estimated at between 146 and 154 kb, with an average size of 150 kb.

The probe pMaCIR 21 108, was found to be strongly homologous to a tobacco putative gene open reading frame (ORF) 1 708 (SHINOZAKI et al, 1986) renamed ORF 2280

(KATAYAM and OGIHARA, 1993) located near the *rpl23* gene. The IR of grass chloroplast genomes carries a deletion of this ORF (KATAYAM and OGIHARA, 1993). Based on its hybridization, KATAYAMA and OGIHARA demonstrated that grass chloroplast genomes were smaller than other known monocots. In the same way, the size of the banana IR, found to be larger than those of rice and maize, can be explained by the presence of undeleted sequences related to other monocots and dicots, including ORF 1 708. In the banana chloroplast, the presence of a sequence homologous to ORF 1 708 supports the suggestion that banana is more related to onion and lily than to the grasses. Based on chloroplast structure, the differentiation between the *Zingiberidae* order and the *Commelidae* order (including the grass family) occurred prior to 'the drastic structural alteration of the chloroplast genome of grass' (KATAYAM and OGIHARA, 1993).

As the banana chloroplast genome seems to be well-conserved, five of the chloroplastic subset of 'universal primers' (DEMESURE et al, 1995; table I) were used on several genotypes (example for the trnD/trnT primer pair is given in figure 6). These primers amplify intronic regions of genes located in the LSC region of higher plant chloroplast genomes. *Theobroma cacao* (cocoa) DNA was used as an external tropical crop reference and human DNA as negative control (figure 6). Amplification products exhibit a length variation throughout banana clones. These variations are very interesting because they prove the possibility of using PCR technology to reveal inter-specific chloroplastic polymorphism in banana in complement to the RFLP technique. Cocoa DNA amplification product size averages are close to those of banana, and those of rice are the smallest (data not shown). These data on the chloroplast genome LSC region and the comparison with rice (total genome size of 134 525 pb, HIRATZUGA et al, 1989) and cocoa (total genome size of at least 110 kb; YEOH et al, 1990) are in accordance with those previously obtained on the IR region and global size estimation.

The presence of amplification products with primers defined in rice genes (figure 4) demonstrates that the same genomic organization of the IR junctions is conserved between rice and banana. Moreover, RFLP analysis of the previously-described pMaCIR 1 046 and *pet A* probes (CARREEL, 1994; figure 9) shows that the size of the banana fragment was approximately 5 kb, as in tobacco. These data suggest that the genomic organization of the banana chloroplasts in the junction regions and in the *pet A* region are very similar to other higher plant chloroplasts.

