CHAPTER 3

GENEALOGICAL RELATIONSHIPS OF BANANA CULTIVARS INFERRED FROM CHLOROPLAST DNA SEQUENCE ANALYSIS

ABSTRACT

Bananas are one of the oldest cultivated crops. It is widely believed that almost all edible seedless banana cultivars have been derived from, or are hybrids of, two wild diploid species, namely *Musa acuminata* Colla (AA) and *M. balbisiana* Colla (BB). Though the banana domestication is believed to have occurred in the Southeast Asian region, the genomic and geographical origin(s) of the ancestral bananas are still uncertain. We are exploring the use of polymorphisms in chloroplast (cp) non-coding loci to determine the domestication history of bananas with special emphasis on the Southeast Asian ABB/BBA complex ('Hin'/ 'Saba'/ 'Namwa'). Four cp-marker loci have been developed—*rpl*16 and *ndh*A introns, *psa*A-*ycf*3 and *pet*A-*psb*J-*psb*L-*psb*F intergenic spacers. Using PCR-RF-SSCP (Polymerase Chain Reaction-Restriction Fragment-Single Strand Conformation Polymorphism or PRS) analysis, several cp-haplotypes were detected among the A and B genomes and several alleles from cultivars and wild relatives were sequenced . By analyzing the four loci-combined sequences, the A and B genome derived alleles could be easily distinguished. Polymorphism within each of the genomes included single nucleotide substitutions and insertion/deletions (indels). To estimate the maternal genealogy based on these data, most parsimonious trees were generated using the maximum parsimony and neighbor-joining methods, and statistical most parsimony-haplotype networks were reconstructed. Four Australimusa bananas, i.e. *M. jackeyi,*

M. textilis and two *Musa* 'Fehi' cultivars, were used as outgroup taxa. Different treatments of indels were compared. Diversity within triploid hybrids was found and their maternal lineages were identified. Five cp-haplotypes were found; two derived from different *M. balbisiana* and three from *M. acuminata*. Particular B genome-rich triploids (ABBs) and a tetraploid (ABBB) were given new genomic designations as BBA and BBBA to indicate their maternal genealogy. The relationships among these alleles and their information concerning the domestication of the Southeast Asian ABB/BBA bananas are discussed.

3.1 INTRODUCTION

Diversity and relationship within the *Musa* complex have puzzled crop scientists for more than 50 years since the time of Cheesman (1947) and Simmonds and Shepherd (1955). As one of the earliest cultivated crops, it was hypothesized that bananas have been domesticated and intensively selected over several millennia by farmers in the Southeast Asian region (Simmonds 1962, 1995; De Langhe and De Maret 1999). It is widely believed that edible seedless banana cultivars have been derived from, or are hybrids of, two wild diploid species in the Eumusa section (basic chromosome number = 11), namely *Musa acuminata* Colla (AA) and *M. balbisiana* Colla (BB) (Cheesman 1947; Simmonds and Shepherd 1955; Shepherd 1990).

The center of origin of hybrids between *M. balbisiana* and *M. acuminata* is subject of debate. Simmonds (1962) argued that AA cultivars, possibly originated in Malaya, were carried into wild *M. balbisiana* distribution areas. He suggested that India and eastern Malaysia were centers of hybrid origin. However, it was also believed

that Neolithic Austronesian people introduced *M. balbisiana* from mainland Southeast Asia into the Philippines and Pacific Islands (Simmonds 1962; Argent 1976; De Langhe and De Maret 1999) while they traveled and colonized the islands between 4,500 and 3,500 years ago (De Langhe and De Maret 1999; Birds et al. 2003). Contrarily to Simmonds' theory, De Langhe and De Maret (1999) proposed that the location of hybridization that gave rise to the AAB hybrids was in eastern Indonesia and the Bismarck archipelago where anthropogenic *M. balbisiana* was brought into the areas where *M. acuminata* ssp. *banksii* was found. They believed that subsequent introductions of the hybrids into other tropical areas over the world, also back to mainland Asia, were attributed to ancient sea travelers.

Controversy exists concerning genome constitutions of ABB/BBB group of cultivars, including e. g. 'Saba', its synonym, 'Hin' (Valmayor et al. 2000), and 'Lep Chang Kut'. Several investigations intending to clarify its taxonomic status have been undertaken, but their origin is still obscure (Simmonds 1962; Rivera 1983; Espino and Pimentel 1990; Shepherd 1990; Jarret and Litz 1986).

The *M. balbisiana* genome is important in banana improvement for assorted quality characteristics as well as for providing hardiness and disease resistance. In general, the B genome gives starchiness and acidity to the fruit, characters commonly considered as cooking quality. *M. balbisiana* is also more tolerant to drought and flooding and more resistant to Panama disease, leaf spot, and nematodes than *M. acuminata* (Simmonds 1987).

Though, none of *balbisiana* genome-containing bananas are significant in the world trade, they are invaluable for rural people in Southeast Asia and other tropical countries. Production of bananas in Southeast Asia, especially for local consumption, is mostly based on triploid hybrid cultivars such as 'Saba' (BBA/BBB) in the Philippines and 'Hin' in Thailand and 'Pisang Kepok' in Indonesia; 'Namwa' (BBA) in Thailand and its synonym 'Pisang Awak' in Malaysia and Indonesia; and 'Chuoi Tay' (BBA) in Vietnam (see Valmayor et al. 2000). Knowledge of the relationships within and among wild accessions, diploid and triploid cultivars, is required for the successful selections of parental resistance to pests and diseases in the breeding programs (INIBAP 2001).

It is undeniable that *M. acuminata* and *M. balbisiana* have played dominant roles in the evolution of edible clones. However, there is still inadequate data about the role that diversity within both species has played in the evolution of hybrid types, or where and when these ancestral hybridizations occurred (Jarret 1990). It is widely accepted that there are several subspecies within *M. acuminata* e.g. ssp. *siamea, burmanica, malaccensis, microcarpa*, and *banksii*, and the first four are found in Thailand. However, up to now, little is known about the extent of genetic diversity within *M. balbisiana*. It was found through isozyme analysis, though, that *M. balbisiana* is a polymorphic species (Lebot et al. 1993), which was in agreement with Kaemmer et al. (1997) who found that the B genome present in plantains (AAB) and the B genome detected in ABB bananas came from different *M. balbisiana* donors. Similarly, Shepherd (1990) suggested that AAB and ABB cultivars could have evolved in different areas at different times. Simmonds and Shepherd (1955) noted that AAB and ABB triploids could have originated from

fertilization of unreduced diploid eggs in sterile AB hybrids and haploid pollen.

Simmonds (1962) concluded that *M. balbisiana* is not native to Thailand because, based on his findings—although is commonly cultivated for male bud consumption and its leaves, which are used as wrappings—*M. balbisiana* lacked a specific vernacular name. This observation implies that AAB and ABB cultivars are not likely to have originated in Thailand. However, combined with the fact that four out of five subspecies of *M. acuminata* were found by Simmonds (1956, 1962) in Thailand—none of which has a vernacular name either—suggests that Thailand is a hotspot of banana diversity (Chomchalow and Silayoi, 1984). In addition, somatic mutations in ABB group of Pisang Awak' or 'Kluai Namwa' has been reported only from Thailand, e.g. 'Kluai Namwa Daeng' with pink flesh; 'Kluai Namwa Khao' with more waxy fruits than usual, 'Kluai Namwa Nuan' with waxy rind, and 'Kluai Namwa Khom' is a dwarf mutant, in addition to common 'Kluai Namwa Luang' with yellow flesh (Chomchalow and Silayoi, 1984).

The identification of banana cultivars has been based on morphological characters and chromosome numbers and referred to notable work by Simmonds and Shepherd (1955). This genome classification of hybrid cultivars has been reinvestigated over the years and still most is used. Several techniques including secondary metabolite analysis (Horry and Jay 1990), genome content (Dolezel et al. 1994; Lysá k et al. 1999), and isozyme markers have been applied and proved successful (Jarret and Litz 1986; Espino and Pimentel 1990; Howell et al. 1994; Lebot et al. 1993). More recently, molecular techniques, i.e. RFLP (Gawel and Jarret 1991a,b; Jarret et al. 1992; Bhat et al. 1994; Carreel et

al. 2002), PCR-RFLP (Nwakanma et al. 2003), STMS (Kaemmer et al. 1997), molecular cytogenetics (Osuji et al. 1997; D'Hont et al. 2000), AFLP (Loh et al. 2000; Ude et al. 2002a,b; Wong et al. 2001; Wong et al. 2002), and dCAPs (derived cleaved amplified polymorphic sequence) markers (Umali and Nakamura 2003), have been applied.

