CHAPTER V

DISCUSSIONS

5.1 DNA extraction

Genomic DNA of 47 tobacco cultivars, comprising of 43 fresh-leaf and 24 cured-leaf samples and three roll-your-own (RYO) tobacco products, was successfully extracted with a commercial DNA extraction kit. The genomic DNA extraction of 39 tobacco cultivars from fresh-leaf samples mostly gave high quantity of the yielded DNA (Figures 4.1-4.3). However, some smear DNA also appeared with these 39 extracted DNA at the bottom of the electrophoretic agarose gels. Moreover, the DNA of four tobacco cultivars, K190, HB01, HBO04P and Pasak, (lanes 3, 8 and 9 in Figure 4.2; lane 1 in Figure 4.3, respectively) also appeared as fainted smear without genomic DNA bands. This smear DNA may be because of some degraded DNA and contaminated RNA. I suspected that the DNA degradation may have come from DNase (DNA endonuclease) within the plant cells which was activated from too much humidity still left in the sample tissues before performing the DNA extraction. In this thesis, the four cultivar specimens (K190, HB01, HBO04P and Pasak) gave low yield of the extracted DNA and their leaf samples seemed to have high humidity compared with those of the other cultivars. So, the high humidity within the leaf material of these four plants may have led to the low quantity of my extracted DNA because the DNase within the plant tissues may have been activated digested the DNA.

Four tobacco cultivars (K326, TN90, TN97 and Chorlare1 in lanes 4, 11, 12 and 15 of Figure 4.2, respectively) gave higher quantity of total genomic DNA and less amount of smear background than the other 35 fresh-leaf materials. The leaf samples of these four cultivars seemed to be drier than the other 35 cultivars. Although all of the fresh-leaf specimens were put into desiccating silica-gel bags to dry the plant tissues rapidly as suggested by Chase and Hills (1991), they were not brought back to the laboratory from crop fields immediately for extracting the genomic DNA. In fact, the DNA extraction should be performed from fresh young plant tissue harvested immediately

before the isolation to give the highest quality and concentration of the extracted genomic DNA (Michiels et al., 2003; Weising et al., 2005). The plant samples should be completely dried within 24 hours to ensure the highest DNA quality because the longer duration of the desiccation process, the lower quality of the isolated DNA. Moreover, water stress and humidity in plant cells also could affect inhibitor compounds like polyphenols (i.e., phenolic compounds) and other secondary metabolites within the plant samples to be increased and lead to low quality of the extracted DNA (Savolainen et al. 1995; Weising et al., 2005). The leaf specimens should also be cut to small pieces and kept in exceed 10:1 in the weight ratio of silica gel to specimens (Weising et al., 2005). However, in my study I put silica gel into the sample bag about 10:3 of the weight ratio and this small amount of silica gel seemed to be not enough to dry the leaf samples in the bag rapidly, especially in the case of some bags which had too many sample pieces. Therefore, in the future study only 2-3 pieces of the harvested plant specimens should be put into the sample bag and be immediately dried with an excessive amount of silica gel. Moreover, an indicating silica-gel which changes its colour from blue to pink or colourless when being humid should be additionally added into the sample bags.

On the other hand, the strong smear bands found within each lane of the electrophoresis gel may be an RNA still persisting after the DNA extraction. This RNA contamination may have happened because an RNase (RNA endonuclease) was not added into the reaction tube in the lysis step of the DNA extraction. I did not use RNase because RNA contamination should not have much effect in the PCR amplification. The contaminated RNA in the PCR reaction usually has much less chance to match with the specific primers than the DNA template. However, such RNA contamination could affect the sequencing step. Any residual RNA which still presents in the DNA template could compete with the DNA template in the sequencing method (Cook, 2007). Furthermore, the residual RNA may have the same effect as excess salt, which decreased the sequence signals and shortened the read lengths. Fortunately, almost all of my sequencing results did not show any problem of the whole sequence signals. Therefore, adding RNase in the lysis step of the DNA extraction is recommended for the future work to prevent any effect from RNA contamination.

In the case of the genomic DNA extraction from cured-leaf specimens (Figures 4.4-4.5) and roll-your-own (RYO) tobaccos (Figure 4.6), all of 24 extracted DNA from cured-leaf samples was in lower concentration than that of the fresh-leaf samples. This may be because the specimens had already passed a curing process (either flue-curing, air-curing, sun-curing or air-and-sun curing method) of which the various conditions such as heat (high temperature), humidity and the curing time may damage the plant DNA. For example, the heat and long period of the curing time may have a strong effect on the DNA denaturation to reduce the quantity and quality of the extracted genomic DNA. The extracted DNA of all Burley cultivars, three local cultivars (E-lueang, Hangkai and Yamueang) and all roll-your-own (RYO) tobacco products were in lower quantity than that of the others. All of the RYO tobacco products gave even less amount of the extracted DNA than Burley and the three local samples.

