Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)
ISSN 1673-1581 (Print); ISSN 1862-1783 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Diversity arrays technology (DArT) for studying the genetic polymorphism of flue-cured tobacco (*Nicotiana tabacum*)*#

Xiu-ping LU[1,2], Bing-guang XIAO†‡[2], Yong-ping LI[2], Yi-jie GUI[1], Yu WANG[1], Long-jiang FAN†‡[1]

([1]*Department of Agronomy, Zhejiang University, Hangzhou 310058, China*)

([2]*Yunnan Academy of Tobacco Agricultural Sciences and China Tobacco Breeding Research Center at Yunnan, Yuxi 653100, China*)

†E-mail: xiaobg@263.net; fanlj@zju.edu.cn

Received Aug. 23, 2012; Revision accepted Nov. 15, 2012; Crosschecked June 24, 2013

Abstract: Diversity arrays technology (DArT) is a microarray-based marker system that achieves high throughput by reducing the complexity of the genome. A DArT chip has recently been developed for tobacco. In this study, we genotyped 267 flue-cured cultivars/landraces, including 121 Chinese accessions over five decades from widespread geographic regions in China, 103 from the Americas, and 43 other foreign cultivars, using the newly developed chip. Three hundred and thirty polymorphic DArT makers were selected and used for a phylogenetic analysis, which suggested that the 267 accessions could be classified into two subgroups, which could each be further divided into 2–4 sections. Eight elite cultivars, which account for 83% of the area of Chinese tobacco production, were all found in one subgroup. Two high-quality cultivars, HHDJY and Cuibi1, were grouped together in one section, while six other high-yield cultivars were grouped into another section. The 330 DArT marker clones were sequenced and close to 95% of them are within non-repetitive regions. Finally, the implications of this study for Chinese flue-cured tobacco breeding and production programs were discussed.

Key words: *Nicotiana tabacum*, Diversity arrays technology (DArT), Genetic diversity, Flue-cured tobacco
doi:10.1631/jzus.B1200227 **Document code:** A **CLC number:** S336

1 Introduction

Genetic resources play a key role in crop breeding and a wide genetic variability within the germplasm pools of crops is critical for good breeding practices. To estimate the genetic diversity of a crop and to understand the population structure or genetic background of crops, several investigations have been carried out in many crops at the country level and worldwide. Traditional investigations of genetic diversity have considered phenotypic traits, isoenzymes, etc., which usually present with relatively low levels of polymorphism, and the emergence of DNA markers in the 1980s (Botstein *et al*., 1980) made these investigations more efficient. Early molecular marker-based studies in crops mainly focused on the use of amplified fragment length polymorphism (AFLP) or randomly amplified polymorphic DNA (RAPD). Along with large-scale sequencing efforts in crops, sequence-based markers, such as simple sequence repeat (SSR) and single nucleotide polymorphism (SNP), have been developed and widely used in studies of crop diversity (Agarwal *et al*., 2008).

Tobacco (*Nicotiana tabacum*) is one of the most economically important nonfood crops. Tobacco breeding programs face challenges with regard to quality characteristics, such as low toxin content, disease resistance, and drought tolerance, and thus,

‡ Corresponding authors

* Project supported by the China National Tobacco Corporation (Nos. 110200701023 and 110201101010-JY-04) and the Yunnan Provincial Tobacco Company (Nos. 08A05 and 2010YN02), China

Electronic supplementary materials: The online version of this article (doi:10.1631/jzus.B1200227) contains supplementary materials, which are available to authorized users