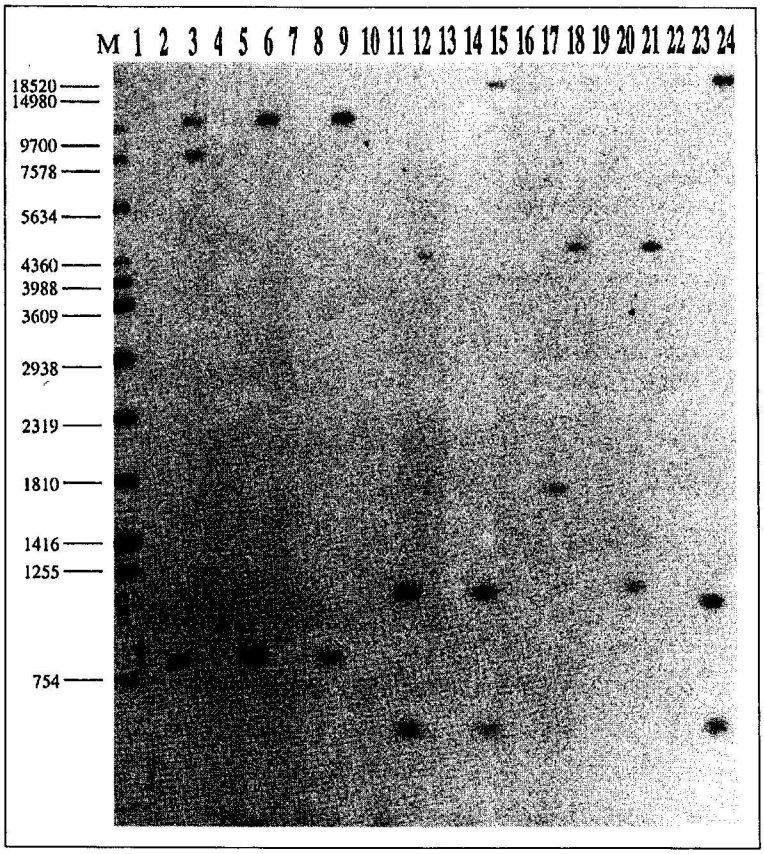

Figure 3
Overnight exposure of pMaCIR 81 008 probe on total endonuclease digested DNA of several representatives of the Eumusa *section.*
Sizes in base pairs are indicated on the right (Lane M). Lanes 1–3: M schizocarpa *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 4–6:* M acuminata *ssp* banksii *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 7–9:* M acuminata *ssp* zebrina *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 10–12:* M acuminata *ssp* malaccensis *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 13–15:* M acuminata *ssp* burmannicoides *restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 16–18:* M balbisiana *'Pisang Batu' DNA restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 19–21:* M balbisiana *'Honduras' DNA restricted with* EcoRI, HaeIII *and* Sau3A *from left to right. Lanes 22–24:* Musa balbisiana *'Butuhan' DNA restricted with* EcoRI, HaeIII *and* Sau3A *from left to right.*

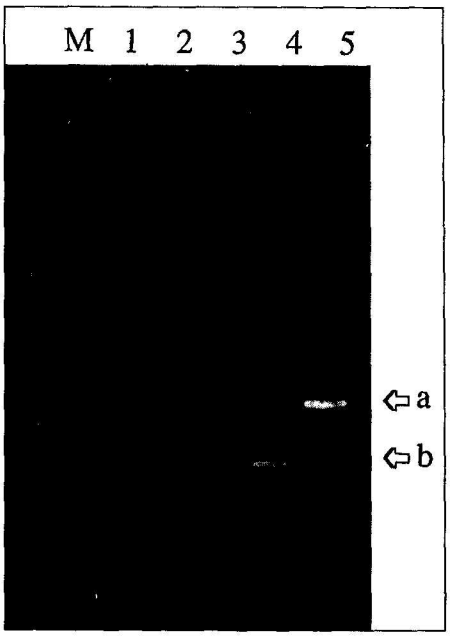

Figure 4
PCR amplification of putative inverted repeats junctions of banana chloroplast genome. Lane 1 to 5: PCR amplification products on banana total DNA using primers AGCH 07, AGCH 08, AGCH 09, primer pairs AGCH 07 / AGCH 08 and AGCH 08 / AGCH 09, respectively. Arrowheads indicate amplification products of IRB/LSC junction (a) and LSC/IRA junction (b). Lane M corresponds to a 100 bp ladder.

coding sequences

Sequences from six chloroplastic probes (coding sequences in table II) exhibit a strong homology with other plant (especially with rice) chloroplastic genes: the large subunit of rubisco (*rbcL* gene), initiation factor IF1 (*infA*), ribosomal protein S8 (*rps8*), alpha subunit of RNA polymerase (*rpoA*), NADH plastoquinone oxido reductase subunit (*psbG*). The pMaCIR 21 108 exhibits a strong homology with a putative tobacco ORF: ORF 1 708 (SHINOZAKI et al, 1986) or ORF 2 280 (KATAYAM and OGIHARA, 1993).