From the genomic DNA extraction of three RYO tobacco products (i.e. Maew, Mae-somsong (red package) and Mae-somsong (white package); Figure 4.6), their DNA quantity was less than that of the other cured-leaf tobacco samples. Normally, RYO tobacco products are made from local cultivars only. They are not only passed the air-and-sun curing process like the local cultivars, but they are also added some additive reagents, such as additive flavours ranging from light, regular, menthol to full flavour, during the blending step. These additive reagents could reduce the efficiency of the chemicals used in the genomic DNA extraction.

In the case of Burley cultivars which gave lower quantity than that of the other cultivars, except RYO, the low DNA quantity may be because the air-curing process for Burley leaves allowed much longer drying period (about four-to-eight weeks) than flue-curing, sun-curing and air-and-sun curing methods of Virginia, Turkish and local cultivars, respectively. Normally, the air-curing process is slowly carried on under virtually natural conditions in which the temperature, humidity, air flow and sunlight can be carefully controlled (Jeffrey, 1940; Jeffrey, 1946). The tobacco leaf which passes the curing method would have more humidity than the other methods. Thus, a high activity of DNase enzyme may occur in the plant tissues within the curing period and may cause more DNA digestion than the other curing methods.

For the DNA extraction results of Turkish and most of local cultivars which were in higher quantity than that of Burley cultivars, I assumed that the sun-curing and air-and-sun curing processes of these two cultivar groups might cause less damage to the genomic DNA within the plant tissues than the air-curing methods. Interestingly, the cured-leaf samples of Virginia cultivars gave the highest quantity of the extracted DNA among all cultivars studied, even though they were passed the flue-curing method which used very high heat (about 75°C) to remove moisture from the harvested leaves (Peele et al., 1995). However, the period for completion of the curing method is rather short, only five-to-seven days. Therefore, the flue-cured leaves may have lowest humidity than those cured with other curing methods and this could reduce the chance of the genomic DNA to be damaged by DNase.

Although the DNA extraction of some tobacco samples may have been affected by various parameters of the curing processes, the quality of the extracted DNA from almost all of the tobacco samples was acceptable for further PCR amplification. Almost all of the PCR reactions successfully generated the amplified products of the expected cpDNA regions. Therefore, the extracted DNA should be pure enough and its purity may have resulted from a high efficiency of the commercial DNA extraction kit used in this study. The handbook of the DNA extraction kit stated that it should give pure extracted genomic DNA suitable for being used in various molecular techniques, such as common PCR, real-time PCR, Southern blotting and RFLP. I therefore recommend using such genomic DNA extraction kit to extract plant genomic DNA for any experiment in the future. This extraction kit does not only provide a quick and easy method for purifying total DNA (including genomic DNA, mitochondrial and chloroplast DNA) from plant tissues, but it also avoids using toxic reagents, such as phenol and chloroform, and does not need any special laboratory equipments.

5.2 Selection of suitable PCR primer-pairs of highly-variable chloroplast DNA regions

The preliminarily PCR-primer screening showed that eight out of nine primer-pairs successfully amplified chloroplast regions of the extracted DNA of three tobacco cultivars which represented three cultivar groups. However, primer *pet*L-*psb*E was not able to amplify most of the DNA samples (lanes 1-3 in Figure 4.7). Only the DNA sample of B1 special cultivar (lane 3 in Figure 4.8) was possibly re-amplified with this primer. Although Shaw et al. (2007) used *pet*L-*psb*E primer pair to amplify the region in many angiosperm taxa, this primer pair failed to amplify with tobacco (*Nicotiana tabacum*) DNA samples in this thesis. Thus, the nucleotide sequences of the primers may not perfectly match with the DNA-template sequences of tobacco enough for successful amplification. Moreover, they also suggested that *pet*L-*psb*E region could not offer high levels of variation as much as the other eight regions. Therefore, this region may not be suitable for genetic-relationship studying among tobacco cultivars because of both of the difficulty to amplify the representative tobacco cultivars and its low PIC values.

Although the other eight regions were successfully amplified in all of the three DNA samples (Figures 4.7 and 4.8), *psb*D-*trn*T$^{(GGU)}$-R and 5′*rps*16x1-*trn*Q$^{(UUG)}$ primers unexpectedly produced several non-specific PCR products (lanes 4-6 and 10-12 in Figure 4.7, respectively). Roux (1995) suggested that using PCR technique to amplify a sample without an optimisation of the PCR conditions can lead to an amplification of multiple unwanted PCR products. To solve this problem one or more of PCR parameters that are known to contribute to primer-template fidelity and primer extension should be varied (Mullis and Faloona, 1987). High on the list of the optimisation variables are Mg^{2+} and dNTP concentrations, pH of buffer and cycling conditions. The situation is even more complicated by the fact that some of the variables are quite interdependent. For example, because dNTPs directly chelate a proportional number of Mg^{2+} ions, an increase in the concentration of dNTPs decreases the concentration of free Mg^{2+} available to influence polymerase function (Roux, 1995). Normally, Mg^{2+} concentration is one most favourable parameter to be manipulated because all variations of the concentration can be run in separate tubes simultaneously. Most suppliers of *Taq*

polymerase now provide $MgCl_2$ solution separated from the rest of the standard reaction buffer to simplify its adjustment. However, optimising the Mg^{2+} concentration is quite sensitive because an excess amount of Mg^{2+} in a PCR reaction would also result in the accumulation of non-specific PCR products (Gibbs, 1990).