there have recently been several molecular marker-based studies on genetic diversity. Initially, as with other crops, AFLP (Ren and Timko, 2001) and then SSR (Yang *et al*., 2007; Moon *et al*., 2009a; 2009b; Fricano *et al*., 2012) and even methylation marker (Zhao *et al*., 2011) were used. Molecular diversity in a worldwide collection of tobacco germplasm revealed the structured populations, basically by different tobacco types (Fricano *et al*., 2012). Flue-cured tobacco accounts for approximately 75% of world tobacco production (Moon *et al*., 2009b) and over 90% of that in China (Liu, 2011). However, flue-cured tobacco contains only a small portion of *N. tabacum* germplasm (Moon *et al*., 2009b). However, its genetic diversity has been steadily decreasing over the past few decades due to strict breeding selection. Moon *et al*. (2009a) investigated a population of 117 American flue-cured cultivars with 71 microsatellite primer pairs and demonstrated that over 50% of the allelic diversity was lost from varieties released between the 1930s and 2005. Although there have been several investigations based on limited SSR or AFLP markers and cultivars (Zhang *et al*., 2006; Yang *et al*., 2007), little is currently known about the genetic variation within Chinese flue-cured tobacco pools or the impact of breeding practices on genetic diversity within Chinese elite cultivars.

Diversity arrays technology (DArT) is a microarray-based marker system that achieves high throughput by reducing the complexity of a DNA sample to obtain a 'representation' of that sample (Jaccoud *et al*., 2001). It has been developed and used in many plants, including wheat and barley, which have complex genomes, while their genome sequences are not yet available for genetic mapping or studies of genetic diversity, etc. (Wenzl *et al*., 2004; Akbari *et al*., 2006; Tinker *et al*., 2009; Alsop *et al*., 2011). Very recently, we developed a tobacco DArT chip that includes 7680 representative sequence tags and which has been successfully used to construct tobacco genetic maps (Lu *et al*., 2013).

In this study, we profiled a germplasm set of 267 flue-cured tobacco cultivars including 121 elite materials over five decades of cultivar development and landraces or farmer varieties, from widespread geographic regions in China and 103 typical flue-cured tobacco accessions from the Americas using this newly developed tobacco DArT chip. We sought to elucidate (1) the genetic variation of flue-cured tobacco germplasms, particular in the Chinese pool, and (2) the genetic background or origin of the main elite cultivars in current tobacco production.

2 Materials and methods

2.1 Plant materials and DNA extraction

A total of 121 Chinese flue-cured tobacco accessions, including elite materials over five decades of cultivar development, and landraces or farmer varieties from widespread geographic regions were collected and used in this study (Table 1; Table S1). To elucidate the evolutionary origin of the Chinese accessions, 103 typical flue-cured tobacco accessions from the Americas and 43 from other countries were also included for the phylogenetic analysis. DNA was extracted from fresh leaf tissue (~200 mg) of these 267 tobacco accessions using a cetyltrimethylammonium bromide (CTAB) protocol (Sambrook and Russell, 2001).

Table 1 Geographic regions of 267 flue-cured tobacco cultivars/landraces used in this study

Geographic region	Country	*n*
Asia	China	121
	Japan, Thailand, Vietnam, DPRK, ROK, and the Philippines	18
North America	USA	82
	Canada	8
South America	Brazil, Cuba, and Argentina	13
Africa	Zimbabwe, Somalia, Tanzania, and Zambia	18
Europe	Poland, Yugoslavia, and England	4
Oceania	Australia	3
Total		267

n: number of flue-cured tobacco cultivars/landraces. ROK: Republic of Korea; DPRK: Democratic People's Republic of Korea

2.2 DArT marker detection

A DArT marker chip for tobacco (Lu *et al*., 2013) was used in this study. Briefly, a genome representation of a mixture of 5 cultivars (HHDJY, Hicks Broad Leaf, Florida301, Burley21, and Turkey Basma) was produced after *Pst*I/*Taq*I digestion and spotted on microarray slides, and the individual genotypes were screened for polymorphism based on fluorescence signals. The 267 tobacco accessions were screened for fingerprinting.

The *Escherichia coli* clones containing the polymorphic DArT markers identified using the discovery arrays were re-arrayed into 384-deep-well microliter plates and grown at 37 °C for 20 h. Plasmid DNA, isolated using the Eppendorf Perfectprep Plasmid 384 procedure, was sequenced in both directions using the M13R (5′-GGAAACAGCTAT GACCATG-3′) and T7-ZL (5′-TAATACGACTCA CTATAGGG-3′) primers. Following an ethanol precipitation cleanup step, the reactions were run on an ABi 3730xl capillary electrophoresis instrument. All sequence reads were assembled and merged to provide one high-quality read per clone where possible. Vector sequences and *Pst*I sites were trimmed so as not to introduce biased similarity among DArT clones in current or future analyses (Tinker *et al*., 2009).

