Proceedings of the Pakistan Academy of Sciences:
B. Life and Environmental Sciences 55 (3): 41–48 (2018)

ISSN: 2518-4261 (print), ISSN 2518-427X (online)

Pakistan Academy of Sciences

Research Article

Exploitation of Polymorphic Sequences in Chloroplast Genome for Identification of Commercial Tobacco Cultivars

Muhammad Ramzan Khan[1,2*], Wajya Ajmal[2], Obaid Ur Rehman[1], Muhammad Aqeel[2], and Ghulam Muhammad Ali[2]

[1]National Centre for Bioinformatics, Quaid-i-Azam University, Islamabad-45500, Pakistan
[2]National Institute for Genomics and Advanced Biotechnology, National Agricultural Research Centre (NARC), Park Road, Islamabad-45500, Pakistan

Abstract: Tobacco is an important commercial commodity. Discrimination of tobacco cultivars at DNA level is extremely important for selling of brand, free of mixing. Chloroplast markers have sufficient resolving power for differentiation at intra-varietal level. In this study only chloroplast genome markers were used to amplify polymorphic regions encoding genes involved in photosynthetic machinery of plants. This region encompassing 9000 to 9600bp harbors SNPs, Indels and SSRs. Different bioinformatics and phylogenetic software were used to investigate relationship of DNA sequences of 11 unknown plant samples with a reference sequence from NCBI. Multiple alignments exhibited sequence conservation encompassing a substantial region. But it also revealed important variety specific SNPs. Sequence similarity index was generated for grouping of unknown plants. The unidentified plants were differentiated into different clusters as revealed by phylogenetic analysis. Furthermore, a neighbor-net network analysis validated the results of Neighbor Joining tree. Our analysis indicates that there exist at least 5 varieties of tobacco. The V1 is the most distant cultivar. V5 is also separated from the rest of cultivars. But V2, V3 and reference sequence make further linage, and can be considered as separate clade. A bigger cluster including V7, S2 and V6 is differentiated into distinct group. The S1, V9, V10 and V8 are the 5th cluster that can be considered as a single variety. Hence there were 5 varieties in total that were mixed in 11 samples.

Keywords: Tobacco, Varietal identification, Chloroplast markers, DNA fingerprinting

1. INTRODUCTION

Tobacco (*Nicotiana tabacum L.*) is one of the most significant commodities in agriculture. Across the world 33 million people depend on tobacco cultivation for their income (International Tobacco Growers Association (ITGA), 1996). The use of tobacco commercial varieties is important for brand selling. Mixing of elite varieties with ordinary may lead to non-acceptable taste and devalue the brand. This leads to economic losses. Therefore, use of pure and authentic varieties is inevitable for further propagation [1,2]. Reliable variety identification is much needed approach in many cultivars. Traditionally, this was achieved by utilizing morphological and biochemical markers but nowadays DNA markers have gained much popularity [3]. There are numerous benefits of employing molecular markers over other options. They are highly variable and co-dominantly inherited. They are easily visible and distributed uniformly over the entire genome. They are constant, economical and simple to use. Usually they need minor quantity of DNA and no prior information related to genome is required [4]. Recently, Yan et al. [5] has reviewed the applications and limitations of universal CpDNA markers in discriminating East Asian evergreen oaks. Sequencing of entire chloroplast genome has unveiled many polymorphic regions that can be selected for varietal identification through PCR amplification and eventually sequencing. These include SNPs, Indels and SSRs in the region between 9000 to 9600bp.

Authentication of tobacco cultivars is an important issue to be addressed to protect the interests of farmers as well as to provide quality

Received: May 2018; Accepted: September 2018
* Corresponding Author: Muhammad Ramzan Khan; drmrkhan_nigab@yahoo.com

product to consumers. Hence, major objective of this study was to discriminate and identify tobacco cultivars of commercial importance. In our study polymorphic regions of chloroplast genome were selected to amplify genes related to energy metabolism and photosynthetic machinery of plants [6]. We used chloroplast markers for the identification/authentication of 11 unknown samples of commercial tobacco cultivars. PCR amplification, sequencing, multiple alignment and phylogenetic reconstruction demonstrate that there exist at least 5 commercial varieties of tobacco that were mixed together. It also indicated evolutionary relationship among samples thus leading to the identification of tobacco samples.