With regard to all parameters in the cycling condition, an annealing temperature is the most important one (Gibbs, 1990). The annealing temperature in the "slow and cold" PCR condition suggested by Shaw et al. (2007) might not be best suitable for PCR amplification of these *psb*D-*trn*T and 5′*rps*16x1-*trn*Q regions. This PCR condition uses a fairly low annealing temperature (50°C) which could generate non-specific products. Roux (1995) suggested that adjusting PCR conditions by reducing the annealing temperature does not only increase sensitivity (i.e., more products), but it also increases the risk of non-specific amplification. Furthermore, the optimisation of annealing temperature was recommended to begin with a temperature increment (2-5°C) straddling a 5°C below the melting temperature (T_m) of the primer-template pairs (Roux, 1995). In this thesis, the PCR optimisation by increasing the annealing temperature from 50°C to 52°C decreased the non-specifically amplified products of both primer-pairs significantly (Figure 4.9). The melting temperatures of *psb*D-*trn*$T^{(GGU)}$-R primer-pair were 58.01°C and 55.75°C while those of 5′*rps*16x1-*trn*$Q^{(UUG)}$ were 54.48°C and 63.93°C. Thus, the melting temperatures of these two primer-pairs were quite closer to the optimised annealing temperature (52°C) than the normally low temperature (50°C), suggesting that an increment of the annealing temperature similar T_m could increase the template-primer binding specificity and reduce the non-product products in the PCR reaction.

Unexpectedly, although the 52°C annealing temperature can reduce non-specific products of *psb*D-*trn*T and 5′*rps*16x1-*trn*Q, the product size of these two regions from B1 special cultivar was incorrectly decreased from about 1350 bp to 1100 bp (lanes 15 and 18 in Figure 4.9). Therefore, the increasing of annealing temperature to 52°C may lead to the incorrect amplification of any other region of either the chloroplast DNA or the nuclear DNA. Fortunately, optimising the annealing temperature to 51°C did not give either the incorrect-size amplified products or any non-specific bands. The

annealing temperature of 51°C therefore should be preferable to be used for an amplification of tobacco DNA with the *psb*D-*trn*T$^{(GGU)}$-R and 5′*rps*16x1-*trn*Q$^{(UUG)}$ primers.

Among the eight primer pairs which successfully amplified the representative DNA samples, six cpDNA regions were successfully amplified and gave clear nucleotide sequences for further analysis with the extracted DNA of all tobacco samples. Only the sequencing results of 5'*rps*16-*trn*Q and 5'*trn*K-3'*rps*16 regions were failed, showing very high noise signals and many unreadable bases which was indicated as "N" throughout the electropherogram profiles. Cook (2007) suggested about the base-calling error in an electropherogram of the sequencing results that there are occasionally miscalls which are obvious to the human eye to recognise but difficult for a computer. As a simple rule, when the sequencing result reaches the point where there is one N per 20 bases, such sequence is not be usable (Cook, 2007). Therefore, 5'*rps*16-*trn*Q and 5'*trn*K-3'*rps*16 regions which had N more than 16 per 20 bases are not recommended for the tobacco genetic relationship study.

Although the annealing temperature was successfully optimised for the amplification of almost all of 62 DNA samples studied, some of the extracted DNA of cured-leaf samples and roll-your-own (RYO) tobacco products were failed to be amplified (Figures 4.20 and 4.21). Only *rpl*32F-*trn*L$^{(UAG)}$ primer-pair successfully amplified the extracted DNA of seven cured samples and two RYO products. This PCR problem of the cured-leaf and RYO tobacco samples may cause by a very low amount of the extracted DNA in the PCR reactions (discussed before). Moreover, various conditions of the curing process such as heat, humidity, cutting and blending might reduce the DNA quantity in the plant tissues (Savolainen et al. 1995; Weising et al., 2005). Furthermore, the presence of polyphenols and secondary metabolites such as tannins, alkaloids, phenolics and terpens increased within the dry plant tissues may interfere with the DNA isolation procedures and other reactions such as DNA restriction, PCR amplification and cloning (Sghaier and Mohammed, 2005).

5.3 The efficiency of the six selected highly-variable regions

Six highly-variable regions (*rpl*32-*trn*L, *pet*A-*psb*J, *ndh*F-*rpl*32, *ndh*C-*trn*V, *atp*I-*atp*H and *psb*D-*trn*T) chosen to be sequenced in this study had been reported to give high Potentially Informative Characters (PICs) (Shaw et al., 2007). These cpDNA markers have been increasingly utilised in plant genetic relationship study as they have more variation than other universal angiosperm primers (Shaw et al., 2007). Similarly, these regions appeared to be potential molecular markers to distinguish tobacco cultivars grown in Thailand. Their efficiency in this tobacco genetic relationship study was discussed below.