2.3 Data analysis

The genetic relationship of the 267 tobacco varieties was investigated using the program NTSYSpc 2.11e (http://www.exetersoftware.com) with the following settings: qualitative data, dice coefficient, sequential agglomerative hierarchical and nested clustering method (SAHN), and unweighted pair-group method with arithmetic average (UPGMA).

The population structure among all 267 tobacco accessions was evaluated with STRUCTURE 2.3.3 (Hubisz *et al*., 2009) using an admixture model with no linkage. Three hundred and thirty DArT markers were used and considered to be haploid for this analysis. All analyses had a burn-in length of 50000 iterations and a run length of 100000 iterations. Three replicates were carried out at each value of K (population number, from 2 to 10). Simulations were run with uncorrelated allele frequencies.

DArT clones were collected and sequenced by the Sanger method in one direction. Repeat sequences were identified and masked by searching against Repbase 16.06 (http://www.girinst.org) with RepeatMasker (http://www.repeatmasker.org/). The non-repetitive DArT sequences were further compared with the tobacco unigene set TobEA (Edwards *et al*., 2010) and GenBank nr database by BLASTX, and the sequences with significant hits (e-value<10^{-7}) were as genic regions. Gene ontology (GO) classification was carried out for those hits (e-value<10^{-7}) using Blast2GO (Conesa *et al*., 2005) and the GO histogram was generated by WEGO (Ye *et al*., 2006).

Two summary statistics parameters were used for diversity measure with a Perl script. The two parameters estimate the population mutation rate per locus based on the number of segregating sites (Watterson, 1975) and the mean value of pairwise divergence per locus (π) (Tajima, 1983), respectively.

3 Results

3.1 Characteristics of DArT markers

A total of 330 high-quality DArT markers were identified from discovery arrays and used in the following diversity analysis. The polymorphism information content (PIC) values of the 330 identified polymorphic markers ranged from 0.07 to 0.50, with an average value of 0.34 (78.8% >0.2) (Table 2).

Table 2 Polymorphism information content (PIC) values for 330 DArT markers

PIC value	DArT	
	Number	Percentage (%)
0.5–0.4	182	55.2
0.4–0.3	48	14.5
0.3–0.2	30	9.1
0.2–0.1	27	8.2
<0.1	43	13.0

To further characterize the 330 DArT markers, their clones were collected and sequenced by the Sanger method. Of the 330 clones, 259 were sequenced successfully and annotated. Bioinformatic analysis indicated that only 4.4% of the marker sequences were masked as repetitive regions by RepeatMasker, and particularly 93 (over 35%) marker sequences, which had significant similarity with the annotated protein-coding genes in the public databases (BLASTX, e-value<10^{-7}), were potentially located in genic regions. The best hits/genes of the 93 markers represent diverse functions according to the GO classification (Fig. 1), and dominated in such as metabolic process and catalytic function. It has been suggested that many DArT markers are within genes or gene-rich regions, according to 1774 oat DArT clone sequences (Tinker *et al*., 2009). We surmise that many other markers should come from functional or nearby regions of genes. The results suggested that the tobacco DArT marker chip developed by us shows a trend similar to those of chips in other plants.