The study of non-coding chloroplast (cp-) sequences has proved to be useful in plant phylogenetic analyses at lower taxonomic levels (Golenberg 1993; Gielly and Taberlet 1994; Kelchner and Clark 1997; Ohsako and Ohnishi 2000; Zhang 2000; Specht et al. 2001; Baumel et al. 2002). These sequences tend to evolve more rapidly than do coding sequences, by accumulation of insertions and deletions (indels) at a rate at least equal to that for nucleotide substitutions (Curtis and Clegg 1984; Wolfe et al. 1987). Compared with the cp-sequence of rice (*Oryza sativa* cultivar group *japonica*, accession number NC001320 (Hiratsuka et al. 1989), the fragments selected in our investigations, were four non-coding regions, i.e. *ndh*A and *rpl*16 introns (N-IN and R-IN) and *psa*A-*ycf*3 and *pet*A-*psb*J-*psb*L-*psb*F intergenic spacers (PY-S and AF-S). The first fragment located in small single copy region and the rest were in large single copy region of the cp-genome (see Appendix A).

The objective of this study was to investigate the cpDNA diversity to infer the genealogy of Southeast Asian banana cultivars. Several alleles from the four loci were assessed using a combined method consisting of PCR, restriction enzyme digestion and SSCP (PCR-RF-SSCP or PRS) (Slabaugh et al. 1997; Schneider et al. 1999; Sato and Nishio 2002). Different PRS haplotypes identified among banana accessions were then sequenced and analyzed to infer their phylogenetic relationships using distance and maximum parsimony methods.

3.2 MATERIALS AND METHODS

3.2.1 Plant materials

Forty-six accessions of *M. acuminata*, *M. balbisiana* (both from the Eumusa section, having basic chromosome number, $x = 11$), cultivars, and hybrids were chosen to represent seven genomic constitutions designated as AA, BB, AB, AAB, ABB, ABBB, BBB, and also to reflect their geographical distributions (Table 3.1). Four accessions from the Australimusa section (basic chromosome number, $x = 10$) i.e. *Musa jackeyi*, *Musa* Fehi 'Aata', *Musa* Fehi 'Tongkat Langit', and *M. textilis*, were included as outgroup taxa. Young curling leaves in pseudostems were collected. DNA was extracted from leaf samples using DNeasy Plant Mini Kit (Qiagen, USA) according to the manufacturer's instructions. Twenty-two extracted DNA samples were generously provided by Dr. Françoise Carreel, CIRAD-FLHOR, Guadeloupe, and by the International Transit Center (ITC) of the International Network for the Improvement of Banana and Plantain (INIBAP), Leuven, Belgium. Two samples from the Philippines were supplied by the Institute of Plant Breeding, Los Baños, College, Laguna, Philippines.

3.2.2 Primers

To amplify the non-coding regions, four primer sets were designed from published genomic DNA sequences (NCBI GenBank). Initially, cp-regions containing mononucleotide repeats in rice and tobacco were compared, then other species, as many as available, were aligned and flanking conserved regions were selected. If all species analyzed retained a mononucleotide repeat, the locus was retained for primer design. Details of these primers are presented in Table 3.2. See appendix A for locations of selected fragments on rice cp-genome.

Table 3.1 List of *Musa* accessions, their genome composition, origins and sources of accessions used in this study.

No.	Species/Cultivar	Code	Origin[a]	Source[b]
1	*M. balbisiana* 'Tani'	BB-CHS	Chachaeng Sao, Thailand	TM083
2	*M. balbisiana* 'Cameroon'	BB-CMR	?	ITC0246 / CIRAD
3	*M. balbisiana* 'Tani'	BB-CYP	Chaiyaphum, Thailand	KU
4	*M. balbisiana* 'Honduras'	BB-HDR	?	ITC0247 / CIRAD
5	*M. balbisiana* 'Lal Velchi'	BB-LVC	India	ITC0626 / CIRAD
6	*M. balbisiana* 'MPL'	BB-MPL	?	ITC0212
7	*M. balbisiana* (wild)	BB-NAN	Nan, Thailand	TM123
8	*M. balbisiana* 'Tani'	BB-NST	Nakorn Si Thammarat, Thailand	KU
9	*M. balbisiana* (wild)	BB-P28	Cagayan, Philippines	IPB-P28
10	*M. balbisiana* 'Pisang Klutuk Wulung'	BB-PKW	Indonesia	ITC1063 / CIRAD
11	*M. balbisiana* 'Pang La Nuan'	BB-PLN	Trang, Thailand	TM121
12	*M. balbisiana* 'Tani'	BB-PRA	Prae, Thailand	KU
13	*M. balbisiana* 'Tani'	BB-SKT	Sukhothai, Thailand	TM127
14	*Musa* 'Balonkawe'	BBBA-BLK	?	ITC0473
15	*Musa* 'Pli Hai'	BBBA-PLH	Chiang Rai, Thailand	KU
16	*Musa* 'Tipparot'	BBBA-TPR	Srisaket, Thailand	KU
17	*Musa* 'Lep Chang Kut'	BBB-LCK	Thailand	KU
18	*Musa* 'Hin'	BBA-HIN	Nakorn Sri Thammarat, Thailand	KU
19	*Musa* 'Moe Nang'	BBA-MNG	Thailand	ITC1066

(continued)

Table 3.1 (Continued) List of *Musa* accessions, their genome composition, origins and sources of accessions used in this study.

No.	Species/Cultivar	Code	Origin[a]	Source[b]
20	*Musa* 'Namwa Sai Daeng'	BBA-NWA	Thailand	KU
21	*Musa* 'Namwa Dum'	BBA-NWD	Thailand	KU
22	*Musa* 'Namwa Khom' (dwarft)	BBA-NWK	Nakorn Ratchasima, Thailand	KU
23	*Musa* 'Namwa Sai Lueang'	BBA-NWL	Thailand	KU
24	*Musa* 'Namwa Mali-Ong'	BBA-NWM	Thailand	KU
25	*Musa* 'Namwa Nuan'	BBA-NWN	Ang Thong, Thailand	KU
26	*Musa* 'Namwa Sai Dum'	BBA-NWS	Thailand	KU
27	*Musa* 'Namwa Khaow'	BBA-NWW	Prae, Thailand	KU
28	*Musa* 'Saba'	BBA-PSB	Philippines	IPB-P93
29	Musa 'Teeb'	BBA-TEB	Kon Kaen, Thailand	KU
30	*Musa* 'Hin'	ABB-HIN	Yala, Thailand	TM128
31	*Musa* 'Hin'	ABB-HIN	Songkhla, Thailand	TM129
32	*Musa* 'Hak Muk Khiaw'	ABB-HMI	Thailand	KU
33	*Musa* 'Hak Muk Khaow'	ABB-HMO	Thailand	KU
34	*Musa* 'Saba'	ABB-ISB	Philippines	ITC0116
35	*Musa* 'Pisang Abuperak'	ABB-PAP	Indonesia	ITC0056
36	*Musa* 'Pisang Gajih Merah'	ABB-PGM	Indonesia	ITC0677
37	*Musa* 'Auko'	AB-AUK	?	ITC0983
38	*Musa* 'Kofi'	AAB-KOF	?	ITC0912
39	*Musa* 'Popoulou'	AAB-POP	Papua New Guinea	ITC0335

(continued)

Table 3.1 (Continued) List of *Musa* accessions, their genome composition, origins and sources of accessions used in this study.