First, *rpl*32-*trn*L was found in this study to be the best choice among all six primer pairs for sequence polymorphism analysis. This region is an intergenic spacer in a small single-copy (SSC) region of the chloroplast genome (see Figure 3.3). In my work, *rpl*32-*trn*L is the most polymorphic region among all six selected regions with 10.47 percentages of variability and 75 PIC values. This region has been noted before as the best region surveyed for low-taxonomic-level molecular studies by Shaw et al. (2007), also having the highest PIC value (64 PICs). Timme et al. (2007) studied pairwise sequence divergence across 25 most divergent noncoding regions of cpDNA in *Helianthus* and *Lactuca* (Asteraceae) and also noted that *rpl*32-*trn*L region is the most highly variable. The PIC value of *rpl*32-*trn*L region of the tobacco (*Nicotiana tabacum*) cpDNA in my study was rather high compared to that of Shaw et al. (2007) which surveyed on as many as seven angiosperm lineages (*Magnolia*, *Prunus*, *Carphephorus*, *Trillium*, *Hibiscus*, *Gratiola* and *Minuartia*).

Recently, more publications have used *rpl*32-*trn*L region as a powerful tool to study molecular genetics, systematics, phylogenetics and identification in many plants. For example, Dunbar-Co et al. (2008) used *rpl*32-*trn*L sequence to study molecular phylogeny and adaptive radiation of the endemic Hawaiian *Plantago* species (Plantaginaceae), suggesting that the *Plantago* lineage was monophyletic and arose from a single long-distance dispersal event. Falchi et al. (2009) studied phylogeography of *Cistus creticus* L. and found that the high mutation rate of *rpl*32-*trn*L region could be explained by the ancient presence of this species, allowing the accumulation of

mutations. Another recent study using *rpl*32-*trn*L region has been performed by Kubota and Ohara (2009) to study the selfing evolution from outcrossing ancestors in a hermaphroditic perennial, *Trillium camschatcense* (Melanthiaceae), in which they proposed possible scenarios for the evolutionary backgrounds of the mating systems among *T. camschatcense* populations. In addition, Sosa et al. (2009) also used *rpl*32-*trn*L region to indicate the differentiation among population structure in *Hunnemannia fumariifolia* (Papaveraceae). Therefore, by now *rpl*32-*trn*L region should be the best cpDNA region to be used as a powerful tool for intraspecific genetic relationship studies, not only among tobacco cultivars, but also in various groups of plants.

Secondly, *pet*A-*psb*J was found in this study to be another highly polymorphic region, second only to *rpl*32-*trn*L region, with 26 PICs and 3.45 percentages of variability. The aligned sequence length of this region was 753 bp, similarly to the sequence length reported in the work of Shaw et al. (2007). *pet*A-*psb*J has also been used before in an intraspecific phylogeographic study of *Trochodendron aralioides* (Huang et al., 2004) to recognise nine haplotypes differentiated among populations of this species. In Thailand, *pet*A-*psb*J region has also been used to infer phylogenetic relationships among various groups of plants. For example, Ngamriabsakul and Techaprasan (2006) studied the molecular phylogeny of Thai *Boesenbergia* (Zingiberaceae) based on this region and showed that *Boesenbergia* species form a monophyletic clade in parsimony and UPGMA analyses. They also found an unpublished new species, *Boesenbergia bambusetorum* was grouped within the clade of the other two populations of *B. longiflora*. In addition, the hybrid evolution in bananas *Musa* (Musaceae) and the diversity assessment within each genome of *Musa* was studied by Swangpol et al. (2007) using *pet*A-*psb*J and three other cpDNA sequences. Although the sequence polymorphisms in this region were rather low in plants (i.e., only 2.57% informative characters in *Boesenbergia* and <2% sequence divergence in *Musa*), *pet*A-*psb*J region could still be suitable to study the differentiation between tobacco cultivar groups.

Interestingly, polymorphic single sequence repeats (SSR) or microsattellite DNA within *pet*A-*psb*J region has been reported in the comparison among closely related pines (Bucci et al., 1998). This SSR was then used in a related study to identify 100-

year-old herbarium specimens of *Pinus brutia* (Pinaceae) (DeCastro and Menale, 2004). Sebastiani et al. (2004) also implicated *pet*A-*psb*J region as a potentially useful microsatellite region in *Castanea* (Fagaceae). Beside of *Pinus* and *Castanea*, Shaw et al. (2007) observed such poly A/T running in all of the seven angiosperm lineages studied and confirmed the likely presence of such SSR region in many other plant taxa. These poly A/T runs were also observed within the *pet*A-*psb*J sequences of 23 tobaccos in my study, such as ATTATT at the 22, 176 and 356 bp aligned sites, TTATTTAT at 66 bp, ATTAATTA at 581 bp and AAATAAAT at 653 bp. However, such poly A/T repeats were found being in the same lengths within this region of all of the tobacco cultivars studied. So, it could not be applied to study genetic population and variation among cultivars of tobacco in any further study.