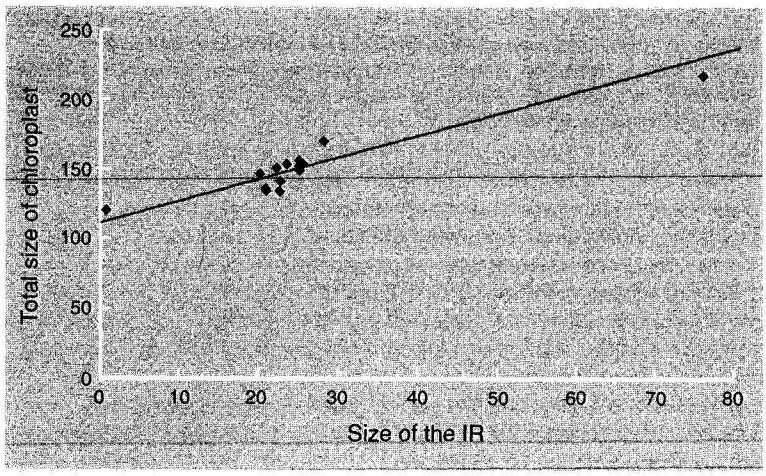

Figure 5
Linear correlation between global size of the chloroplast genome and inverted repeats. Data on IRs and global sizes of chloroplasts were plotted as a linear regression. Parameters and bestfit equation were [total chloroplast size] = 1.3657 × [IR size] + 114.53, with r = 0.941.

Probe pMaCIR 31 315 contains 991 bp homologous to the *rbcL* gene (94% nucleic acid homology with *Phoenix reclinata*, table II).

As rubisco seems to be a good molecular clock (GIELLY and TABERLET, 1994), this partial coding sequence was translated, and 270 amino acids of the same part of the putative protein were aligned with *rbcL* gene sequences of phylogenetically-related tropical species, and representatives of the grass section, with dicots as external references. Figure 7 shows the tree constructed from these data. As expected, the separation between monocots and dicots is readily observed. *Musa acuminata* is placed in the monocot part of the tree but is not strongly related to the grass species. The most related species to banana is *Phoenix reclinata* which is separated from another representative of the *Arecale* order. These data combined with those concerning the IR structure suggest that banana could be more distant from the grass section than expected by morphologically-based phylogenetic studies (CRONQUIST, 1988). Similar data were obtained in the *Zingiberale* order and in monocots (DUVALL et al, 1993; SMITH et al, 1993) using partially PCR sequenced *rbcL* genes. Due to the PCR amplification, nucleotide data acquisition was less accurate than with our cloned fragment. This could explain why the already available banana partial *rbcL* gene sequence was not found to be the most homologous sequence in a database search.

The other partial coding sequences were translated using the standard genetic code producing putative polypeptide sequences. Those sequences were compared to protein databases. Strong homology occurs with protein sequences, confirming the high level of chloroplast genic sequence homology already found between banana and other plant species (GAWEL and JARRET, 1991a, 1991b; FAURÉ et al, 1993a). Interestingly, rice was found to be the most homologous concerning all these other protein sequences, except for ORF 1 708 which is absent in rice. The five partial gene sequences homologous to rice were translated using the standard genetic code, and the same parts of homologous genes from the rice chloroplast were compared for codon usage. Only the strictly homologous amino acid positions were checked. As previously described, prefe-

rences of the third base of the codon follow the rules of tRNA wobble (table III). Even if chloroplastic DNA sequences are well-conserved between rice and banana, based on sequence comparison data, codon usage differs between the two species. Arginine is coded by six potential codons. Neither in rice nor in banana, was preference on the first base detected. Interestingly, the third base of the arginine codon in banana appears to be preferentially A or U, while rice exhibits no preferences. This situation is reversed for the serine codons, where banana exhibits no preferences while rice prefers A or U. The last case concerns the lysine codon where the AAA codon is preferred to the AAG codon in rice, while banana exhibits an equiprobable distribution.

The codon usage preference was determined with strongly-expressed proteins in which it is well known that this preference reflects the relative abundance of tRNAs in the cell (WATSON et al, 1987). Furthermore, using genetic transformation with heterologous sequences, the differential preferences of codon usage between donor and host organisms may be responsible for inhibition of expression of the transfected sequences (SOLTES-RAK et al, 1995). Thus, determination of the preferential codon usage in chloroplasts is very important for banana genetic engineering, if plastid is the transfection target. As the chloroplast is partitioned from the nucleus in the cell, codon usage may differ and extrapolation to nuclear codon usage is risky. However, these data can serve as a starting point for further investigations concerning general codon usage in banana.

non-coding sequences

Probes pMaCIR 10 113 and pMaCIR 20 016 also contain the 5' regions of the genes *infA* and *rps8*, respectively. There is no strong relation between the two promoter regions on these probes. However, sequence comparison between the *rps8* promoter sequences of rice and banana shows that, interestingly, these sequences are homologous between the two species, especially for the putative TATA box and the ribosome binding sites (figure 8). Moreover, duplication of a 8 bp region in banana replaces a different sequence in rice, suggesting that