No.	Species/Cultivar	Code	Origin[a]	Source[b]
40	*Musa* 'Teeb Kum'	AAB-TBK	Thailand	ITC0667
41	*M. acuminata* ssp. *banksii* 'Madang'	AA-BKS	Papua New Guinea	CIRAD
42	*M. acuminata* ssp. *burmannicoïdes* 'Calcutta 4'	AA-BUR	India	CIRAD
43	*M. acuminata* ssp. *malaccensis*	AA-MAL	Malaysia	CIRAD
44	*M. acuminata* ssp. *siamea*	AA-SAM	Phrae, Thailand	KU
45	*Musa* 'Grande Naine'	AAA-GNN	?	ITC0180 / CIRAD
46	*Musa* 'Hom Thong'	AAA-HOM	Thailand	QSG
47	*M. jackeyi*	TT-JCK	Northeast Queensland, Australia	CIRAD
48	*Musa* Fehi 'Aata'	TT-FAA	Papua New Guinea	CIRAD
49	*Musa* Fehi 'Tongkat Langit'	TT-FTL	Indonesia	CIRAD
50	*M. textilis*	TT-TEX	Philippines	CIRAD

[a] ? indicates unknown origins

[b] CIRAD= Centre de Coopération Internationale en Recherche Agronomique pour le Dévelopement, Guadeloupe; ITC= International Transit Center, International Network for the Improvement of Banana and Plantain (INIBAP), Leuven, Belgium; KU= Kasetsart University Student Training Farm, Nakorn Ratchasima, Thailand; TM= Germplasm collected by the authors in Thailand; QSG= Queen Sirikit Garden, Bangkok, Thailand; IPB= Institute of Plant Breeding, University of the Philippines, Los Baños, College, Laguna, Philippines

Table 3.2 Primers for amplifications

Name	Nucleotide sequence	Tm (°C)	Amplified region	Sequence code	Location[a]	Protein encoded
*rpl*16-F	5'GCTATGCTTAGTGTGTGACT3'	55.3	*rpl*16 intron	R-IN	LSC	ribosomal protein L16
*rpl*16-R	5'CATTCTTCCTCTATGTTGTTT3'	52.0				
*psa*A-F	5'AAATCGTGAGCATCAGCATG3'	54.7	*psa*A- *ycf*3 exon3 intergenic spacer	PY-S	LSC	photosystem I P700 apoprotein A1 and photosystem I assembly protein Ycf3
*psa*A-R	5'CCGAGGAGAACAGGCCATTC3'	58.6				
*ndh*A-F	5'GCTGCTCAATCTATTAGTTATGA3'	55.3	*ndh*A intron	N-IN	SSC	NADH dehydrogenase subunit 1
*ndh*A-R	5'TGTGCTTCAACTATATCAACTGT3'	53.7				
*pet*A-F	5'TATGAAAATCCACGAGAAGC3'	51.3	*pet*A-*psb*J-*psb*L-*psb*F intergenic spacer	AF-S	LSC	cytochrome f, photosystem II protein J and L, and cytochrome b559 beta chain, respectively
*pet*A-R	5'TATCAGCAATGCAGTTCATC3'	51.7				

[a] LSC = large single copy chloroplast region and SSC = small single copy chloroplast region

3.2.3 PRS analysis

With some modifications, analyses of PRS were done based on techniques reported previously (Orita et al. 1989; Jordan et al. 1998; Sato and Nishio 2002). PCR was performed in a total reaction mixture of 30 μL, containing 200 μM dNTPs (Promega), 1.5 mM $MgCl_2$, 0.5 μM each primers, 1x *Taq* DNA polymerase buffer, 1 unit of *Taq* DNA polymerase (Qiagen) and 20-50 ng of total genomic DNA as template. Amplification was carried out at 94°C for 3 min, followed by 32 cycles of 45 sec at 94° C, 45 sec at suitable annealing temperature for each primer set, 90 sec at 72°C, and final extension at 72°C for 5 min using a T1 Thermocycler 96 (Whatman, Biometra, Germany). The PCR product was checked on 1% agarose gel using 1xTAE buffer (40 mM Tris, 0.114% glacial acetic acid, 1 mM EDTA, pH 8). PCR products of PY-S was not digested, the R-IN and N-IN fragments were digested with the *Hinf*I restriction enzyme, and the AF-S fragment was digested with *Taq*I (all from New England BioLabs, Inc, USA), under the conditions recommended by the supplier. Two volumes of denaturing solution (95% (v/v) formamide, 0.025% bromphenol blue, 0.025% (w/v) xylene cyanol and 10 mM NaOH) were mixed with PCR or digested products, then denatured for 10 min at 95°C, and immediately placed on ice to stabilize single strands. Then 3.5 μL aliquots were loaded on a non-denaturing polyacrylamide gel (Sequagel MD, National Diagnostics, USA). Electrophoresis was performed in 1x TBE buffer using Hoefer SQ3 Sequencer (Amersham Pharmacia Biotech, USA) and run in a 4°C refrigerator at constant 8 watt for 12-14 hr. The gel was then silver stained using standard procedure (Guillemette and Lewis 1983). Accessions showing polymorphic bands were grouped as haplotypes and one to six accessions from each haplotype representing geographical distributions were selected for cloning and sequencing.

3.2.4 Cloning and sequencing of selected haplotypic fragments

The undigested PCR fragments were purified using MinElute PCR Purification Kit (Qiagen) and cloned into *E. coli* 'DH10B' using either TOPO TA Cloning Kit for sequencing (Invitrogen) or pGEM-T Vector Systems (Promega). Plasmids were extracted using Wizard® *Plus* SV Minipreps DNA Purification System (Promega) and sent for sequencing (Macrogen, Inc, Seoul, South Korea).

3.2.5 Sequence alignments and analyses

Nucleotide sequences of the chloroplast DNA were first aligned with the GeneDoc program version 2.6.002 (Nicholas and Nicholas 1997) and manually edited. Sequences from all four chloroplast loci were combined and analyzed as a single locus (Olmstead and Sweere 1994; Soltis and Soltis 1998). Base compositions were determined by analysis of the aligned matrix (excluding gaps and missing data) using the DnaSP program version 4 (Rozas et al. 2003). Singletons, if not confirmed by sequencing from multiple clones, were eliminated. Variable regions with mononucleotide repeats and where alignments cannot be satisfactorily done were omitted from further phylogenetic analyses (Zhang 2000).

3.2.6 Phylogenetic analyses

Phylogenetic relationships among *Musa* cpDNA were inferred, with the Australimusa accessions as outgroup, by the neighbor-joining (NJ) and maximum parsimony (MP) methods using PAUP* version 4.0b.10 (Swofford 2002). Bootstrap and majority-rule consensus from 500 replicates were conducted to evaluate the support of branches.

The NJ analysis was done using Kimura-2-parameter model with all indels excluded from the data matrix. Meanwhile, prior to MP analysis, three different coding methods were applied for gaps: (i)

complete deletion of indel regions, (ii) retaining all insertions and treating deletions as missing data, and (iii) removing all indels from the data matrix, identifying every potential mutation event and applying unordered multistate characters (0/1/2) for each event, then adding the characters back to the matrix (Baum et al. 1994; Peralta and Spooner 2001).

The MP trees were constructed with all characters equally weighted. Heuristic search with 100 random addition analyses and tree bisection-reconnection (TBR) branch-swapping (Steepest descent option in effect), stepwise addition, and MULPARS options was used in the search for most-parsimonious trees. The amount of homoplasy was evaluated with the consistency index (CI), excluding uninformative characters, and the retention index (RI). A strict consensus tree was generated from the equally most-parsimonious trees. To evaluate the branch supports, bootstrap and 50% majority-rule consensus from 500 replicates were conducted.

3.2.7 Network reconstruction

Gene genealogies were estimated using network reconstructed by statistical parsimony algorithm (Templeton et al. 1992) generated by the program TCS 1.13 (Clement et al. 2000). With this method, unrooted cladograms that have a high probability (>95%) of being true based on a finite-site model of DNA evolution are identified. Each indel, regardless its size, was counted as one mutational event and treated as fifth state. Nucleotide substitutions within indels were treated as additional events.

3.3 RESULTS

3.3.1 Chloroplast haplotypic analyses by PRS

PRS analyses of the four cp-regions revealed five haplotypes in the Eumusa section (Table 3.3). All four loci clearly distinguished *M. balbisiana* and *M. acuminata*. Thereupon, the polymorphisms between B and A genome obviously differentiated hybrids. The maternal contributions in autopolyploid cultivars of the two species were easily identified e. g. 'LCK' (BBB) showed the same pattern as other B accessions and 'HOM' (AAA) was grouped with other A accessions. Two cp-haplotypes were found in B (BI and BII) and three in A accessions (AI, AII, and AIII). In B accessions, two cp-haplotypes each were found in N-IN digested with *Hinf*I and AF-S digested with *Taq*I, while only one was found in R-IN digested with *Hinf*I and undigested PY-S. Cp-haplotype BI included six *M. balbisiana*, two BBA and all tetraploid hybrids, while cp-haplotype BII comprised of seven *M. balbisiana* including 'SKT', BBA 'HIN' and all 'Namwa'. On the other hand, within A accessions, three haplotypes each were found in undigested PY-S and AF-S digested with *Taq*I. Surprisingly, the 'Saba' and 'Hin' accessions represented two different PRS patterns within their cultivars—BBA 'PSB' and 'HIN' contained B-like chloroplast (BII) and ABB 'ISB', 'HIN', and 'HIS' contained A-like (AII).