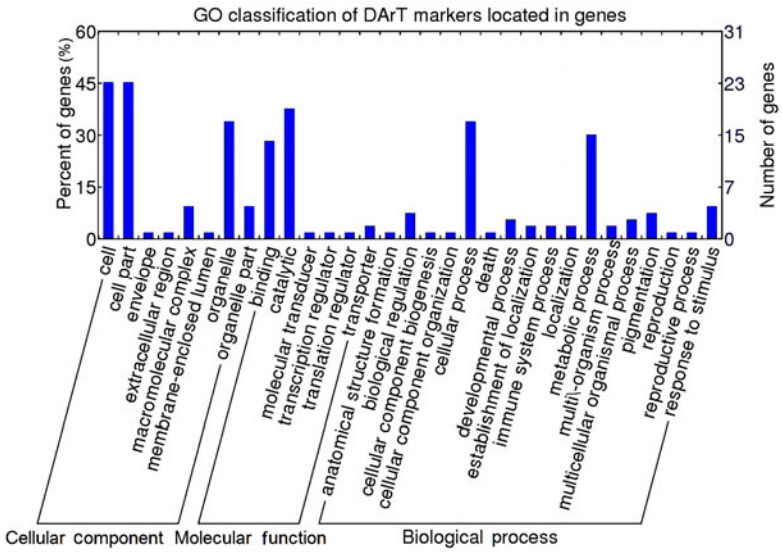

Fig. 1 Functional classification of genes that are located by 93 DArT markers

3.2 Cluster analysis of 267 flue-cured tobacco accessions

Based on the genotyping results of 330 DArT markers, the phylogenetic tree and population structure of 267 flue-cured tobacco accessions from China and other countries (Table 1) were investigated. In general, both clustering results (Figs. 2 and 3; detailed information was provided in Table S1) fit the pedigree of the cultivars used in this study. For example, 581 were bred by selection from Dajingyuan and they initially clustered together. Yunyan85 and Yunyan87 have the same parental line (K326) and initially they clustered together. The 267 accessions appeared to be classified into two subgroups according to the phylogenetic analysis (Fig. 2). Each subgroup can be further classified into 2–4 sections according to a phylogenetic tree (Table 3; Fig. 2) or population structure analysis (Fig. 3). In Subgroup 1, two sections (Sections A and B or #1 and #2) could be observed by both analyses. In contrast, Subgroup 2 is relatively complex, and at least 2–4 sections can be ascertained (Table S1). There are two main sections (C and D) in the phylogenetic tree or three sections (#3–#5) by population structure analysis in this subgroup and Section C includes particularly diverse accessions (144 of 267 accessions). In general, Sections A, B, and D by phylogenetic analysis included most accessions of Sections #2, #1, and #4 of population structure analysis respectively, while Section C covered #3 and #5 (detailed corresponding section pairs are provided in Table S1).

Interestingly, the main cultivars in China are all contained in Subgroup 1. For example, eight cultivars (K326, Yunyan85, Yunyan87, Yunyan97, Zhongyan100, Longjiang911, HHDJY, and Cuibi1) from this subgroup accounted for 83% of the total area of tobacco production in 2010 (Liu, 2011). Furthermore, among the eight elite cultivars, two with high quality (HHDJY and Cuibi1) were grouped into a section (Section A by phylogenetics) that included the American cultivars K346, NC89, etc., while the other six with high yield were grouped into another independent section (Section B by phylogenetics) that included the American cultivars K326, NC71, etc. (Table 3). To compare genetic diversity with the Chinese tobacco cultivars (released from 1950 to now) and landraces, the eight cultivar set has the smallest values of the statistic parameter (π) and number of segregating sites in the 330 DArT marker sites (Table 4; Table S2), suggesting that they only included part of the genetic diversity of Chinese tobacco accessions. These results indicate that current Chinese tobacco breeding programs and tobacco production reflect a narrow genetic diversity or background.