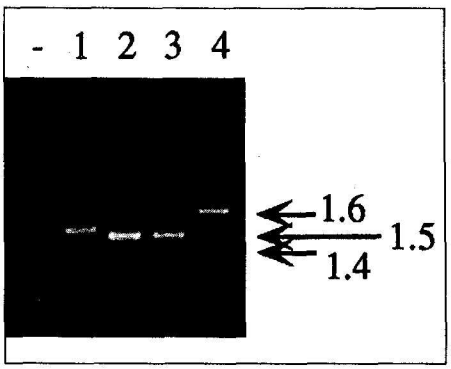

Figure 6
PCR amplification of the trnD/trnT region of banana chloroplast genome. Agarose gel electrophoresis of PCR amplification products using the trnD/trnT primer pair (table I). Lane -: human DNA as negative control. Lanes 1 to 3: banana DNA: Musa acuminata, Musa balbisiana, *AAA triploid cultivar Cavendish, respectively. Lane 4: cocoa DNA. Arrowheads indicate size of the products in kb.*

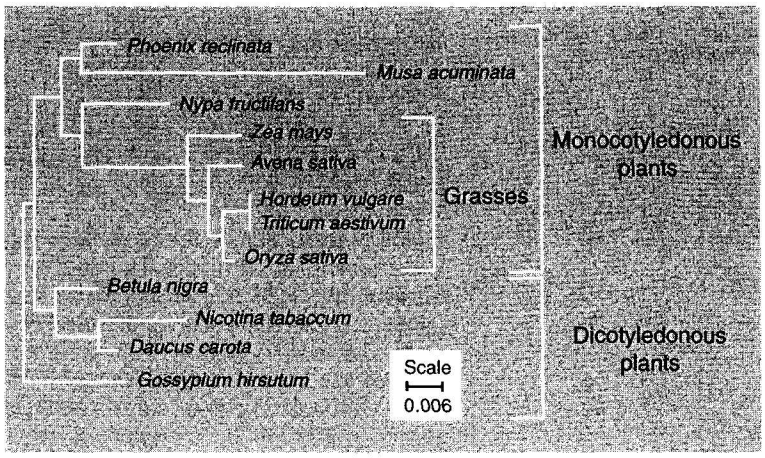

Figure 7
Place of Musa acuminata in a phylogeny of some plant species based on rubisco. 270 amino acids (No157 to 426) from the rice rubisco protein (accession number: X 15 901) were used as a base for clustal *alignment (see Materials and methods). Other plant species' rubisco, homologous to this fragment were used: Avena sativa (P48 684), Betula nigra (L01 889), Daucus carota (D44 566), Gossypium hirsutum (P14 958), Hordeum vulgare (P05 698), Musa acuminata (banana clone pMaCIR31 315, EMBL X 99 497), Nicotiana tabaccum (Z00 044), Nypa fruticans (P28 261), Phoenix reclinata (P28 262), Triticum aestivum (P11 383) and Zea mays (X 86 563). Scale: relative genetic distance,* clustal *standard parameters.*

the size of this region has to be conserved. It is possible that a transcription factor or a DNA binding protein involved in the transcription process might be implicated in these regions with different properties between the two species.

An AT rich sequence of approximatively 40 bp is present in the promoter region of banana and completely absent in rice, suggesting a different regulation process between the two species. It has been reported that AT rich insertions or deletions could contribute to RFLP polymorphisms in chloroplasts (ZURAWSKI et al, 1984; DALLY and SECOND, 1990; HOOGLANDER et al, 1993; KANNO et al, 1993). Moreover, recent advances in chloroplastic simple sequence repeat (SSR) analyses demonstrate that A or T repeats can be highly polymorphic at the inter- or intraspecific level in pine (CATO and RICHARDSON, 1996). Furthermore, the most polymorphic restriction enzyme site found in banana RFLP studies was *DraI* (AAATTT). This AT-rich