3.3.2 CpDNA sequence analyses

A total of 11 B accessions and 7 A accessions (Table 3.3) were selected as representatives of their haplotypes and geographical distribution for DNA sequencing, altogether with four Australimusa bananas. The complete cp-sequences are displayed in Appendix B.

Table 3.3 Chloroplast haplotypes from PCR-RF-SSCP (PRS) analyses of *Musa*.

No.	Chloroplast haplotypes [a]				Accession code[b]
	R-IN/*Hf*I	PY-S	N-IN/*Hf*I	AF-S/*Tq*I	
BI	B	B	B1	B1	BB-CHS, **BB-CMR**, BB-MPL, **BB-NAN**, BB-NST, BB-PRA, **BBA-MNG**, BBA-TEB, BBBA-BLK, **BBBA-PLH**, BBBA-TPR, **BBB-LCK**
BII	B	B	B2	B2	BB-CYP, BB-HDR, BB-LVC, **BB-P28**, **BB-PKW**, BB-PLN, **BB-SKT**, **BBA-HIN**, **BBA-PSB**, BBA-NWA, BBA-NWD, BBA-NWK, BBA-NWL, BBA-NWM, **BBA-NWN**, BBA-NWS, BBA-NWW
AI	A	A1	A	A1	ABB-ISB, ABB-HMI, ABB-HMO, **AA-SAM**
AII	A	A2	A	A2	**ABB-HIN**, ABB-HIS, ABB-PAP, **ABB-PGM**, **AB-AUK**, AAB-KOF, AAB-POP
AIII	A	A2	A	A3	ABB-TBK, **AA-BKS**, AA-BUR, **AA-MAL**, AAA-GNN, **AAA-HOM**

[a] Chloroplast haplotypes of four regions; R-IN/*Hf*I = *rpl*16 intron digested with *Hinf*I restriction enzyme, PY-S = undigested *psa*A-*ycf*3 intergenic spacer, N-IN/*Hf*I = *ndh*A intron digested with *Hinf*I restriction enzyme, and AF-S/*Tq*I = *pet*A-*psb*J-*psb*L-*psb*F intergenic spacers digested with *Taq*I restriction enzymes

[b] Accession code appears according to Table 3.1. Eighteen accessions in bold-type face were selected for DNA sequencing.

The combined cpDNA sequences consisting of four fragments were mostly non-coding (73%) (Fig. 3.1). Sequence variability of the four cpDNA regions is summarized in Table 3.4.

The sequence length ranged from 5,584-5,631 bp in B accessions and 5,461-5,506 bp in A accessions. After alignment, the sequences resulted in a final data matrix with 5,774 positions by introducing 39 gaps, among which 10 were due to outgroup indels. Total gap length was 404 bp, which is about 7% of the aligned sequence length. Seven gaps were single nucleotide indels, 17 were small (2-10 bp), and 15 were large, ranging from 11-54 bp. The average GC content of the combined sequences is 0.34.

An inverted repeat, which could form a hairpin secondary structure with a 24-bp loop and 31-bp stem, was found in AF-S fragment at position 5,237-5,329, 29 bp downstream of *psb*J coding region (Fig. 3.2). Short inverted repeats were also found within this loop. In addition to the indels in the loop, one nucleotide difference at position 5295 was found among the two sections.

3.3.3 Phylogenetic analyses

In the neighbor-joining distance analysis (Fig. 3.3) of the aligned DNA sequence data set, the occurrence of the B and A clades was supported by 100% bootstrap values for both clades (bootstrap value not shown). The analysis presented three lineages within A clade. Unsolved relationship within the B clade was shown, with bootstrap value of 59%.

Table 3.4 Average base composition, and length comparison of sequences and indels in combined chloroplast sequences of *Musa*. Aligned data matrix appears in Appendix B.

Sequence Code[a]	Regions	Positions on Data Matrix	Aligned Length (bp)	No. of Indels	Indel Length (bp)	Average Base Composition			
						A	C	G	T
R-IN	rpl16 exons	1-12 1081-1122	54	-	-	0.48	0.19	0.19	0.14
	rpl16 intron	13-1080	1068	7	113	0.40	0.14	0.17	0.29
PY-S	psaA	1123-1289	167	-	-	0.25	0.23	0.23	0.29
	psaA-rcf3 spacer	1290-1970	681	3	51	0.31	0.18	0.14	0.37
	rcf3	1971-2124	154	-	-	0.26	0.27	0.16	0.30
N-IN	ndhA exons	2125-2185 3347-3387	102	-	-	0.33	0.18	0.14	0.35
	ndhA intron	2186-3346	1161	19	104	0.37	0.13	0.17	0.33
AF-S	pet A	3388-4219	832	-		0.31	0.18	0.22	0.29
	petA-psbJ spacer	4220-5358	1139	11	136	0.34	0.14	0.17	0.35
	psbJ	5359-5481	123	-	-	0.41	0.26	0.17	0.16
	psbJ-psbL spacer	5382-5606	125	-	-	0.27	0.22	0.15	0.36
	psbL	5607-5722	116	-	-	0.39	0.14	0.17	0.30
	psbL-psbF spacer	5723-5745	23	-	-	0.37	0.11	0.13	0.39
	psbF	5746-5774	29	-	-	0.34	0.16	0.18	0.32
Total			5774	39	404	0.34	0.16	0.18	0.32

[a] Sequence code appears according to Table 3.2.

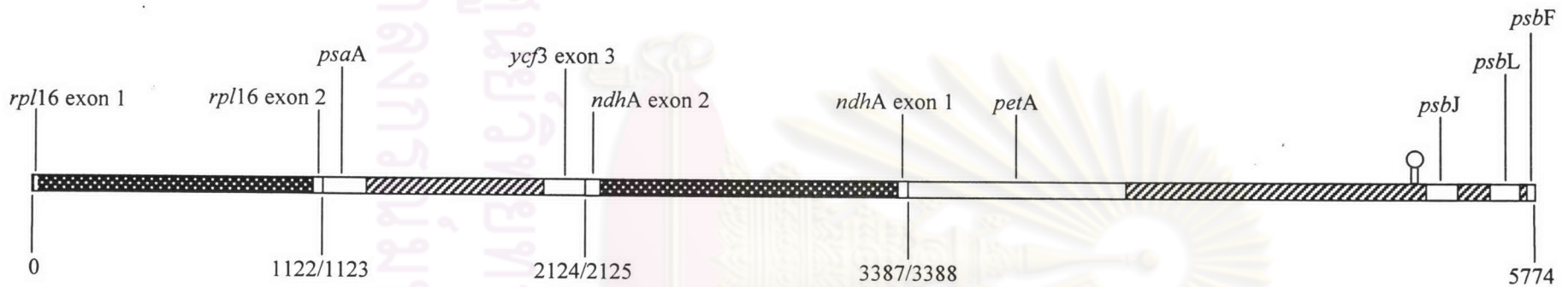


Fig. 3.1 Diagram representation of the combined chloroplast sequence used in the phylogenetic analysis of *Musa*. The total of 5.4-kb sequence contained 1.0-kb *rpl*16 intron (R-IN), 1.0-kb *psa*A-*ycf*3 intergenic spacer (PY-S), 1.1-kb *ndh*A (N-IN), and 2.3-kb *pet*A-*psb*J-*psb*L-*psb*F intergenic spacer (AF-S) fragments. Numbers indicate positions on the aligned data matrix and boxes are exons (□), introns (▩), and intergenic spacers (▨). Inverted repeats found at 29-bp downstream of *psb*J (see text) is indicated by hairpin structure ( ).