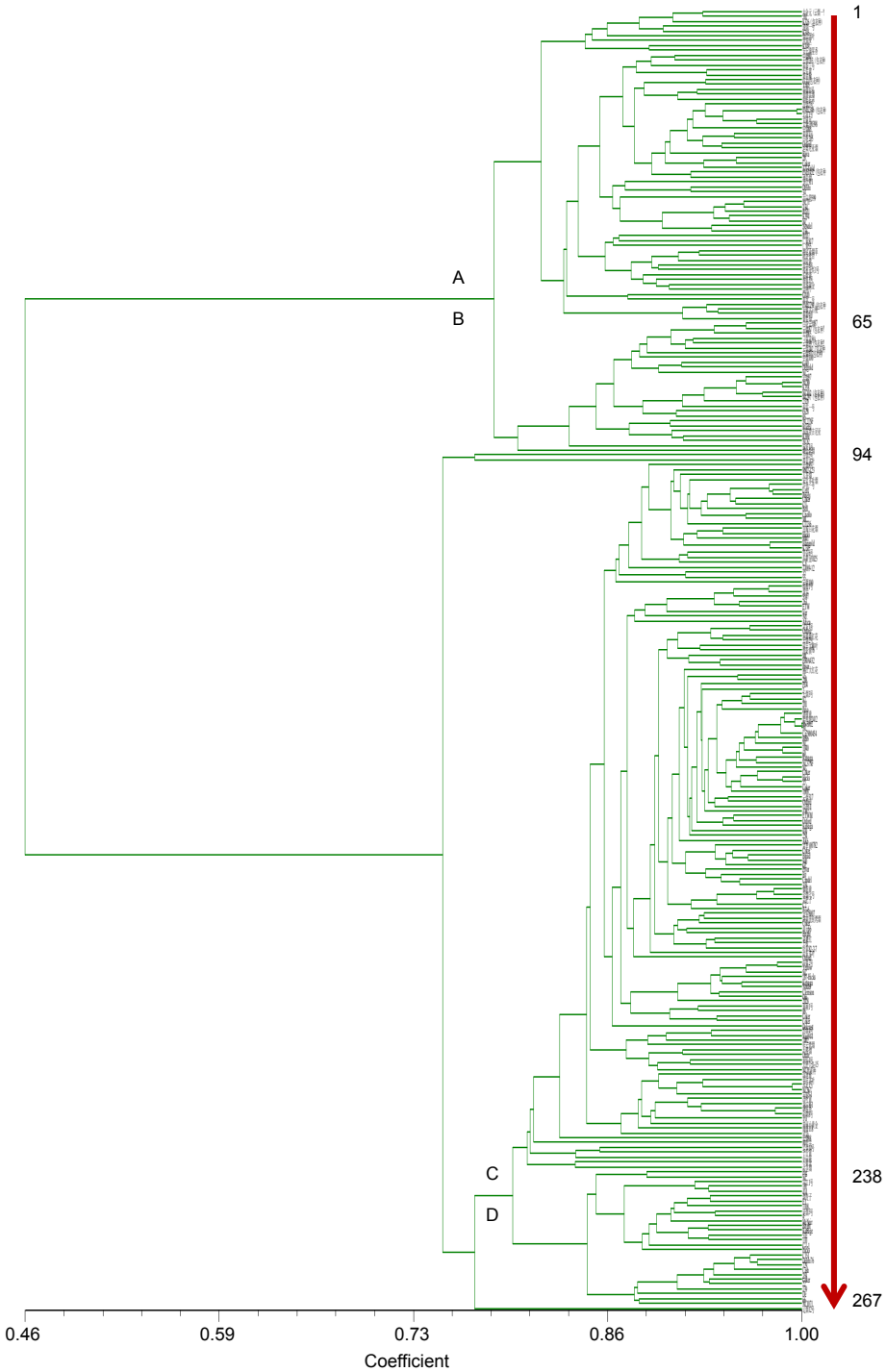

Fig. 2 DArT-based UPGMA dendrogram of 267 flue-cured tobacco cultivars
Four sections and phylogenetic codes of 267 accessions are shown. Detailed information of the four sections and their accessions was provided in Table 3 and Table S1