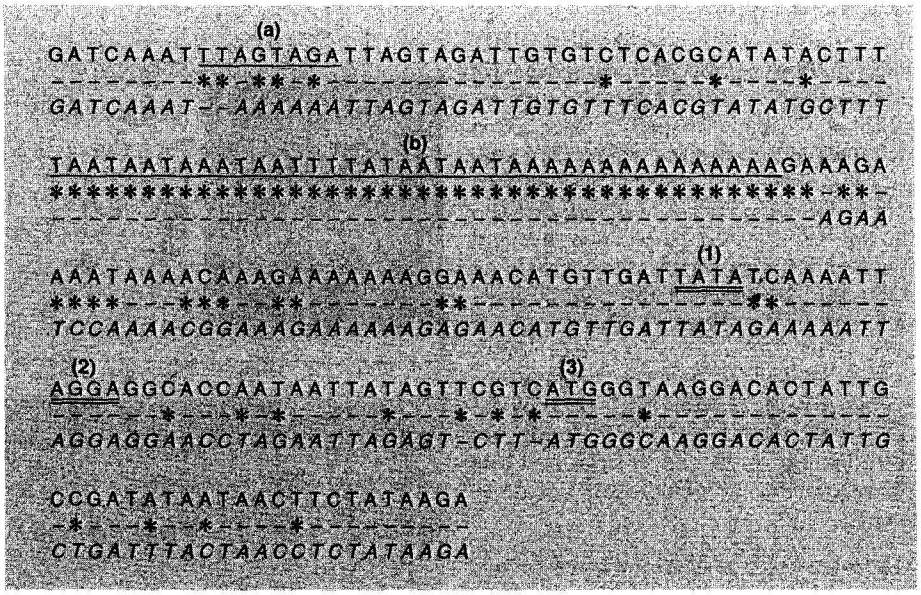

Figure 8
Nucleotide sequence alignment of the 5' region of the rps8 chloroplast gene between banana and rice. Stars indicate differences between the two sequences, dashes correspond to putative deletions in the rice sequence. Underlined regions correspond to banana specific parts of he 5' region: (a): duplicated 8 bp region; (b): AT rich region. Double lines correspond to putative promoting elements as TATAbox (1), ribosome binding site (2) and ATG start codon (3). Rice sequence is visualized in italics.

region presents homologies with transcriptional factors (eg, TFII A, B, D, E and F by browsing the TESS transfac database) suggesting different expression of this gene between rice and banana. Moreover, it was reported that some structural rearrangements in the chloroplast genome structures could also be involved in the transcriptional regulation of genes (KATAYAMA and OGIHARA, 1993).

The 5' region of the *infA* banana chloroplast gene is not homologous to another chloroplast *infA* regulatory region. Protein alignment suggests that *infA* from banana is shorter than that of rice and maize. The presence of the transcriptionnal factor IRF1 binding site (AAGTGA; FUJITA et al, 1987) in the 5' region of the banana gene, immediatly before the putative TATA box while this sequence is absent in rice, suggests a probably different regulation of *infA* expression.

Analysis of all these data suggests that chloroplastic non-coding sequences, and especially the 5' promoting regions of genes, could be very interesting, due to their potentially differential expressions throughout plant species. SSR data suggest that small variations in these hypothetically regulatory regions could appear at the intra-specific level (CATO and RICHARDSON, 1996).

Other non-coding chloroplast sequences (table II) were compared to EMBL data bank contents. Including 5' regions of genes, they have a mean homology of 70% with rice

Table III
Codon usage in banana chloroplast genes.

	Ala	*Arg*	*Asp*	*Glu*	*Gly*	*Ile*	*Leu**	*Leu*	*Lys*	*Ser*	*Thr*	*Val*
Banana	A or U	A or U	U	A	A or U	Py	U	A or U	nd	nd	Py	A or U
Rice	A or U	nd	U	A	A or U	Py	U	A or U	A	A or U	Py	A or U

Amino acids from five proteins were scored and codon usage was determined for all of the 20 possible amino acids. Preferential usage of the first base (Arg, Leu, Ser) and the third base was assayed. Distribution of purines vs pyrimidines and (A / U) vs (C / G) was determined against equiprobability using the χ^2 test. Only distributions differing significantly ($P = 0.001$) from equiprobability were scored and preferential nucleotide usage was determined. * the first nucleotide of the codon. nd, no preference detected.