2. MATERIALS AND METHODS

2.1. Plant Material Selection

Seeds of 13 commercial varieties were obtained from a local commercial tobacco company. These were designated as S1, S2, S3, S4, V1, V2, V3, V5, V6, V7, V8, V9 and V10. Plants were grown in the field of National Institute of Genomics and Advanced Biotechnology, National Agricultural Research Centre, Islamabad under suitable conditions. Plant materials (leaves) were harvested for DNA extraction.

2.2. DNA Extraction

Genomics DNA was extracted from 13 fresh tobacco leaf samples using CTAB method [7], and confirmed by running in gel electrophoresis. DNA samples were quantified using Nanodrop spectrophotometer (Thermo Scientific, Inc.) to check the quality as well as the quantity of extracted DNA.

2.3. Primer Designing

The chloroplast genome of tobacco was retrieved from NCBI database (https://www.ncbi.nlm.nih.gov/nucleotide/). The primers were designed manually by selecting the most polymorphic region in the chloroplast genome of tobacco. Primer sequences were forward primer 5'-GAGAAAGAGCTTCATTGCTTGGTGT-3' and reverse primer 5'-CCGGCTGGGTACTGACCAGACCAG-3'.

2.4. PCR Amplification

DNA samples were diluted and subjected to thermocycler (Applied Biosystems Inc) with primers for the amplification of the desired fragments. Each PCR reaction (50 µl) contained 10 ng DNA template, 5 µl 10 X reaction buffer, 5 µl MgCl2, 1 µl dNTPs, 1 µl of each primer, and 0.5 µl of Taq DNA polymerase (Promega, Madison, WI, USA). The PCR conditions set for amplification were 95°C for 5 min to activate taq polymerase, followed by 36 cycles of 94°C for 1 min (denaturation), 62°C for 1 min (annealing), 68°C for 1 min (extension) and a final 68°C for 5 min followed by 12°C. The PCR products were run on 1.5% agarose gel and confirmed.

2.5. Sequence Analysis, Multiple Alignment SNP Detection

After getting the results from PCR amplification and subsequent purification, 11 samples were sent to Macrogen (Korea) for sequencing. The sequence files obtained were edited and analyzed with Bio-Edit software [8]. Blastn was executed for target identification in NCBI database (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The sequences were trimmed from both ends to remove any mismatched or flanking regions using the BioEdit software [8]. The resulting sequences were edited and analyzed with different tools. Sequence identity matrix was calculated. For this purpose the multiple alignment was loaded in ClustalX; output tree format option was chosen and % identity matrix was selected. Finally, a table was generated.

In order to observe the sequence conservation, multiple sequence alignment was performed using ClustalW program [9]. SNP (single nucleotide polymorphism) analysis of tobacco cultivars was carried out and a table was generated using BioEdit software [8].

2.6. Phylogenetic Analysis

Phylogenetic reconstruction was executed on the aligned sequences using Neighbor Joining method in MEGA 6.0 software [10] for finding the evolutionary connection among different samples. The evolutionary relationship, recombination, hybridization and homoplasmy are well exhibited

by phylogenetic network instead of a tree like illustration [10]. In this regard, a neighbor-net network algorithm was employed in SplitTree4 software [11] using default parameters.

3. RESULTS

3.1. Divergence in Amplified Sequences is Detectable among Different Tobacco Samples

In order to amplify DNA fragment, PCR was employed using standard conditions. Fig. 1 illustrates that it is possible to amplify a fragment of 650 bp from 11 out of 13 samples of tobacco. Two samples S3 and S4 could not be amplified mainly due to poor DNA concentration as leaves became already withered and dried. The band intensity or quantity and quality of amplified PCR products were high enough to be directly sequenced without purification step. Blastn search revealed that all the hits were targeting the tobacco chloroplast genome, thus indicating the authenticity of targets.