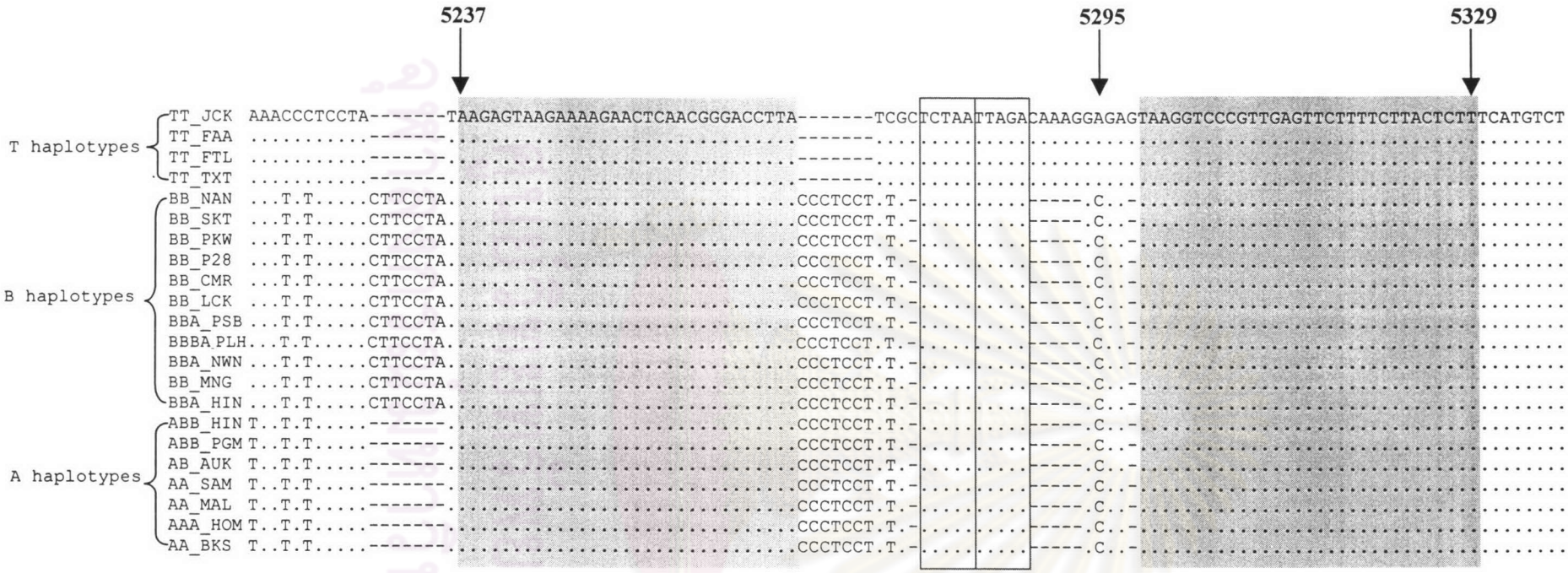


Fig. 3.2 Inverted repeats within *pet*A-*psb*J intergenic spacer region in *Musa*. Sequence alignment showing region of inverted repeats (in shaded boxes) located at position 5237-5329, 29 bp downstream of *psb*J coding region. Nucleotide substitutions occurred at position 5295. In addition of nucleotide substitutions and indels, the loop contains another short (5-bp) inverted repeat (indicated by open boxes) in between. Sequence codes appear according to Table 3.1. Complete data matrix is presented in Appendix B.

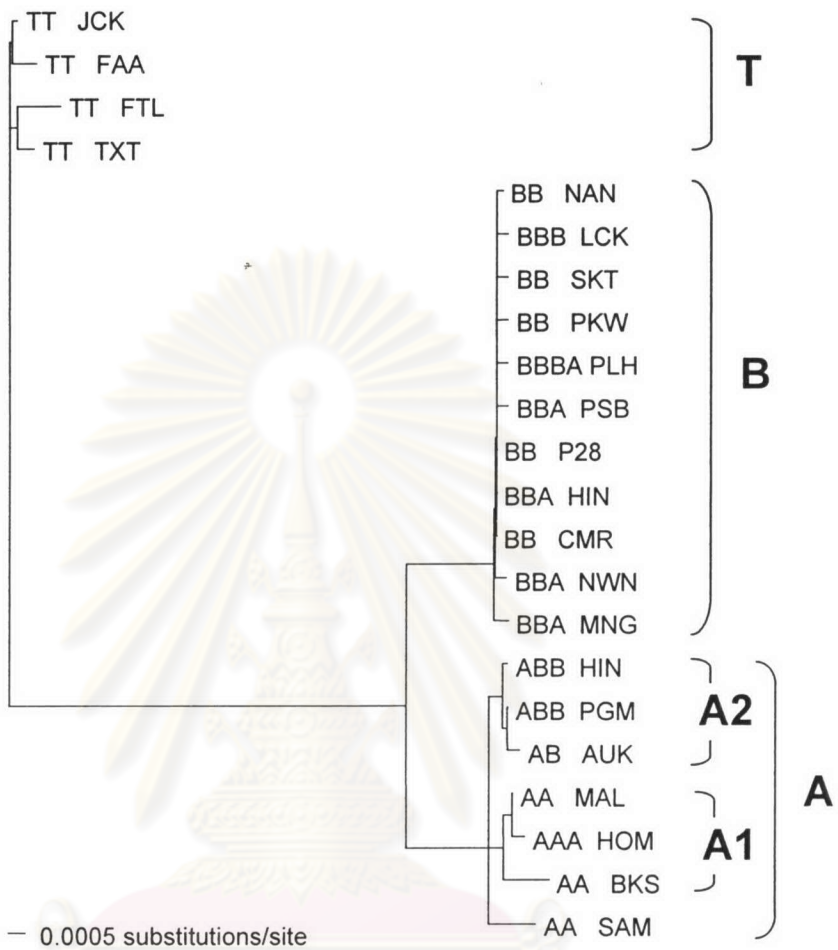


Fig 3.3 Neighbor-joining tree of *Musa* chloroplast combined sequences, rooted at the Australimusa bananas. *M. balbisiana* and *M. acuminata* clades are indicated by B and A. Within A clade, two subclades are designated as A1 and A2. Four accessions from Australimusa section indicated by 'T' are outgroup. Accession codes appear according to those in Table 3.1.

Maximum parsimony analyses with three different methods of gap coding yielded slightly different topologies; the first two analyses were similar whereas the last one was different. In general, rooted at the four Australimusa, the Eumusa bananas formed two subclades, B and A , with bootstrap value of 100%. In the first analysis, when indels were completely excluded, there were 144 variable sites of which 97 were parsimony informative. Twenty-one equally most-parsimonious trees were obtained, with tree length = 147, CI = 0.9700, and RI = 0.9929. The A clade in the strict consensus tree (Fig. 3.4) is monophyletic with two subclades. Interestingly, one of the subclade consisting solely of diploid *M. acuminata* and its autopolyploid 'HOM' with moderately high bootstrap support (60%), while another contained diploid and triploid hybrids with high bootstrap support. The relationship within the B clade, though, forming the clade with 100% bootstrap support, was not resolved.

In the second analysis, when indels were retained in the data set but coded as missing, there were 153 variable sites with 103 parsimony informative (figure not shown). The analysis yielded nine equally parsimonious trees with tree length = 156, CI = 0.9717, and RI = 0.9931. Strict consensus tree presented similar topology to the first analysis.

For the third analysis, which indels were removed, recoded as unordered multistate characters (coding matrix, Fig. 3.5), and added back to the data matrix, all 39 gaps were among 136 parsimony informative of 183 variable sites. The analysis resulted in only one most-parsimonious tree (Fig. 3.6) of length 189, with CI = 0.9577 and RI = 0.9908. A subclade consisting of 'NAN', 'CMR', 'LCK', and 'MNG' was resolved as polytomy within the B clade, with bootstrap value of 65%. The A clade consisted of three subspecies lineages, (i) *M. acuminata* ssp.

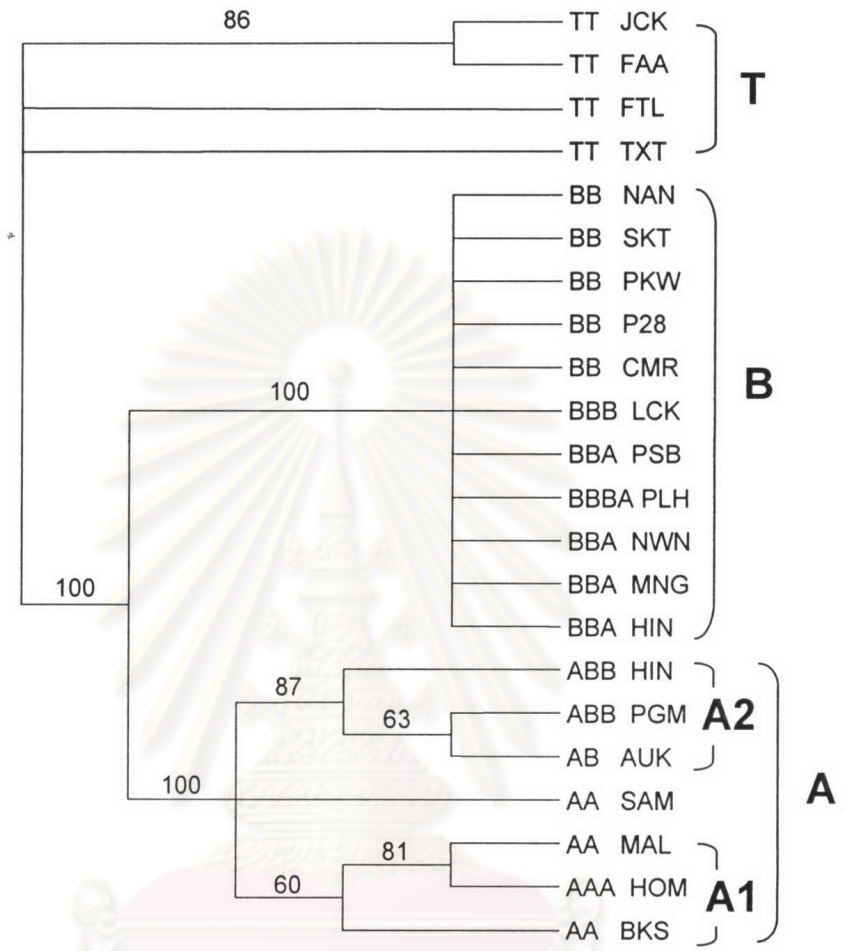


Fig. 3.4 Strict concensus tree of 21 most parsimonious trees from the analysis of *Musa* combined cpDNA sequence data with all gap excluded. Percentage of 500 bootstrap replicates is given above branch with majority-rule concensus. Letters T, B, A, A1, and A2 are designated to clades as in Fig. 3.3. Names of taxa appear according to Table 3.1.