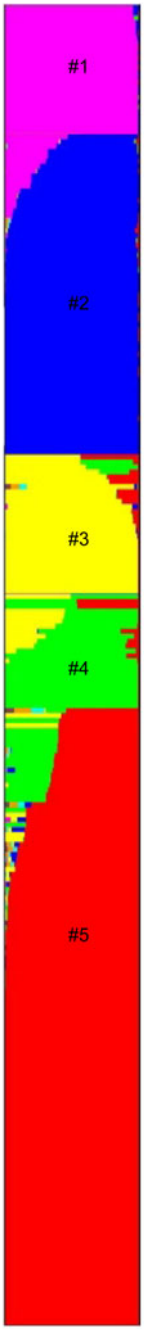

Fig. 3 Population structure of 267 flue-cured tobacco accessions by STRUCTURE program
Each horizontal bar in the graph represents a single accession and its inferred proportion of genetic admixture. The parts with different colors represent five different subpopulations. Detailed information of five groups and their accessions were provided in Table S1 (Note: for interpretation of the references to color in this figure legend, the reader is referred to the web version of this article)

Table 3 Phylogenetic clusters of Chinese and American flue-cured tobacco and representative cultivars

Subgroup	Section	Number	Representative cultivar	
			China	USA
1	A	64	HHDJY, Cuibi1	K346, NC89
	B	27	Yunyan85, Yunyan87, Hongyan100	K326, NC71
2	C	144	Yunyan2, Yunyan3	Hicks Broad Leaf, Coker176
	D	29	Yunyan98, Yunyan100	Coker139, Oxford2

Detailed information of accessions was provided in Table S1

Table 4 Genetic diversity of Chinese flue-cured tobacco accessions in various breeding periods

Group	Number of accessions	Number of segregating sites	π^*
Cultivars (1950–now)	72	302	0.412
Eight-elite set[#] (now)	8	140	0.233
Landraces	49	315	0.385
1950–1970s	36	294	0.395
1980–1990s	24	282	0.412
2000s	10	237	0.325

[#] Eight elites include K326, Yunyan85, Yunyan87, Yunyan97, Zhongyan100, Longjiang911, HHDJY, and Cuibi1; [*] π: average pairwise nucleotide diversity (Tajima, 1983)

Our cluster analysis appears to show the following: (1) The 267 cultivars were not first clustered by their geographical origins, i.e., cultivars from the same region or country were not usually grouped together. The results suggested that germplasm introduction and cross-breeding programs have given flue-cured tobacco cultivars a mixed genetic background in different countries. On the other hand, the elite cultivars that are bred and used in different countries show a similar genetic background. (2) Farmer varieties show a wide genetic diversity (e.g., 315 segregating sites; Table 4) and cover each section, including even nodes that are further downstream in the phylogenetic tree. These varieties, together with foreign cultivars/germplasm (detailed information for diversity measures was provided in Table S2), provide a valuable genetic pool to meet the challenge that faces Chinese flue-cured tobacco breeding programs. (3) Elite cultivars that were widely planted at different time in the US and China were usually grouped together in the same section (Table 3). In general, the trend of reduction in the number of segregating sites and π value over time was observed in the Chinese tobacco breeding program (Table 4). The results suggest that over the past five decades, the Chinese tobacco breeding program has followed that of the US (Moon *et al*., 2009a). However, two new cultivars that were just released by YATAS (Yunyan98 and Yunyan100) are located in an independent section (Section D in the phylogenetic tree) of Subgroup 2 and present a unique genetic background to other elites.

4 Discussion

In this study, we investigated the genetic diversity of 267 flue-cured accessions, including 121 Chinese cultivars/landraces, 103 cultivars from the Americas and 43 other foreign cultivars as controls, by using 330 DArT markers. To our knowledge, this is the first attempt to investigate the genetic diversity of Chinese tobacco with the largest cultivar sample size and number of molecular markers. Our results are generally consistent with those of other studies of diversity based on SSR markers in the American and Chinese accessions. Moon *et al*. (2009a; 2009b) reported similar cluster results and limited genetic diversity for the American flue-cured tobacco cultivar used in this study. Their investigation (Moon *et al*., 2009a) suggested that the allelic diversity in American flue-cured tobacco cultivars has been declining since the 1930s. A narrow genetic diversity has also been reported in Chinese flue-cured tobacco based on AFLP and inter-simple sequence repeat (ISSR) (Zhang *et al*., 2006; Yang *et al*., 2007).

As mentioned above, a narrow genetic diversity, which usually makes the development of molecular markers more difficult, was observed in tobacco. In addition to AFLP, RAPD, and SSR markers, DArT provides an alternative for tobacco, as shown here. The DArT technology combines a reduction of genome complexity with high throughput and

cost-effective hybridization-based polymorphism detection. DArT is not sequence-based, and is thus particularly helpful for species such as tobacco for which the genome sequence is not available or sequence data are limited. Many DArT markers are within genes or gene-rich regions, as shown in this study and others (Tinker *et al*., 2009), which is also particularly important for quantitative trait locus (QTL) mapping and molecular-assisted breeding of species with complex genomes (big, repetitive, and polyploid), such as wheat, barley, oat, and tobacco.

Our results offer several implications for Chinese flue-cured tobacco breeding programs and production. First, in Chinese tobacco production, cultivars with different genetic backgrounds could be planted in reasonable proportions or distributions. For example, an increase in the distribution of two new cultivars (Yunyan98 and Yunyan100) should improve the current situation in which only a limited number of elite cultivars with an extremely narrow genetic background are being planted. As mentioned above, the two cultivars are located in an independent section of Subgroup 2 and present a unique genetic background to other elites. If they eventually become elite cultivars like K326 and HHDJY in China, they will be the first cultivars with a novel genetic background or Chinese genetic characteristics. Meanwhile, even for cultivars from Subgroup 1, an increase in the distribution of cultivars with high quality (HHDJY and Cuibi1) should also be helpful for easing the current situation in China. Second, farmer varieties should be given more attention in Chinese tobacco breeding programs. They provide the same wide genetic novelty as foreign germplasm. In Subgroup 2, several landraces are particularly diverse with respect to others (Fig. 2; Table S1), and therefore provide tremendous promise for genetic improvement in current tobacco breeding programs in China.

Compliance with ethics guidelines

Xiu-ping LU, Bing-guang XIAO, Yong-ping LI, Yi-jie GUI, Yu WANG, and Long-jiang FAN declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Agarwal, M., Shrivastava, N., Padh, H., 2008. Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep.*, **27**(4):617-631. [doi:10.1007/s00299-008-0507-z]

Akbari, M., Wenzl, P., Caig, V., Carling, J., Xia, L., Yang, S., Uszynski, G., Mohler, V., Lehmensiek, A., Kuchel, H., *et al*., 2006. Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theor. Appl. Genet.*, **113**(8):1409-1420. [doi:10.1007/s00122-006-0365-4]

Alsop, B.P., Farre, A., Wenzl, P., Wang, J., Zhou, M., Romagosa, I., Kilian, A., Steffenson, B.J., 2011. Development of wild barley-derived DArT markers and their integration into a barley consensus map. *Mol. Breed.*, **27**(1):77-92. [doi:10.1007/s11032-010-9415-3]

Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.*, **32**(3):314-331.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**(18):3674-3676. [doi:10.1093/bioinformatics/bti610]

Edwards, K.D., Bombarely, A., Story, G.W., Allen, F., Mueller, L.A., Coates, S.A., Jones, L., 2010. TobEA: an atlas of tobacco gene expression from seed to senescence. *BMC Genomics*, **11**(1):142. [doi:10.1186/1471-2164-11-142]

Fricano, A., Bakaher, N., del Corvo, M., Piffanelli, P., Donini, P., Stella, A., Ivanov, N.V., Pozzi, C., 2012. Molecular diversity, population structure, and linkage disequilibrium in a worldwide collection of tobacco (*Nicotiana tabacum* L.) germplasm. *BMC Genet.*, **13**(1):18. [doi:10.1186/1471-2156-13-18]

Hubisz, M.J., Falush, D., Stephens, M., Pritchard, J.K., 2009. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.*, **9**(5):1322-1332. [doi:10.1111/j.1755-0998.2009.02591.x]

Jaccoud, D., Peng, K., Feinstein, D., Kilian, A., 2001. Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res.*, **29**(4):E25. [doi:10.1093/nar/29.4.e25]

Liu, G., 2011. Practical Guide to Leaf Tobacco Production Techniques in China. China National Leaf Tobacco Corporation, Beijing (in Chinese).