The target sequences were analyzed using Bio-Edit program. In order to detect the sequence conservation among different cultivars trimmed sequences were subjected to multiple alignments using Clustal W tool (Fig. 2). The alignments revealed conservation among all the sequences but different mutations including substitutions and even small deletions at the 5' and the 3' were also detected. But this may be due to selection of sequence length. Anyhow substitution mutations were most predominant. Table 1 shows that

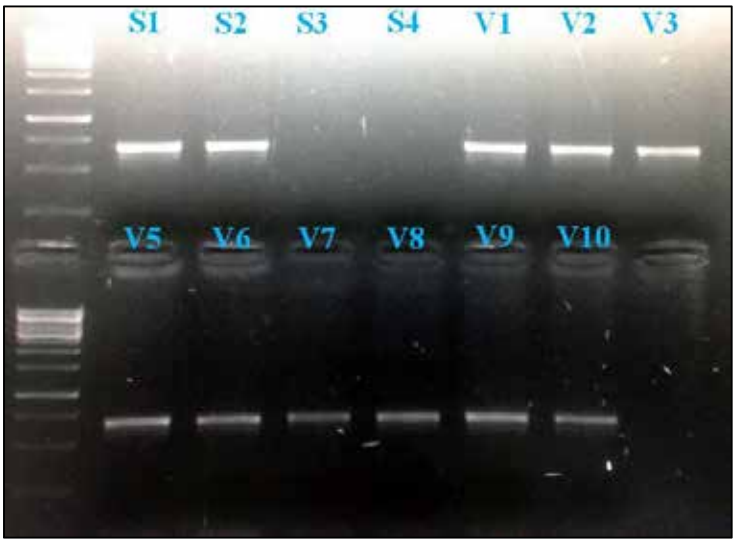

Fig. 1. Amplified PCR products have been labeled. 'M' denotes the 1kb ladder. S3 and S4 bands are missing because their DNA was degraded. Rest of the samples show clear bands

Table 1: Sequence identity matrix calculated using Bio-Edit software

Seq	S1	S2	V1	V2	V3	V5	V6	V7	V8	V9	V10	NCBI_ Z00044.2
S1	ID	98.8	90.4	98.3	98.3	98.6	99.3	99.0	99.5	99.0	99.1	98.1
S2	98.8	ID	90.9	99.1	98.5	98.5	99.5	99.5	99.0	99.1	99.0	98.0
V1	90.4	90.9	ID	91.0	90.4	91.0	90.7	90.7	89.9	90.0	89.9	89.5
V2	98.3	99.1	91.0	ID	99.0	97.6	99.0	98.6	98.1	98.3	98.1	98.5
V3	98.3	98.5	90.4	99.0	ID	97.6	98.3	98.3	98.5	98.6	98.5	98.8
V5	98.6	98.5	91.0	97.6	97.6	ID	98.6	98.6	98.5	98.3	98.5	97.5
V6	99.3	99.5	90.7	99.0	98.3	98.6	ID	99.6	99.1	99.3	99.1	98.1
V7	99.0	99.5	90.7	98.6	98.3	98.6	99.6	ID	99.1	99.3	99.1	98.1
V8	99.5	99.0	89.9	98.1	98.5	98.5	99.1	99.1	ID	99.5	99.6	98.6
V9	99.0	99.1	90.0	98.3	98.6	98.3	99.3	99.3	99.5	ID	99.5	98.5
V10	99.1	99.0	89.9	98.1	98.5	98.5	99.1	99.1	99.6	99.5	ID	99.0
NCBI_ 00044.2	98.1	98.0	89.5	98.5	98.8	97.5	98.1	98.1	98.6	98.5	99.0	ID

The percentage values of sequences showing maximum identity are highlighted in blue.

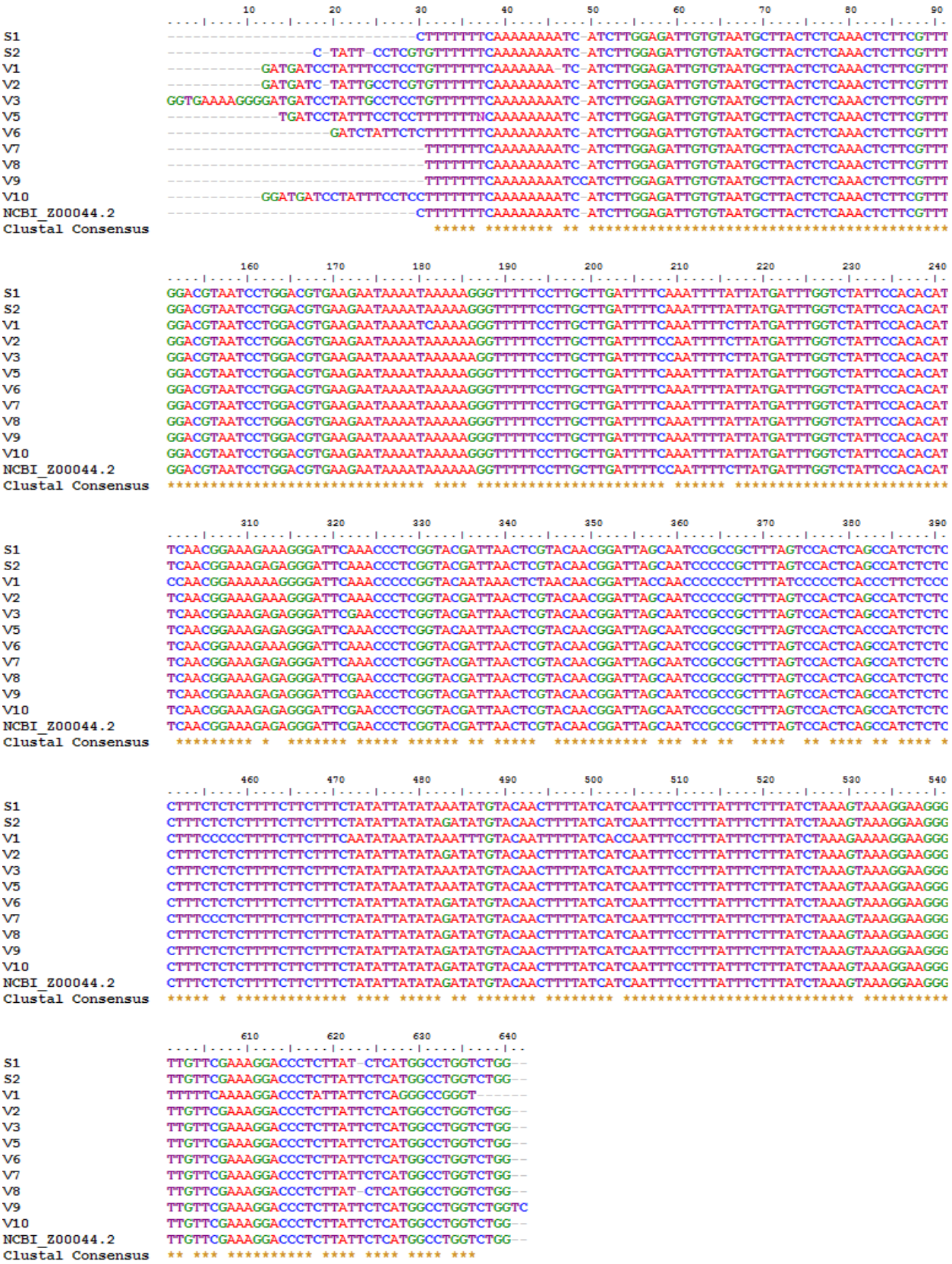

Fig. 2. Multiple alignment of tobacco samples with the reference sequence (NCBI_Z00044.2) using ClustalW program in BioEdit. Different colors represent four different nucleotides. Red is for adenine, blue for cytosine, green for guanine and purple for thymine. The consensus below indicates the conservation of nucleotides.

substitution mutations are distributed along the entire length of the selected region. A similarity matrix could better reveal substitution differences between the cultivars. Table 2 demonstrates that similarity ranged from 90.4 to 99.6 percent. A cut off value was set as indicator of different cultivar. Ninety nine percent identities indicate that these are the same cultivars but values below this index may be taken as different cultivars.

3.2. Chloroplast markers can differentiate tobacco cultivars

Phylogenetic analysis revealed the evolutionary relationship among the sequence samples thus leading to the identification of tobacco cultivars (Fig. 3). Our results indicate that some samples are closely related to each other; e.g., V8, V9, V10 and S1 are forming one clade showing that their origin is same, and hence are more close to each other than the rest of the samples. Similarly V6, S2 and V7 are forming second clade while V2, V3 and reference sequence (NCBI_Z00044.2) constitute the third clade (Fig. 3). The V1 and V5 are the outgroup sequences as they are present at the root of the tree indicating that they are the most divergent of all other samples. Comparison of V1 and V5 reveals that V1 is the most divergent of all the samples. The neighbor-net network analysis also demonstrated the similar results (Fig. 4), thus validating our results.