intergenic sequences. This, added to the greater homology (90%) of coding sequences between rice and banana, suggests a strong global similarity at the sequence level, an interesting observation in view of the phylogenetic relationship between banana and the grass species.

conclusion

In total, 14 homologous probes uniformly cover the banana chloroplast genome (figure 9). Sequence data and structural analysis of those probes added to structural information on IRs and IR-LSC junctions indicate that, overall, the banana chloroplast genomic organization is similar to previously described monocot chloroplasts, but that local characteristic differences occur, as illustrated by comparison with rice and tobacco. The probes described can be used for RFLP studies. Sequence data are available for all of these probes and the primer pairs have been tested. PCR technology is an important tool that obviates the need to use radiolabelled elements and large quantities of DNA (BAKER and PALUMBI, 1994). Furthermore, crude DNA extracts can be used even with quantitative PCR methods (BAURENS et al, 1996). These advantages of PCR open the way for the use of interesting chloroplast markers in a molecular breeding approach to a banana improvement program.

acknowledgments

We want to thank P Gauthier for providing *Lotus* chloroplast DNA. E Jenczewski and F Bakry are gratefully acknowledged for critical reading of the manuscript. This work was supported by a Cirad grant.

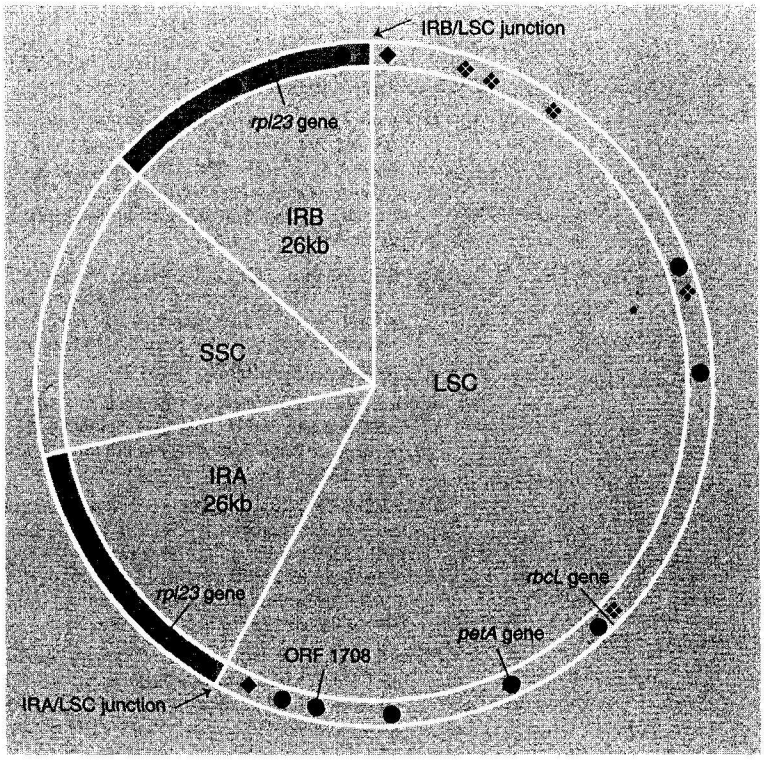

Figure 9
Putative banana chloroplast map, length 150 kb.
Localization of probes and primers on the putative banana chloroplast map: ● *indicate banana probes. From the IRB/LSC junction, pMaCIR 10 008, pMaCIR 10 003, pMaCIR 31 315, pMaCIR 1046, pMaCIR 70 036, pMaCIR10 113, pMaCIR 20 016, pMaCIR 81 008, pMaCIR21 108, pMaCIR21 108, pMaCIR 81 008, respectively (clockwise).* ◆ *indicate primers defined on rice chloroplast (sequences in table I).* ❖ *indicate primer pairs from the universal set of chloroplast primers (DEMESURE et al, 1995).*
Starting clockwise from the IRB/LSC junction, trnH-trnK, psbC-trnS, trnS-trnfM, trnD-trnT, trnM-rbcL, respectively (table I). Indicative positions of some important elements are indicated.

references

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *J Mol Biol* 215, 403-410

Baker CS, Palumbi SR (1994) Which whales are hunted? A molecular genetic approach to monitoring whaling. *Science* 265, 1538-1539