The third highly polymorphic region found in this study was *ndh*F-*rpl*32 which is adjacent to *rpl*32-*trn*L region. This region shows high variation within its sequences with 2.99% variability and 23 PICs and it has been noted as highly variable by Timme et al. (2007). Although their result in *Helianthus* and *Lactuca* sequence analysis showed that *ndh*F-*rpl*32 region had less *p*-distance (proportion of basepairs differing between two sequences) value than *rpl*32-*trn*L region, they still proposed that the genetic variation in this region was high enough for using in a species-level study of plants. Moreover, Shaw et al. (2007) also suggested that this region is among the best choices for low-level molecular investigation, especially if coamplified with *rpl*32-*trn*L (as *ndh*F-*rpl*32-*trn*L) to give 2 kb of the potentially more informative region. Therefore, *ndh*F-*rpl*32 region has been widely coamplified with *rpl*32-*trn*L region to study molecular phylogeny and phylogeographic analysis of many plant species. For example, Dunbar-Co et al. (2008) used *ndh*F-*rpl*32 and *rpl*32-*trn*L sequences to analyse the phylogenetic tree of the endemic Hawaiian *Plantago* species (Plantaginaceae), suggesting the rapid phenotypic diversification of major lineages in this group. In addition, Kubota and Ohara (2009) also used *ndh*F-*rpl*32 together with *rpl*32-*trn*L to investigate the evolution of outcrossing ancestors of *Trillium camschatcense* (Melanthiaceae) based on the reproductive and geographical features of the mating systems. Another study of *ndh*F-*rpl*32 and *rpl*32-*trn*L regions was to phylogeographically study the endemic population of Mexican tulip poppy *Hunnemannia fumariifolia* (Papaveraceae) in the Sierra Madre Oriental (Sosa et

al., 2009). Therefore, although *ndh*F-*rpl*32 region alone had less PIC value than *rpl*32-*trn*L and *pet*A-*psb*J regions in this study of tobacco, this region could be another effective molecular marker after coamplified with *rpl*32-*trn*L region.

Next, *ndh*C-*trn*V was found being a moderately good marker because it gave 2.63 variability percentages and 18 PICs. Takahashi et al. (2005) firstly used this region to analyse genetic variation among *Saccharum* species (Poaceae). This region was also noted as a highly variable region by Timme et al. (2007), showing that *ndh*C-*trn*V region has moderately high *p*-distance value of 0.130. Interestingly, Shaw et al. (2007) observed several small indels (6-56 bp) in *ndh*C-*trn*V region. Likewise, a 41 bp deletion was discovered in this region of three wild potato species, *Solanum chiquidenum*, *S. chomatophilum* and *S. jalcae* (Ames et al., 2007) and this deletion was proposed to be a marker for *Solanum* (Solanaceae). In fact, insertions and deletions (indels) are common in intergenic spacer regions of chloroplast DNA and they can provide important phylogenetic characters for closely related species. Although my study found only two small insertions (8 and 7 bp) within *ndh*C-*trn*V region of tobacco, they were still be useful for distinguishing tobacco cultivar groups, like the other three regions before.

*atp*I-*atp*H intergenic spacer of tobacco chloroplast DNA was found in my study to have less variation (1.48% variability and 12 PICs) than the other regions above. Shaw et al. (2007) reported the average variability of 46% and average 42 PICs from *atp*I-*atp*H region across seven angiosperm lineages and also observed poly A/T running in the region, similar to *pet*A-*psb*J region. This poly A/T was found being rather long (24 bp) and could cause alignment problems during sequencing. Provan et al. (2004) successfully developed such microsettlelites or SSRs of *atp*I-*atp*H region to reveal intraspecific variation of grasses and cereals (Poaceae). These chloroplast SSRs could be a valuable tool to study high-resolution cytoplasmic diversity in important range of plant taxa. In my study, some poly A/T repeats were also observed within the alignment of *atp*I-*atp*H region, but they were quite short, not differentiated among tobacco cultivars and did not cause such sequencing problem. Therefore, *atp*I-*atp*H region could only be used to distinguish tobacco cultivar-groups like the other regions, but can not be developed as cpSSR marker for studying genetic relationship of tobacco cultivars.

The least variable region in my study was *psb*D-*trn*T which had only 1% variability and 7 PIC values in the sequences. Although *psb*D-*trn*T revealed the lowest polymorphism among all six regions studies, it was suggested as being highly variable by Shaw et al. (2007). They reported that *psb*D-*trn*T gave high variation with 36% variability and 49 PICs after compared among seven angiosperm lineages. Ronikier et al. (2008) conducted an experiment using sequence-data of this region to elucidate the phylogeography of *Campanula alpine* (Campanulaceae) and revealed congruent groups of distinct cpDNA haplotypes which supported a phylogeographical AFLP pattern. Additionally, *psb*D-*trn*T region has also been used to study the evolution of selfings from outcrossing ancestors of *Trillium camschatcense* (Melanthiaceae) by Kubota and Ohara (2009). Although several previous studies successfully used *psb*D-*trn*T region to analyse the population genetic variations in some plants, I would recommend not to use this region to study genetic variation of tobacco cultivars since it gave the least polymorphism sequences among tobacco cultivars in my study.

Even though the six highly-variable noncoding regions had different levels of the efficiency for genetic relationship study, all six regions could be used as molecular markers to differentiate tobacco cultivar groups with the various degrees of polymorphism in nucleotide sequence characteristics. For example, although *psb*D-*trn*T region has least sequence polymorphism among tobacco cultivars than the other regions, it could distinguish Virginia tobacco cultivars from the other cultivars. On the other way round, *rpl*32-*trn*L was both the best region for genetic relationship study and the best marker to differentiate all tobacco cultivars to three groups: Virginia group, special-seven-local group and the group of the other cultivars leaf. Moreover, an opportunity to implement other techniques, such as the coamplification of *ndh*F-*rpl*32-*trn*L region, to improve genetic relationship study of tobacco cultivars in future should be investigated.

5.4 Genetic relationships of tobacco cultivars in Thailand

From the six aligned sequence results of the cpDNA of 23 tobacco cultivars, the degrees of polymorphism in nucleotide sequence characteristics among different tobacco cultivars in sequence lengths were from 1 to 66 bp differences and the amounts of base substitutions were from 0.37% to 1.26%. These six regions distinguished two tobacco cultivars in Virginia cultivar-group (K326 and PVH03) from the others. The distinctive sequences of Virginia cultivars found in this study agreed well with the previous AFLP study of Siva et al. (2008) which could separate flue-cued (Virginia) cultivars from air-cued (Burley) cultivars. Although the AFLP genetic polymorphism presenting among Virginia cultivars of tobacco was low (15.35%), they suggested that the markers found specific to Virginia cultivars could be used in an identification of the genotypes in trade and commerce. In addition, Sarala and Rao (2008) could distinguish eight flue-cued cultivars from two Burley cultivars by using RAPD method. They reported that the genetic similarity among flue-cured Virginia cultivars found to be higher (0.70) than that of Burley (0.65).

When the total 51 tobacco cultivars were analysed by using *rpl*32-*trn*L region, the nucleotide sequences of this region could strongly group three Virginia cultivars (K326, PVH03 and PV09) together with 100% bootstrap on the NJ tree (Figure 4.23). This K326 cultivar was ensured to be Virginia cultivar by the officers of Lamphun tobacco station. They confirmed that the K326 plant materials collected in the field were grown from 30-day old seedlings which they germinated and distributed to local farmers supported by Thailand Tobacco Monopoly. PVH03 and PV09 cultivars were also surely identified as in Virginia cultivar group since their cultivation was strictly controlled by the private leaf-processing companies. Therefore, the uniqueness in sequence characteristics of these three cultivars could be implemented as sequence markers specific to Virginia cultivar-group.

However, the aligned sequences of *rpl*32-*trn*L region showed that the sequences of three Burley cultivars (HB01, HBO04P and TN97) were also identical to those of the Virginia cultivars (K326, PVH03 and PV09). This misplacing of some Burley samples within the Virginia group led to the question of whether they were really Burley

cultivars. Although the seedlings of these three cultivars were collected from a greenhouse of Meajo Experiment Station, the officers were not completely sure about the history of cultivation of these cultivars. The similar misidentification may have occurred with K187, K190 and Coker326 Virginia cultivars of which the nucleotide sequences were also identical to Burley cultivars. Likewise, the regional officers were not able to judge whether these K187, K190 and Coker326 were of Burley or Virginia cultivar-group. Thus, K187, K190 and Coker326 tobacco cultivars may actually be Burley cultivars which were misidentified while HB01, HBO04P and TN97 should be Virginia cultivars.

In the case of Turkish cultivars, Samsun and Xanthiyaka, all of their sequences from *rpl*32-*trn*L region were the same as those of Burley and were grouped together on the cluster I of the NJ tree (Figure 4.23). The sequence results revealed a close relationship between Burley and Turkish cultivar groups. This finding was similar to the work of Yang et al. (2007) who used ISSR and IRAP markers to determine the genetic diversity of tobacco germplasms. Their result from the genetic distance analysis suggested that there was low genetic diversity within and between different cultivar groups. Moreover, their phylogeny result revealed the grouping between Burley cultivar group and oriental (Turkish) group. Therefore, Turkish cultivar group may have the same origin of germplasms with Burley group because the genetic polymorphism between the two tobacco cultivar groups was very low. This hypothesis was not totally agreed with a common believe of the officers of Thailand Tobacco Monopoly since there are many phenotypic variations among the two cultivar groups which could be classified by different chemical characteristics and the methods of leaf curing (Akehurst, 1981; Goodman, 1993). However, the differences between both cultivar groups in morphological and chemical characteristics may have come from their different cultivation approaches. For example, Turkish cultivars have been generally grown on poor soil and stressful climatic condition and then would have led them being very small, as well as having unique aromatic property (Akehurst, 1981; Davis and Nielsen, 1999). On the other hand, Burley cultivars which have cultivated on much more fertile locations all over the world commonly have larger leaves with sweeter flavour (Davis and Nielsen, 1999).

Not only most of Burley and Turkish cultivars were grouped together on the NJ tree (Figure 4.23), but most of local cultivars were also clustered with this Burley-and-Turkish group. This result was similar to the previous study reported by Setaphan (2007) using ISSR markers to study genetic relationship of 40 tobacco cultivars grown in Thailand. She suggested that almost all of local cultivars were grouped together with some Burley and Turkish imported cultivars. Thus, both ISSR and sequencing results suggested that some of local tobacco cultivars may have been descended from Burley and/or Turkish imported cultivars. This could have happened because some Burley and Turkish cultivars have been imported by Thailand Tobacco Monopoly to be promoted and given tobacco farmers for cultivation since 50 years ago (ปราโมทย์, 2524; วรวิชย์ และคณะ, 2549). Then, the farmers may have kept tobacco seed by themselves for next cultivating season if the promoted tobacco produced high yield of tobacco leaves. This genetic relationship between some local and imported cultivars also agreed well with the fact that some local cultivars were morphologically similar to some imported cultivars. Therefore, some local cultivars were probably not truly local cultivars, although they have their own Thai names and have been cultivated in local areas for a long time. Moreover, these local tobacco specimens should be pronounced as "imported cultivars" instead.

Interestingly, seven local cultivars (Hangkai, Kan, Petmakhuea, Ubon Ratchathani, Petkhangsink, Baitung, Bailai) and one RYO tobacco product (Mae-somsong) were different from the other 22 local cultivars by having a large 66-bp insertion within *rpl*32-*trn*L sequences. Such special insertion of these cultivars was not found in the other five regions in this study. This large nucleotide insertion may significantly indicate the genetic differences between these special-local cultivars and other local cultivars in both genotypic origin and the history of import. As a result, these seven-special-local cultivars may have been imported into Thailand for a longer period of time than other local cultivars. Unfortunately, the evolutionary or substitution rate of *rpl*32-*trn*L region has never been studied before. Therefore, I cannot estimate the time when they were firstly imported into the country.

Some authors have previously addressed the issue of indel frequencies and the relative rates of indel mutations in noncoding cpDNA sequences, even though several

conflicting hypotheses have been put forward. While Gielly and Taberlet (1994) suggested that indels in noncoding cpDNA occur with nearly the same frequency as nucleotide substitutions, Clegg et al. (1994) and Britten et al. (2003) argued that the indels may occur more frequently than the nucleotide substitutions. Nevertheless, all of them demonstrated that indels within nucleotide sequences are thought to be a major driving force in the evolution and molecular clock of noncoding cpDNA sequences. For example, Yamane et al. (2006) analysed the entire chloroplast genome sequence from sugarcane, maize and rice, and successfully calculated the average rate of indel mutation in the intergenic regions of chloroplast genomes to be $0.8\pm0.04\times10^{-9}$ per site per year. This estimated rate of indel mutation was as slow as previous estimates reported in other organisms (e.g. primates; Saitou and Ueda, 1994). Although the indel mutation rate of Yamane et al. (2006) was not based on the chloroplast genome of tobacco, I calculated a putative divergence time for the seven-special tobacco cultivars and other cultivars to be approximately 1.0~1.2 million years. This putative divergence time seems to be too much because it should be only in the range of thousand years, considering the history of tobacco cultivation. Nevertheless, the accumulation of the genetic variation within *rpl*32-*trn*L sequences of these seven special-local cultivars could be considered as have been taking for a very long time. According to the legal definition, only the tobacco cultivars which have been cultivated in Thailand for a long time can be identified as "local cultivars" (วรวิชย์ และคณะ, 2549). Therefore, I proposed that these special-seven-local cultivars presumably had a long-history of growing in Thailand and they could be legally pronounced as "true" local cultivars.

5.5 Preliminary experiment for multiplex PCR

In the previous experiment, *rpl*32-*trn*L region was successfully used as the best molecular marker showing different sequencing signals in the combined electropherogram between Virginia and other cultivars, with 5 bp substitutions of Virginia specific-marker on the forward sequencing result (Figure 4.24). This sequence marker could be a very efficient tool to identify any Virginia cultivar which may be mixed with other tobacco cultivars. For example, such Virginia-specific sequence marker could be

applied to investigate an illegal mixing of Virginia tobacco into roll-your-own (RYO) tobacco products, which must be made only from local tobacco cultivars by law. If any RYO specimen is sent to be analysed using such Virginia sequence marker and then produces an electropherogram similar to the 5 combined sequence signals above, the tested sample may have been illegally mixed with Virginia tobacco. In spite of the benefit of using this 5-bp sequence marker, the marker also has a disadvantage that it could not be developed as two-separated bands on a common agarose gel electrophoresis. Only the nucleotide substitution polymorphism was found between *rpl*32-*trn*L sequences of Virginia and other tobacco cultivars studied without any indel polymorphism.

Fortunately, the large 66-bp insertion within *rpl*32-*trn*L sequences of some special-local cultivars could be used as a specific marker in both sequencing reaction and gel separation. The combined PCR products of Virginia cultivars and those special-local tobaccos could successfully be separated, showing as two not-so-clear but distinguishable bands on 1.8% agarose gel (lane 2 in Figure 4.28). He et al. (2007) compared different concentration of common agarose gel matrices to resolve PCR-amplified DNA fingerprint profiles. They suggested that the 0.5-2% concentration typically used for a large range of separation, from 100 bp to 23 kb, could give only relatively low-resolution results after compared with some high-resolution agarose gels. Likewise, my 1.8% common agarose gel could only give the low-resolution result of two different PCR fragments of the tobacco cultivars. Thus, changing the electrophoresis medium to a high-resolution agarose gel should increase sharpness of the low-resolution PCR bands.