The above results allow us to infer that there exist at least 5 varieties of tobacco in these samples. The groups of samples in the same variety are listed in the table 3. V1 is the most distant cultivar. V5 is also separated from the rest of cultivars. But V2, V3 and NCBI sequence make further linage and can be considered as separate clade. A bigger cluster including V7, S2 and V6 is differentiated into distinct group. S1, V9, V10 and V8 constitute the 5th cluster that can be considered as a single variety. It can be extrapolated that these actually belong to the same variety as their sequences exhibit an evolutionary link i.e., having the same ancestral species. Hence there is chance of mixing, if these are considered as separate varieties physically. S2 and V7 may be originally treated as different but analysis indicated that this is the same variety but, names were given differently.

4. DISCUSSION

Molecular and bioinformatics data analysis of 11 unknown commercial samples of tobacco enabled us to ascertain that there exist at least 5 commercial varieties of tobacco that were mixed together. In this study it was possible to amplify a fragment of 650 bp through PCR. These sequences features mostly substitution mutations but deletions are also detectable at upstream and downstream terminal regions. Noman et al. [12] used similar CpDNA marker and found mutations in the variable regions of olive chloroplast. Marker sequence analysis of 965 unknown olive plants resulted in identification of 515 plants differentiated into 27 varieties successfully. Interestingly, the similarity matrix ranged from 90.4 to 99.6 percent. This indicates that differences not only account for existence of different varieties but may even different species. These observations were further validated through phylogenetic reconstruction.

Molecular phylogeny can be inferred in Muscari (*Asparagaceae*) species based on CpDNA sequences [13]. Liu et al. [14] successfully detected the intraspecific polymorphism in Phalaenopsis equestris cultivars. Their results are analogous to

Table 2: SNP (single nucleotide polymorphism) analysis of tobacco cultivars

	1	7	19	156	179	186	258	283	292	305	334	352	374	385	390	402	426	447	453	516	519	548	593
V1	G	T	-	G	A	C	A	A	A	A	C	C	A	C	A	A	C	A	A	A	A	A	T
V2	G	T	-	A	C	C	A	A	A	G	C	G	A	C	G	A	T	T	G	A	A	A	T
V3	G	T	-	A	C	C	-	G	G	G	G	G	A	C	G	G	T	T	A	A	A	A	T
V5	T	N	-	G	A	A	A	G	A	A	G	C	G	T	A	A	T	A	A	A	A	A	T
V6	T	T	-	G	A	A	A	A	A	G	G	G	A	T	G	A	T	T	G	A	A	A	T
V7	T	T	-	G	A	A	A	G	A	G	G	G	A	T	G	A	C	T	G	A	A	A	T
V8	T	T	-	G	A	A	A	G	G	G	G	G	G	T	G	G	T	T	G	A	A	A	-
V9	T	T	C	G	A	A	A	G	G	G	G	G	A	T	G	G	T	T	G	A	A	A	T
V10	T	T	-	G	A	A	A	G	G	G	G	G	G	T	G	G	T	T	G	A	G	A	T
S1	T	T	-	G	A	A	A	A	A	G	G	G	G	T	G	G	T	T	A	A	A	A	-
S2	G	T	-	G	A	A	A	G	A	G	C	G	A	T	G	A	T	T	G	A	A	A	T
NCBI	T	T	-	A	C	C	A	G	G	G	G	G	G	C	G	G	T	T	G	G	G	G	T

NCBI is a reference identifier. SNPs are highlighted in cyan.

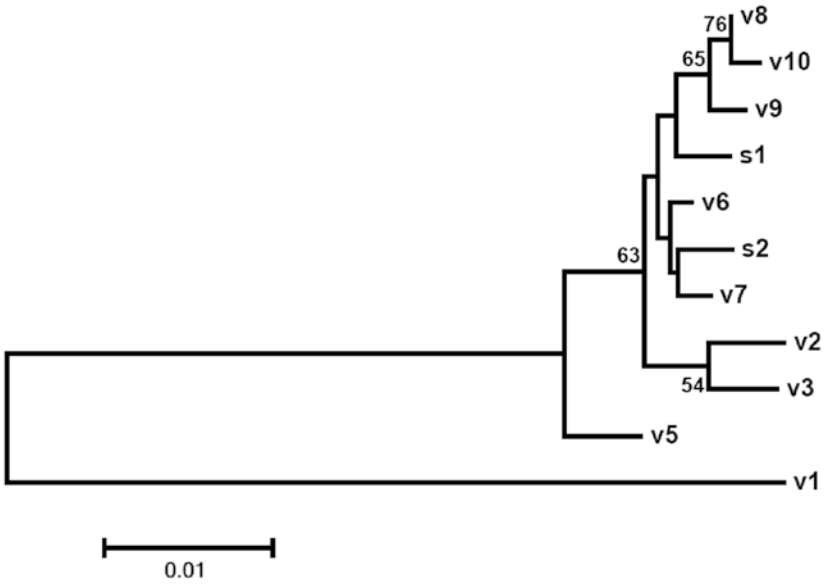

Fig. 3. Phylogenetic tree using Neighbor Joining algorithm (default parameters) to differentiate the samples. The bootstrap value was set to 1000 to check the robustness of tree. Scale bar shows nucleotides substitution per site.

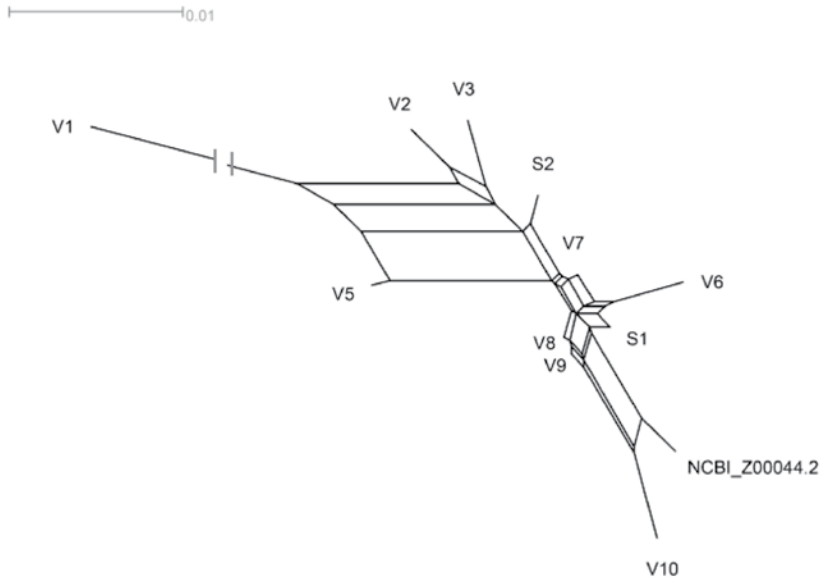

Fig. 4. Neighbor-net network analysis using SplitTree4 package (default parameters). Scale bar shows nucleotide substitution per site. The figure depicts V1 as outgroup OTU (Operational taxonomic unit) as compared to the rest of tobacco samples.

Table 3: Differentiation of tobacco samples

Variety No.	Identifiers included in the same variety
1	V1
2	V5
3	V2, V3
4	V7, S2, V6
5	S1, V9, V10, V8

A total of 5 varieties (cultivars) can be distinguished. V and S denote different samples.

our as they were able to evaluate 11 orchid cultivars that could be separated into six distinct groups. Our data demonstrates that there exist at least 5 varieties of tobacco in these samples. The groups of samples in the same variety are listed in the table 3. V1 is the most distant cultivar. V5 is also separated from the rest of cultivars. But V2, V3 and NCBI sequence make further linage and can be considered as separate clade. A bigger cluster including V7, S2 and V6 is differentiated into distinct group. S1, V9, V10 and V8 constitute the 5th cluster that can be considered as a single variety. It can be extrapolated that these actually belong to the same variety as their sequences exhibit an evolutionary link i.e., having the same ancestral species. Hence there is chance of mixing, if these are considered as separate varieties physically. S2 and V7 may be originally treated as different but analysis indicated that this is the same variety but, names were given differently. For further validation of these results morphological data is inevitable.

As high throughput sequencing is becoming cheaper and cheaper with time it is now possible to do the varietal profiling [15,16]. In this regard the role of SNP variations is inevitable. This is particularly useful for plant breeder's rights and varieties approval programs. Thus, chloroplast genome profiling is much cheaper and rapid strategies for varietal detection than nuclear genome that is still expensive and requires high quality expertise in genome informatics. In line with this assertion Zhang et al. [17] established the differentiation of rubber dandelion species from weedy relatives by using chloroplast high throughput NGS sequencing and molecular markers.