Baurens FC, Noyer JL, Lanaud C, Lagoda PJL (1996) Use of competitive PCR to assay copy number of repetitive elements in banana. *Mol Gen Genet* 253, 57-64

Carreel F (1994) Etude de la diversité génétique des bananiers (genre *Musa*) à l'aide des marqueurs RFLP. Paris, France, Cirad-Flhor, Ina Grignon, doctorat, 120 p

Cato SA, Richardson TE (1996) Inter-and intraspecific polymorphism at chloroplast SSR loci and the inheritance of plastids in *Pinus radiata* Don. *Theor Appl Genet* 92, in press

Cronquist A (1988) The evolution and classification of flowering plants. New York, USA, The New York Botanical Garden (ed), 350 p

Dally AM, Second G (1990) Chloroplast DNA diversity in wild and cultivated species of rice (genus *Oryza*, section *Oryza*). Cladistic-mutation and genetic-distance analysis. *Theor Appl Genet* 80, 209-222

Demesure B, Petit R J, Sodzi N (1995) A set of universal primers for amplification of non-coding regions of mitochondrial and chloroplast DNA in plants. *Mol Ecol* 4, 129- 131

Duvall MR, Clegg MT, Chase MW, Clark WD, Kress WJ, Hills HG, Eguiarte LE, Smith JF, Gaut BS, Zimmer EA, Learn GH (1993) Phylogenetic hypotheses for the monocotyledons constructed from *rbcL* sequences data. *Ann Missouri Bot Gard* 80, 607-619

Fauré S, Noyer JL, Carreel F, Horry JP, Bakry F, Lanaud C (1993a) Maternal inheritance of chloroplast genome and paternal inheritance of mitochondrial genome in bananas (*Musa acuminata*). *Curr Genet* 25, 265-269

Fauré S, Noyer J L, Horry J P, Bakry F, Lanaud C, González de León D (1993b) A molecular marker-based linkage map of diploid bananas (*Musa acuminata*). *Theor Appl Genet* 87, 517-526

Forcioli D (1995) Différenciation génétique comparée des informations nucléaires et cytoplasmiques dans les populations naturelles de *Beta maritima* : importance des flux géniques et conséquences sur la stérilité mâle. Paris, France, université Paris-VI, 190 p

Fujita T, Shibuya H, Hotta H, Yamanishi K, Taniguchi T (1987) Interferon-beta gene regulation: tandemly repeated sequences of a synthetic 6 bp oligomer function as a virus-inducible enhancer. *Cell* 49, 357-367

Gawel NJ, Jarret RL (1991a) Cytoplasmic genetic diversity in bananas and plantains. *Euphytica* 52, 19-23

Gawel NJ, Jarret RL (1991b) Chloroplast DNA restriction fragment length polymorphisms (RFLPs) in *Musa* species. *Theor Appl Genet* 81, 783-786

Gielly L, Taberlet P (1994) The use of chloroplast DNA to resolve plant phylogenies: non coding versus *rbcL* sequences. *Mol Biol Evol* 11(5), 769-777

Hanahan D (1983) Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol* 166, 557

Hiratzuka J, Shimada H, Whittier R, Ishibashi T, Sakamoto M, Mori M, Kondo C, Honji Y, Sun C-R, Meng B-Y, Li Y-Q, Kanno A, Nishizawa Y, Hirai A, Shinozaki K, Sugiura M (1989) The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intramolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of cereals. *Mol Gen Genet* 217, 185-194

Hooglander N, Lumaret R, Bos M (1993) Inter-intraspecific variation of chloroplast DNA of european *Plantago* spp. *Heredity* 70 (3), 322-334

Kanno A,Watanabe N, Nakamura I, Hirai A (1993) Variation in chloroplast DNA from rice (*Oryza sativa*) differences between deletions mediated by short direct-repeat sequences within a single species. *Theor Appl Genet* 86, 579-584

Katayama H, Ogihara Y (1993) Structural alteration of the chloroplast genome found in grasses are not common in monocots. *Curr Genet* 23, 160-165

Maier RM, Neckermann K, Igloi GL, Kossel H (1995) Complete sequence of the maize chloroplast genome, gene content, hotspot of divergence and fine tuning of genetic information by transcript editing. *J Mol Biol* 251, 614-628

Palmer JD, Nugent JM, Herbon LA (1987) Unsusual structure of geranium chloroplast DNA: a triple size inverted repeat, extensive gene duplications, multiple inversions, and two repeat families. *Proc Natl Acad Sci USA* 84, 769-773

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 85, 2444-2448

Shinozaki K, Ohme M, Tanaka M, Wagazugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-shinozaki K, Ohto C, Torazawa K, Meng BY, Sugita M, Deno H, Kamogashira T, Yamada K, Kusuda J, Takaiwa F, Kato A, Tohdoh N, Shimada H, Sugiura M (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J* 5, 2043-2049

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: a laboratory manual.* New York, USA, J Sambrook, EF Fritsch, T Maniatis (eds), Cold Spring Harbor Laboratory Press, 3 vol, 1 680 p

Smith JF, Kress WJ, Zimmer EA (1993) Phylogenetic analysis of the Zingiberales based on *rbcL* sequences. *Ann Missouri Bot Gard* 80, 620-630

Soltes-Rak E, Kusher DJ, Dudley Williams D, Coleman JR (1995) Factors regulating *cryIVB* expression in the cyanobacterium *Synechococcus* PCC 7942. *Mol Gen Genet* 246, 301-308.

Sugiura M (1989) The chloroplast chromosome in land plants. *Annu Revi Cell Biol* 5, 51-70

Thompson JD, Higgins DG, Gibson TJ, Clustal W (1994) Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680.

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M (1994) Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of black pine *Pinus thunbergii*. *Proc Natl Acad Sci USA* 91, 9794-9798

Watson JD, Hopkins NH, Roberts JW, Argetsinger Steitz J, Weiner AM (1987) Molecular biology of the gene. Menlo Park, CA, USA,The Benjamin/Cummings publishing Cie Inc (ed), 4th edition, 850 p

Yeoh HH, Chung DK, Fritz PJ (1990) *Theobroma cacao* chloroplast DNA: isolation, molecular cloning, and characterization. *Café Cacao Thé* XXXIV (3), 173-178

Zurawski G, Clegg MT, Brown AHD (1984) The nature of nucleotide sequence divergence between barley and maize cpDNA. *Genetics* 106, 735-749

Utilización de sitios etiquetados por su secuencia para esbozar la estructura genómica del cloroplasto de plátano.

RESUMEN

INTRODUCCIÓN. Desde el punto de vista socioeconómico, la obtención de resistencias es el principal objetivo de los seleccionadores de plátanos y bananos. Para controlar racionalmente la selección asistida por marcadores moleculares, hay que conocer las características genéticas de los genitores. Es obligatorio cartografiar el genoma del plátano para definir las estrategias de selección molecular. En las plantas superiores, la información genética está localizada en los genomas cloroplásticos, mitocondriales y nucleares. El genoma cloroplástico (Gc), que es el menos complejo, se utiliza corrientemente en biología molecular. En el caso del plátano, la utilización de técnicas de RFLP reveló varios grupos cloroplásticos en los tipos silvestres y cultivados, pero los estudios anteriores dieron informaciones insuficientes sobre la estructura del cromosoma cloroplástico. Este documento muestra los principales elementos conocidos de la estructura general del cloroplasto de los *Musa*. **MATERIAL Y MÉTODOS.** La estructura del Gc del plátano se estudió a partir de ocho sondas correspondientes a sitios etiquetados por su secuencia (STS) homólogos y de cinco pares de iniciadores «universales». **RESULTADOS Y DISCUSIÓN.** Debido a las dos regiones, repeticiones inversas, RI, detectadas y las homologías de secuencia, el tamaño del cromosoma cloroplástico del plátano, a pesar de parecerse, no está estrechamente emparentado con la hierba. El análisis de secuencias no codificantes en 5' de genes cloroplásticos y el uso de codones indican que una transcripción diferenciada pudo ocurrir en el plátano y el arroz. **CONCLUSIÓN.** Globalmente, se puede utilizar una colección de once sondas RFLP de plátano, cubriendo de manera uniforme el genoma cloroplástico (14 loci).

PALABRAS CLAVES

Musa, ADN, cloroplasto, mapas genéticos, marcadores genéticos.