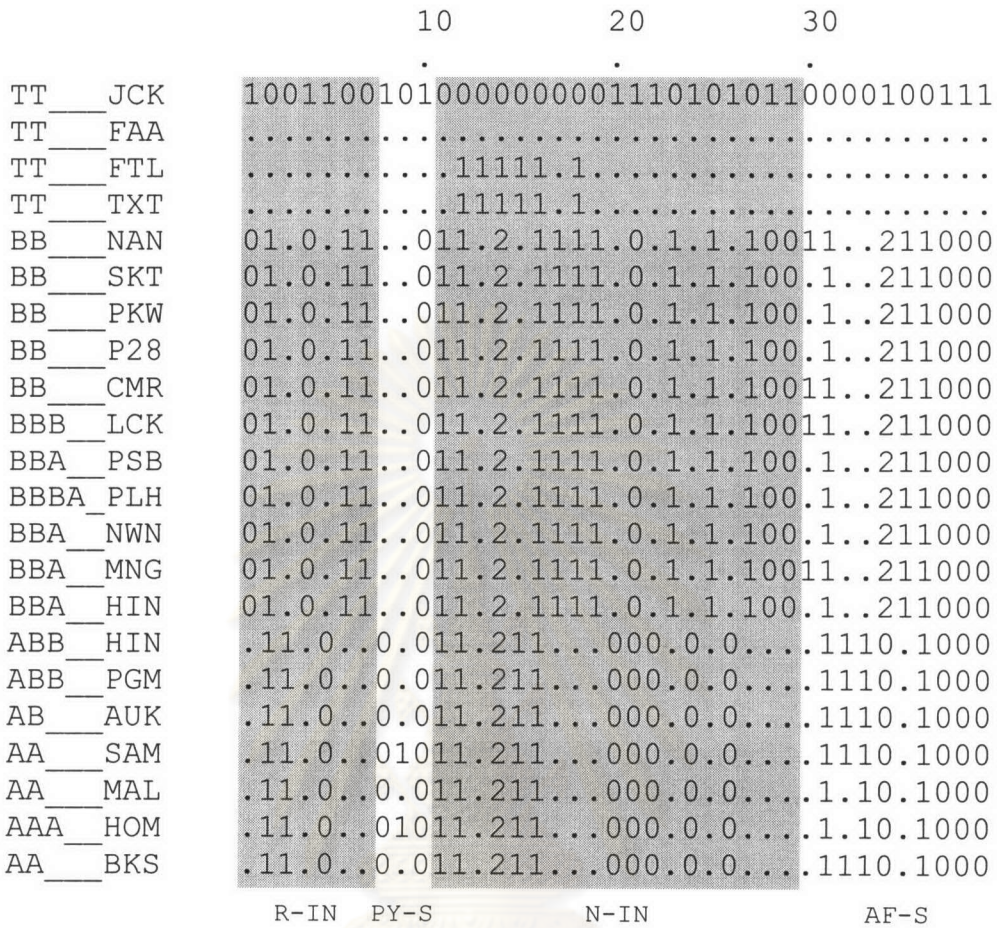


Fig. 3.5 Indel matrix of *Musa* combined chloroplast sequences. Unordered multistate characters indicated are 0=deletion, 1=insertion, 2=insertion with extra nucleotide difference. Dot symbol (.) show a site with the same character variant to that of the first sequence. Positions of the 39 indels with reference to the aligned data matrix (Appendix B) in four cp-regions are as follows:
R-IN (*rpl*16 intron): 54, 55-108, 419-427, 596, 798-819, 820-837, 956-963;
PY-S (*psa*A-*ycf*3 intergenic spacer): 1447-1469, 1649-1671, 1712-1716;
N-IN (*ndh*A intron): 2611-2614, 2615-2616, 2617-2621, 2622-2629, 2630-2634, 2635-2638, 2639-2651, 2654-2664, 2665, 2667-2680, 2681-2703, 2849-2855, 2919, 2920-2925, 2926-2931, 2932-2936, 2946-2950, 3143, 3222;
AF-S (*pet*A-*psb*J-*psb*L-*psb*F intergenic spacers): 4256-4299, 4801-4814, 4832-4842, 4867-4880, 4881-4908, 5229-5235, 5268-5274, 5278, 5289-5293, 5298.

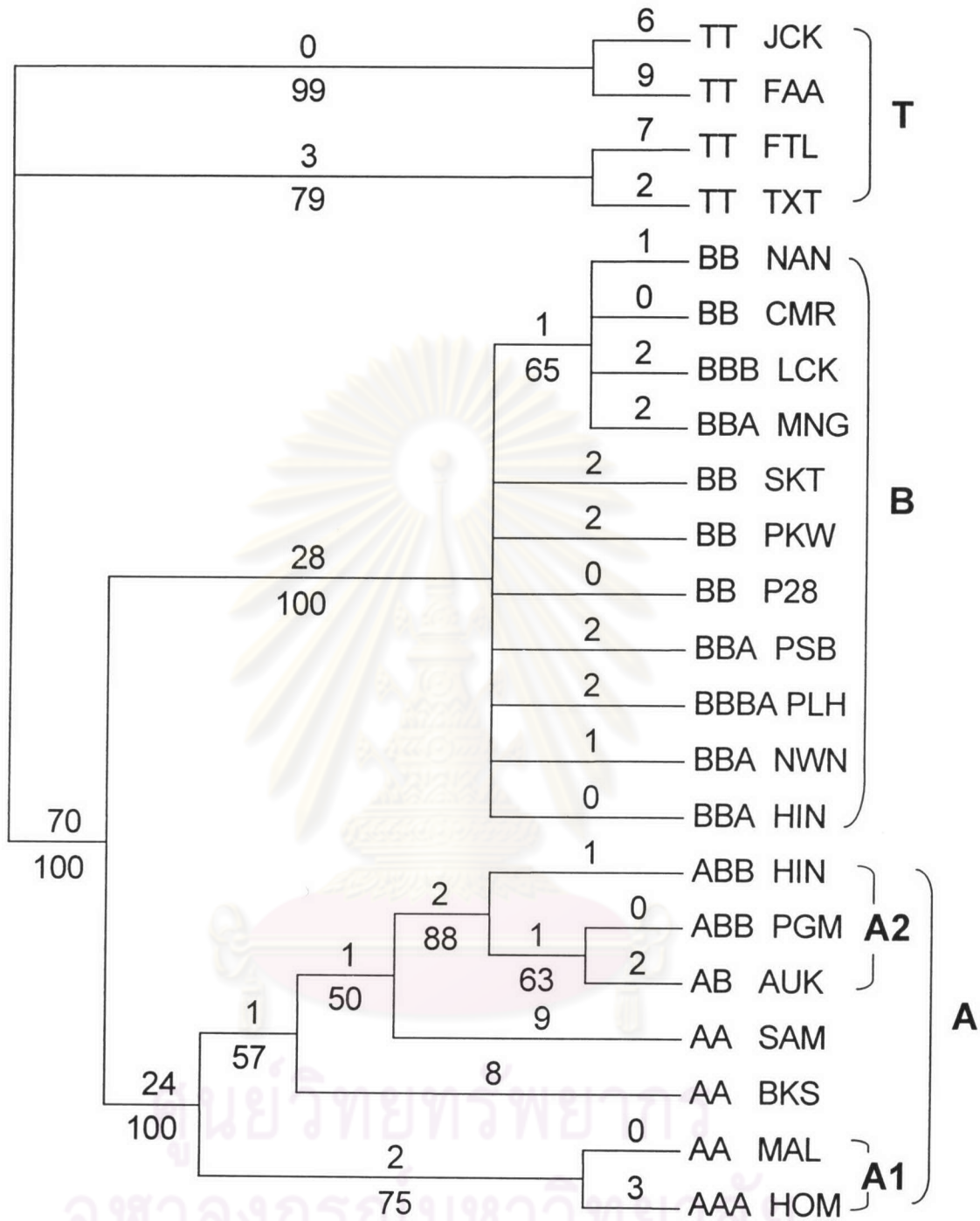


Fig. 3.6 The single most-parsimonious tree from the analysis of *Musa* combined cpDNA sequence data with gaps recoded. Number above branches show supporting characters for the clades and number below branches are bootstrap values (%) with 50% majority-rule concensus from 500 replicates. Tree length = 189, CI (excluding uninformative characters) = 0.9577, RI = 0.9908. Letters T, B, A, A1, and A2 are designated to clades as in Fig. 3.3. Names of taxa appear according to Table 3.1.