Several high-resolution agarose gels, for example Nusieve 3:1 agarose and Metaphor agarose, have been used to achieve better electrophoresis results than common agarose gel. Nusieve 3:1 agarose could ensure fine resolution of DNA fragments from 10 bp upto 1500 bp whereas Metaphor agarose could resolve DNA fragments in the range of 200 to 800 bp (He et al., 2007). Nagaraju et al. (2002) successfully used Nusieve 3:1 agarose mixed with Sigma agarose to resolve FISSR-PCR (Fluorescence Inter-Simple Sequence Repeat PCR) markers of F_2 offspring derived from a cross of two divergent silkworm strains. Their result also showed that using Nusieve

agarose mixed for FISSR-PCR marker assay could be a method of choice for large scale screening of varieties/cultivars and highthroughput genotyping in mapping of genomes where microsatellite information is scanty or absent. In the case of Metaphor agarose, He et al. (2007) compared the resolution of Metaphore with polyacrylamide gel and vinyl-polymer of polyacrylamide. They found that the resolution of Metaphore gel and polyacrylamide gel was not as high as that of vinyl-polymer of polyacrylamide gel which was much sharper and tighter.

From the band separation tests with Nusieve 3:1 and Metaphor, both special gels helped resolving the two unclear PCR bands of Virginia and special-local cultivars. However, both of the high resolution gels were uneasily prepared, since the gels took much longer time for gel solidification, also being too soft and uneasy to handle, especially in the case of Metaphor. Moreover, the PCR fragments were mobilised through both high-resolution gels two-to-three time slower than through a common agarose gel. Furthermore, the costs of Nusieve 3:1 and Metaphor gels were approximately four-fold and five-fold more expensive than common agarose gel, respectively (Lab Focus; the supplier). Interestingly, the two PCR bands could be better separated in the 1.8% common agarose gel after two-hour running than being separate in Nusieve 3:1 and MetaPhor. This may be because the two high-resolution gels need much longer running time that could make the gel be partially melted and the PCR bands may be leaked to the buffer solution. Therefore, using a 1.8% normal agarose gel for the two *rpl*32-*trn*L band separation would be the best recommendation to gain an easy gel-preparation and to save the analysing cost.

From the result of the multiplex PCR experiment, *ndh*F-*rpl*32 region was successfully co-amplified with *rpl*32-*trn*L of different tobacco cultivars. Since *ndh*F-*rpl*32 region could also give the 19-bp insertion specific for Virginia cultivar group (lanes 1 and 4 in Figure 4.30) and this indel marker was not in the same length range of *rpl*32-*trn*L fragments (830-850 bp vs 1270-1200 bp, respectively), the multiplex PCR co-amplification could nicely show four separated PCR bands on the gel (lane 3 in Figure 4.31). Some researchers have introduced multiplex PCR approaches to rapidly identify variants of organisms instead of using common PCR and DNA sequencing. For example, Cousins et al. (1996) developed multiplex PCR strategy to detect and identify

multiple mycobacterial of *Mycobacterium avium* and *M. intracellulare*. More recently, Oliveira and de Lencastre (2002) designed two primer-pairs for using in multiplex PCR to rapid identify structural types and variants of the *mec* element in methicillin-resistant *Staphylococcus aureus* (MRSA).

The successfully multiplex PCR co-amplification of *ndh*F-*rpl*32 and *rpl*32-*trn*L regions in this thesis could effectively reduce the time and cost of Virginia cultivar analyses in the large scale. Such multiplex PCR technique is much easier to perform and take less time than the DNA sequencing-marker approach. The multiplex PCR may give a simple gel result within only about 5 hrs for PCR and electrophoresis experiments whereas the DNA sequencing method must spend more than 3-day period of time. Moreover, the DNA sequencing technique would have much higher cost than the multiplex PCR, additionally needing 5-US-dollar payment per reaction for the sequencing service (Macrogen Inc., Korea). In practical, this multiplex *ndh*F-*rpl*32 and *rpl*32-*trn*L marker could be easily identify unknown tobacco leaf sample whether it is Virginia cultivar, or even investigate a RYO tobacco product whether it has been illegally mixed with Virginia cultivar. Moreover, this technique may be further developed for tobacco seed samples which would rather be more difficult to be analysed than leaf specimens because the tobacco seed is very small and then contains less genetic information from cpDNA than fully-developed leaf specimens. Another usage of this multiplex-specific marker would be to confirm the genotypic correctness of Virginia tobacco cultivars, such as K326 cultivar, which have been grown in many locations in Thailand from the promotion of Thailand Tobacco Monopoly. This genotypic confirmation would then help strengthening the quality control of tobacco cultivation and curing in the cigarette production industry.