Lu, X.P., Gui, Y.J., Xiao, B.G., Li, Y., Tong, Z.J., Liu, Y., Bai, X.F., Wu, W.R., Xia, L., Huttner, E., *et al.*, 2013. Development of DArT markers for linkage map of flue-cured tobacco. *Chin. Sci. Bull.*, **58**(6):641-648.

Moon, H.S., Nicholson, J.S., Heineman, A., Lion, K., Hoeven, R., Hayes, A.J., Lewis, R.S., 2009a. Changes in genetic diversity of US flue-cured tobacco germplasm over seven decades of cultivar development. *Crop Sci.*, **49**(2):498-508. [doi:10.2135/cropsci2008.05.0253]

Moon, H.S., Nifong, J.M., Nicholson, J.S., Heineman, A., Lion, K., Hoeven, R., Hayes, A.J., Lewis, R.S., 2009b.

Microsatellite-based analysis of tobacco (*Nicotiana tabacum* L.) genetic resources. *Crop Sci.*, **49**(6):2149-2159. [doi:10.2135/cropsci2009.01.0024]

Ren, N., Timko, M.P., 2001. AFLP analysis of genetic polymorphism and evolutionary relationships among cultivated and wild *Nicotiana* species. *Genome*, **44**(4):559-571. [doi:10.1139/g01-060]

Sambrook, J., Russell, D.W., 2001. Molecular Cloning: A Laboratory Manual, 3rd Ed. Huang, P.T. *et al.*, translator, 2002. Science Press, Beijing, p.547-610 (in Chinese).

Tajima, F., 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, **105**(2):437-460.

Tinker, N.A., Kilian, A., Wight, C.P., Heller-Uszynska, K., Wenzl, P., Rines, H.W., Bjornstad, A., Howarth, C.J., Jannink, J.L., Anderson, J.M., *et al*., 2009. New DArT markers for oat provide enhanced map coverage and global germplasm characterization. *BMC Genomics*, **10**(1):39. [doi:10.1186/1471-2164-10-39]

Watterson, G.A., 1975. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, **7**(2):256-276. [doi:10.1016/0040-5809(75)90020-9]

Wenzl, P., Carling, J., Kudrna, D., Jaccoud, D., Huttner, E., Kleinhofs, A., Kilian, A., 2004. Diversity arrays technology (DArT) for whole-genome profiling of barley. *PNAS*, **101**(26):9915-9920. [doi:10.1073/pnas.0401076101]

Yang, B.C., Xiao, B.G., Chen, X.J., Shi, C.H., 2007. Assessing the genetic diversity of tobacco germplasm using inter-simple sequence repeat and inter-retrotransposon amplification polymorphism markers. *Ann. Appl. Biol.*, **150**(3): 393-401. [doi:10.1111/j.1744-7348.2007.00139.x]

Ye, J., Fang, L., Zheng, H., Zhang, Y., Chen, J., Zhang, Z., Wang, J., Li, S., Li, R., Bolund, L., 2006. WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.*, 34(Web Sever):W293-W297. [doi:10.1093/nar/gkl031]

Zhang, H.Y., Liu, X.Z., He, C.S., Yang, Y.M., 2006. Genetic diversity among flue-cured tobacco (*Nicotiana tabacum* L.) revealed by amplified fragment length polymorphism. *Bot. Stud.*, **47**(3):223-229.

Zhao, J.H., Zhang, J.S., Wang, Y., Wang, R.G., Wu, C., Fan, L.J., Ren, X.L., 2011. DNA methylation polymorphism in flue-cured tobacco and candidate markers for tobacco mosaic virus resistance. *J. Zhejiang Univ.-Sci. B (Biomed. & Biotechnol.)*, **12**(11):935-942. [doi:10.1631/jzus.B1000417]

List of electronic supplementary materials

Table S1 Information of tobacco accessions used in this study

Table S2 Genetic diversities of accession groups divided by different methods