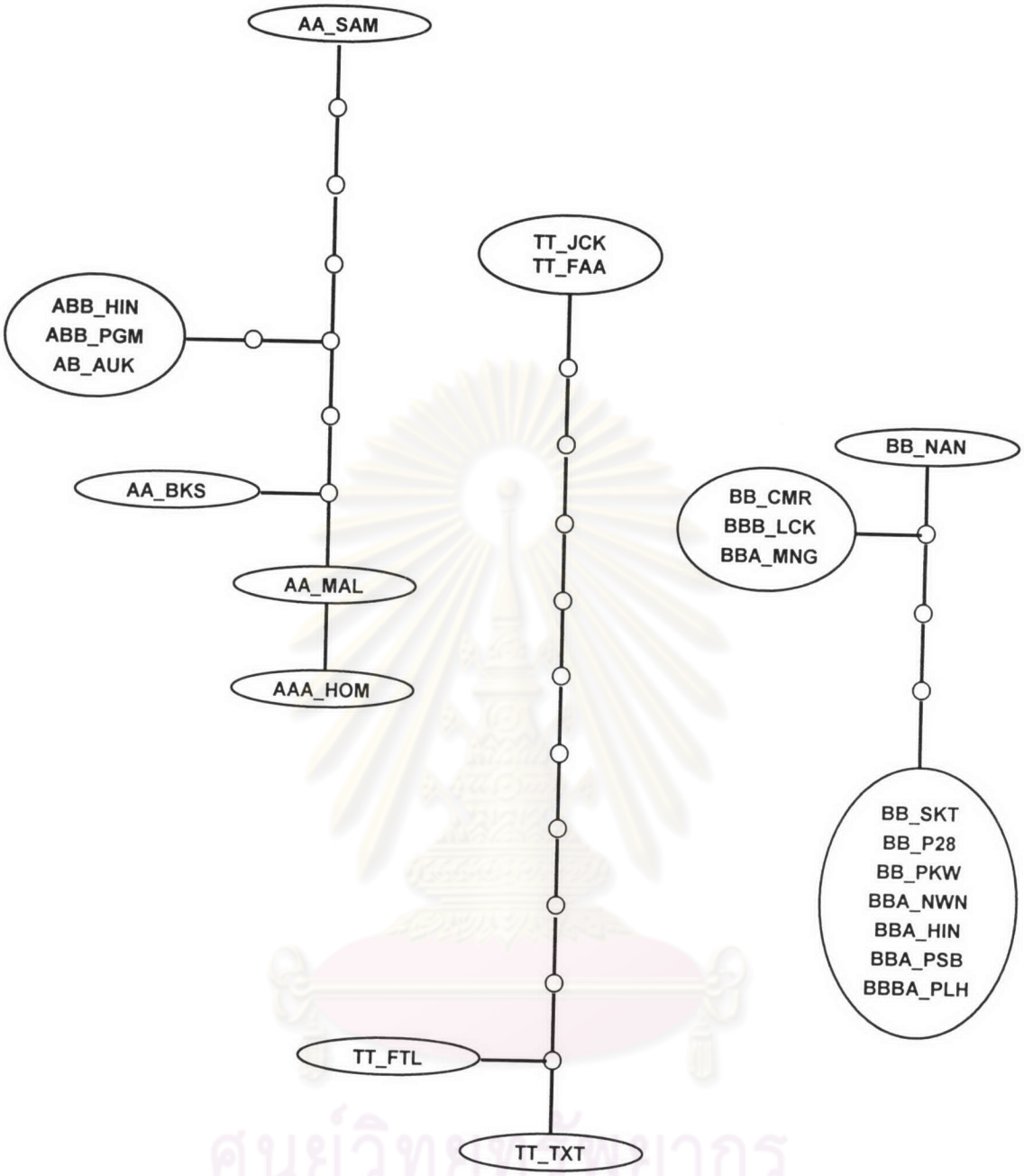
banksii, (ii) *M. acuminata* ssp. *malaccensis*, and (iii) *M. acuminata* ssp. *siamea.* As in the previous two analyses, 'HOM' formed a subclade with *M. acuminata* ssp. *malaccensis*, with bootstrap value of 75%. Also similar to the two analyses above, the hybrids formed an exclusive subclade with high bootstrap supported (88%). The subclade was sister to *M. acuminata* ssp. *siamea*.

Three separated networks representing A , B , and T clade were obtained from the statistical parsimony haplotype network analysis (Fig. 3.7). The network analysis showed the placement of members in each clade almost identical to the third method in MP analysis (i.e. with gaps excluded and recoded). It should be noted, however, that *M. acuminata* ssp. *siamea*, ssp. *banksii*, 'HOM', and *M. balbisiana* 'NAN' were placed at the tips of the network, which is probably due to several autapomorphic characters, including single nucleotide substitutions, found in these taxa (Fig. 3.7).

3.4 Discussion

3.4.1 Efficiency of PRS technique

PRS technique proved useful in the preliminary detection of diversity within the Eumusa section, especially to distinguish between B and A genomes where large size differences in sequence length were found. Without prior knowledge of the sequences, PRS analysis of DNA fragments differing in some nucleotides and length may be detected with highest efficiency using fragments around 100-400 bp (Orita et al. 1989; Jordan et al. 1998). Sato and Nishio (2002) reported that this method has a high ability in detecting one to six nucleotide changes in intraspecific



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Fig. 3.7 Statistical parsimony haplotype networks of cpDNA sequences. Accessions in the same haplotypes appear in ovals. Each mutational change represents as internode and small open circles indicate missing intermediates with >95% statistical parsimony support. The three networks are not connected due to large number of mutations. Accession codes appear according to those in Table 3.1.

cultivars of cabbage and rice. However, prediction models for polymorphic bands and suitable conditions for high resolution in SSCP detections are not well established (Jordan et al. 1998). The treatment conditions and selection of restriction enzymes may have affected resolution in our experiment where only five haplotypes were found. Nevertheless, cpDNA of *M. balbisiana* and M. *acuminata* are markedly different from each other that classification can be based on simple PRS technique with specific primers. Hypervariable regions of cpDNA that easily separated the two species can be found in *ndh*A intron where nucleotide length differences can be as many as 60 bp and in *pet*A-*psb*J-*psb*L spacer where more than 45-bp difference is common. In *psa*A, more than 20-bp length difference was found among A accessions.

3.4.2 Non-coding cp-sequence evolution in *Musa* cultivars

In our sequence matrix, several small indels and at least six large (27-132 bp) labile sites, e. g. at position 55-108, 2611-2706, and 2845-2856, contained, or flanked by mononucleotide repeats of adenosine/ thymine. These repeats are possibly involved in insertions and deletions by slipped-strand mispairing mechanism and frequently caused length variations in non-coding cp-sequences (Levinson and Gutman 1987; Morton 1995).

Visual examination of the indels indicated that most are simple direct repeats of adjacent or nearby sequences. However, it seems likely that, in several positions, the original or the repeat sequence has undergone subsequent evolution. For example, an insertion unique to *M. acuminata* ssp. *siamea* at 4453-4474 (GTAAT AGACA ATAGA CATAC AA) is homologous to the sequence at 4431-4452 shared by the Eumusa bananas and not that from the Australimusa bananas (GTAAT

AGGCA ATAGA CATAC AA). It seems that this insertion is a direct repeat recently occurred after the two sections have separated.