The results of this endeavour have repercussions in on-farm preservation of tobacco germplasm, and in availability of authentic plant material for the propagation of reliable varieties. This approach has implications in cultivar identification of any other species in plant kingdom.

5. CONCLUSIONS

Chloroplast markers used in this study have sufficient resolving power to discriminate 13 commercial samples of tobacco into 5 cultivars at intra-varietal.

6. ACKNOWLEDGEMENT

Mr. Muhammad Riaz at NCB, QAU is acknowledged for editing of this manuscript.

7. REFERENCES

1. Davalieva, K., I. Maleva, K. Filiposki, O. Spiroski & G.D. Efremov. Genetic variability of Macedonian tobacco varieties determined by microsatellite marker analysis. *Diversity* 2: 439-449 (2010).
2. Denduangboripant, J., T. Piteekan & M. Nantharat. Genetic polymorphism between tobacco cultivar-groups revealed by amplified fragment length polymorphism analysis. *Journal of Agricultural Science* 2: 41 (2010).
3. Caguiat, X.G.I., J.C. Yabes & D.A. Tabanao. Phylogenetic Similarity of Popular Rice Varieties from Different Sources. *Journal of Phylogenetics & Evolutionary Biology:* 1-4 (2016).
4. Giannoulia, K., F. Gazis, N. Nikoloudakis, D. Milioni & K. Haralampidis. Breeding, molecular markers and molecular biology of the olive tree. *European Journal of Lipid Science and Technology* 104: 574-586 (2002).
5. Yan, M., Y. Xiong, R. Liu, M. Deng & J. Song. The application and limitation of universal chloroplast markers in discriminating East Asian evergreen oaks. *Frontiers in plant science* 9 (2018).
6. Mariotti, R., N.G. Cultrera, C.M. Díez, L. Baldoni & A. Rubini. Identification of new polymorphic regions and differentiation of cultivated olives (*Olea europaea L.*) through plastome sequence comparison. *BMC Plant Biology* 10: 211 (2010).
7. Doyle, J.J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* 19: 11-15 (1987).
8. Hall, T.A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41: *[London]: Information Retrieval Ltd., c1979-c2000.*: 95-98 (1999).
9. Thompson, J.D., D.G. Higgins & T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22: 4673-4680 (1994).
10. Tamura, K., G. Stecher, D. Peterson, A. Filipski & S. Kumar. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology*

and Evolution 30: 2725-2729 (2013).
11. Huson, D.H. & D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution* 23: 254-267 (2006).
12. Noman, M., W. Ajmal, M.R. Khan, A. Shahzad & G.M. Ali. Exploitation of Concatenated Olive Plastome DNA Markers for Reliable Varietal Identification for On-Farm Genetic Resource Conservation. *American Journal of Plant Sciences* 6: 3045 (2015).
13. Dizkirici, A., O. Yigit, M. Pinar & H. Eroglu. Molecular phylogeny of Muscari (Asparagaceae) inferred from cpDNA sequences. *Biologia*: 1-10 (2018).
14. Liu, Y.-C., B.-Y. Lin, J.-Y. Lin, W.-L. Wu & C.-C. Chang. Evaluation of chloroplast DNA markers for intraspecific identification of Phalaenopsis equestris cultivars. *Scientia Horticulturae* 203: 86-94 (2016).
15. Chen, X., J. Zhou, Y. Cui, Y. Wang, B. Duan & H. Yao. Identification of Ligularia Herbs Using the Complete Chloroplast Genome as a Super-Barcode. *Frontiers in Pharmacology* 9: 695 (2018).
16. Osuna-Mascaró, C., R.R. de Casas & F. Perfectti. Comparative assessment shows the reliability of chloroplast genome assembly using RNA-seq. *Scientific Reports* 8: 17404 (2018).
17. Zhang, Y., B.J. Iaffaldano, X. Zhuang, J. Cardina & K. Cornish. Chloroplast genome resources and molecular markers differentiate rubber dandelion species from weedy relatives. *BMC Plant Biology* 17: 34 (2017).