Short inverted repeats were found in the region of *pet*A in the cp-sequence, which is similar to those found in other plants, e.g. in *trn*K intron in *Fagopyrum* (Ohsako and Ohnishi 2000), *rpl*16 intron in some bamboos (Kelchner and Wendel 1996), *psb*A-*trn*H intergenic spacer in *Paeonia* (Sang et al. 1997) and *atp*β-*rbc*L intergenic spacer in epacrids (Ericales) (Crayn and Quinn 2000). The region which could cause hairpin secondary structure has been implicated in the occurrence of "hotspots" of length mutation in non-coding cpDNA (vom Stein and Hachtel 1988; Clegg et al. 1994; Gielly and Taberlet 1994).

3.4.3 Phylogenetic treatment of gaps

In our analyses of the Eumusa bananas and their relatives, we found indels to be particularly useful in resolving terminal clades in the topology and strengthening support for certain branches. For example, the analyses by nucleotide substitutions failed to distinguish low diversity within the B clade (Fig. 3.4). Inclusion of gaps as multistate characters in the analysis, however, resulted in splitting of the B's clade in half with moderately strong bootstrap value support (65%). Gaps may provide some genetic information about relationships that cannot be resolved by the former two analyses. Additional indels may provide important evidence of lineages e.g. insertion at 1649-1671, in the *psa*A locus, unique to *M. acuminata* ssp. *siamea* and AAA 'Hom Thong', hinted at the close relationship between them. On the other hand, the removal of indels would result in exclusion of phylogenetic information e. g. at 4897 where nucleotide substitutions have occurred within indels. However, it was cautioned that parallelism and reversal of length mutations occur at

relatively high rates in non-coding DNA comparing to coding DNA. Multiple coding of single indel events could add homoplasy to the data set (Golenberg 1993; Zhang 2000).

3.4.4 Phylogeny of *Musa* cultivars as inferred from cpDNA

The combined cpDNA data (Fig. 3.4-3.7) solidly supported the theory proposed by Cheesman (1947) and Simmonds and Shepherd (1955) that hybrid and polyploid cultivars came from two wild species, *M. balbisiana* and *M. acuminata*. The Eumusa clades formed a strong monophyletic clade indicated that they shared common ancestor. In addition, the cpDNA sequence data support the observed lower level of genetic variability within *M. balbisiana* compared to within *M. acuminata* (Simmonds 1962; De Langhe and De Maret 1999).

However, B clade was resolved into moderate bootstrap support (65%). This finding suggested that cpDNA of the B accessions varied but little, though, variability within *M. balbisiana* was reported previously (Shepherd 1990; Lebot et al. 1993; Sotto and Rabara 2000; Ude et al. 2002b). Nevertheless, the result connoted that cultivated *M. balbisiana* 'Tani' ('SKT') from the Sukhothai province may not be derived from wild *M. balbisiana* 'NAN'. Morphologically, these two accessions were distinct with light greenish, waxy pseudostem in the former and heavy purple-blotchy pseudostem in the latter. The presence of distinct *M. balbisiana* 'NAN' in the remote valley in northern Thailand indicated strongly that *M. balbisiana* is wild and native to Thailand. Associated in the same clade with the *M. balbisiana* 'NAN' (Fig. 3.6) were other diploid 'Pisang Klutuk Wulung' (PKW) from Indonesia and 'Cameroon' (CMR) whose geographic origin was doubtful. 'CMR'

accession may have been introduced into the international collection from Thailand and cultivated 'PKW' might have also arisen from Thailand.

Triploid hybrids, 'Hin', required special review. As inferred from phylogenetic analyses, the 'HIN' clones originated from two different maternal lineages. One collected from Yala Province in southernmost part of Thailand, was associated within the A clade, and another, originally collected from Nakhon Si Thammarat Province in central southern Thailand, within the B clade. The 'A Hin' seemingly descended from the same origin as of Indonesian 'Pisang Gajih Merah' (PGM) and Papua New Guinean 'Auko' (AUK). This incidence of diversity within the ABB, though was reported before in RFLP (Gawel and Jarret 1991a) and isozyme (Lebot et al. 1993) analyses, was not addressed clearly. To reflect their maternal origins, we, therefore, proposed that the two different 'Hin' cultivars, whose genomes were previously designated as ABB triploids, should be re-designated as BBA for 'B Hin' and ABB for 'A Hin'. This instance is also applicable for the 'Namwa' cultivar group (BBA) and 'Moe Nang' (BBA). Our cp-sequence data provided strong evidence for a better classification of these triploids and cautioned the selection of triploid hybrid materials for future genetic analyses, especially 'Saba' and 'Hin', which in each cultivar, inherited multiple maternal origins.

In A clade, there were two lineages and two sub-clades. 'Hom Thong', a triploid AAA cultivar, is likely to be a mutant of wild *M. acuminata* ssp. *malaccensis*. Due to different samples used, our report of close relationship between the Cavendish group and the Malay Peninsular subspecies agreed largely, but not completely, with previous reports

(Gawel and Jarret 1991a; Howell et al. 1994; Lebot et al. 1993; Bhat et al. 1994; Ude et al. 2002a, b).

CpDNA sequence of *M. acuminata* ssp. *siamea* was unusual, as to two unique single nucleotide substitutions and one large (22 bp) indel found in its sequence. These autapomorphic characters was detected by the MP analyses (9 characters, Fig. 3.6) and the TCS network (Fig. 3.7). The PRS pattern (AI, Table 3.3) of *M. acuminata* ssp. *siamea* shared by ABB 'Saba' from ITC and two 'Hak Muk' cultivars suggested that the Thai subspecies and these ABBs also shared common ancestor.

One of the two sub-clades belonged solely to the hybrid cultivars, including two Pacific-originated, 'Pisang Gajih Merah' and 'Auko', and Thai 'A Hin'. PRS pattern (AII, Table 3.3) of these hybrids were similarly found in ABB 'Pisang Abuperak' and two AABs 'Kofi' and 'Popoulou', which confirmed their common Pacific origin. *M. acuminata* ssp. *banksii* believed to be an ancestor of the Pacific AAB plantains (Lebot et al. 1993), was unlikely being so for these investigated Pacific hybrids. Though the maternal source species of their inherited genomes was not clearly indicated by our phylogenetic analyses, based on Fig 3.6, these hybrids were closer related to *M. acuminata* ssp. *siamea* than to other subspecies.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

3.4.5 Network analysis

In general, the network analysis of the *Musa* cultivars agreed mostly with the MP analyses. The networks (Fig. 3.4) are not connected, reflecting fixed differences between the three groups and lack of shared polymorphism (Posada and Crandall 2001). Applying the hypothesis from coalescent theory to the network, rare haplotypes (including

singletons) occur preferentially at the tips of the cladograms and more frequent haplotypes in the interior. In our case, cpDNA of *M. acuminata* ssp. *siamea* and ssp. *banksii*, as well as *M. balbisiana* 'NAN' contained more autapomorphic characters, thus possessed more mutational events from the common alleles. The network distinguished two related clades of *M. balbisiana*: (i) contained 'NAN', 'CMR', 'LCK', and 'MNG' and (ii) contained the rest of the clade. This result is congruent with MP analysis with gap excluded and recoded because of the similar treatment of informative gaps. This circumstance indicated two different origins of B cultivars, i.e. one in Thailand and another in the Southeast Asian Islands. In the A clade, the hybrid cultivars, AB and two ABBs, which clustered in one clade originated from common maternal lineage.

3.5 CONCLUSION

Our results indicated that allotriploid cultivars, ABBs, of Southeast Asia, originated from two distinct maternal lineages and should be re-designated differently as BBAs and ABBs. This discovery pronounces the knowledge of multiple maternal origins and complicate evolutionary history within the banana complex, which have been suspected previously.

The low diversity found in cpDNA is the indication of high functional constraint of the cp-genome (Golenberg 1993; Clegg et al. 1994). In the case of *Musa*, low cpDNA polymorphism may have caused by the common practice of vegetative propagation. Moreover, selection force imposed by domestication may act upon nuclear genome rather than on cp-genome. This low diversity might be a limit constraint to the study of their maternal lineages. However, DNA sequences, as a direct genetic

comparison between accessions possessed advantages among other techniques, e.g. RFLPs, AFLPs, and CAPs, especially to infer gene genealogies. The sequence analysis could overcome uncertainty in the B and A genome classification, which persisted in *Musa* research community for more than 50 years. In addition, it is rather promising for further polyploid and genome evolutionary studies. As more *Musa* DNA sequences accumulated and become freely available, their analyses would ease explanations of unsolved inquiries in different research fields, e.g. classification, evolution, hybridization, genetics, breeding, and others.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย