GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Large-scale detection and application of expressed sequence tag single nucleotide polymorphisms in *Nicotiana*

Y. Wang[1]*, D. Zhou[2]*, S. Wang[2] and L. Yang[3]

[1]College of Agriculture, Shandong Agricultural University, Taian, China
[2]College of Plant Protection, Shandong Agricultural University, Taian, China
[3]Agricultural Big-Data Research Center and College of Plant Protection, Shandong Agricultural University, Taian, China

*These authors contributed equally to this study.
Corresponding author: L. Yang
E-mail: lyang@sdau.edu.cn

Genet. Mol. Res. 14 (3): 7793-7800 (2015)
Received October 31, 2014
Accepted April 28, 2015
Published July 14, 2015
DOI http://dx.doi.org/10.4238/2015.July.14.5

ABSTRACT. Single nucleotide polymorphisms (SNPs) are widespread in the *Nicotiana* genome. Using an alignment and variation detection method, we developed 20,607,973 SNPs, based on the expressed sequence tag sequences of 10 *Nicotiana* species. The replacement rate was much higher than the transversion rate in the SNPs, and SNPs widely exist in the *Nicotiana. In vitro* verification indicated that all of the SNPs were high quality and accurate. Evolutionary relationships between 15 varieties were investigated by polymerase chain reaction with a special primer; the specific 302 locus of these sequence results clearly indicated the origin of Zhongyan 100. A database of *Nicotiana* SNPs (NSNP) was developed to store and search for SNPs in *Nicotiana*. NSNP is a tool for researchers to develop SNP markers of sequence data.

Key words: *Nicotiana*; Single nucleotide polymorphism; EST; Database; Evolution; Primer

INTRODUCTION

Nicotiana tabacum L. is a commercial crop in many countries (Cao et al., 2013), and is widely used as a model in molecular plant-microbe interactions and plant biology (Bombarely et al., 2012; Sierro et al., 2013). There are two main classification methods for *Nicotiana*: one states that *Nicotiana* contains three subgenera, 14 sections, and 66 species (Goodspeed, 1945; Goodspeed and Thompson, 1959); and the other classifies *Nicotiana* into 13 sections and 76 naturally occurring species (Knapp et al., 2004; Lewis, 2011). Therefore, there is a large number of species in the *Nicotiana*, and the relationships between them are unclear.

Compared with other plant models, studies on *Nicotiana* breeding, gene function, and evolution are few because of a lack of effective tools, although the draft genomes of *Nicotiana benthamiana* (Bombarely et al., 2012), *Nicotiana sylvestris*, and *Nicotiana tomentosiformis* (Sierro et al., 2013) have been sequenced. The main molecular markers that have been used were simple sequence repeats (SSRs) (Bindler et al., 2011; Tong et al., 2012), or SSRs combined with intron length polymorphisms (Cao et al., 2013). Therefore, the development of effective single nucleotide polymorphisms (SNPs) is required for *Nicotiana.*

SNPs have been widely used in studies of gene function, phenotypes, evolution, and species diversity, because of their high density and convenience (Li et al, 2009; Hirakawa et al., 2013; Qi et al., 2013). An increasing number of high-density genetic linkage maps have been constructed in the peach (Martínez-García et al., 2013), oilseed rape (Delourme et al., 2013), and the cucumber (Qi et al., 2013), based on SNP mining from sequences or arrays. Furthermore, SNPs have been used to distinguish between similar taxonomic groups and to investigate evolutionary relationships between plants (Lu et al., 2013; Wang et al., 2014). Because of their wide distribution in genomes, SNPs can be used to infer gene function (Hirakawa et al., 2013; Guimaraes et al., 2014).

However, there are not enough arrays or resequence data to detect SNPs in *Nicotiana*. Fortunately, a large number of assembled expressed sequence tags (ESTs) have been reported (Dong et al., 2004), most of which came from well-known genes, so there are probably many similar ESTs among *Nicotiana* species. In the present study, we developed a series of procedures to mine SNPs in *Nicotiana*, based on ESTs. A database of SNPs in *Nicotiana* (NSNP) that we developed will provide a detailed annotation of *Nicotiana* SNPs, and advance the study of *Nicotiana* breeding and evolution.

MATERIAL AND METHODS

Source of sequences

The assembled ESTs of 10 *Nicotiana* species were downloaded from PlantGDB (http://www.plantgdb.org/), and their sequences can be downloaded at http://biodb.sdau.edu.cn/nsnp/data.html.

Plant materials

Fifteen species of *Nicotiana* were obtained from the tobacco lab Shandong Agricultural University (Table 1). Eight of them were wild species, and belonged to three subgenera

and six sections. Five cultivated species were divided into two groups, based on their origin: four were Chinese varieties and NC 95 was an American variety.

Table 1. Characteristics of the 15 species of *Nicotiana* investigated in this study.

Species	Subgenus	Section	Wild/cultivated
N. glauca	Rustica	Paniculata	Wild
N. paniculata	Rustica	Paniculata	Wild
N. otophora	Tabacum	Tomentosae	Wild
N. undulata	Petunioides	Undulatae	Wild
N. sylvestris	Petunioides	Alatae	Wild
N. acuminata	Petunioides	Acuninatae	Wild
N. benthamiana	Petunioides	Suaveolentes	Wild
N. goodspeedii	Petunioides	Suaveolentes	Wild
N. pani	Rustica	Paniculata	Wild
N. benthamiana	Petunioides	Suaveolentes	Wild *double
N. tabacum cv. Guiyan 11	Tabacum	Genuinae	Cultivated
N. tabacum cv. NC 95	Tabacum	Genuinae	Cultivated
N. tabacum cv. Cuibiyihao	Tabacum	Genuinae	Cultivated
N. tabacum cv. Yunyan 85	Tabacum	Genuinae	Cultivated
N. tabacum cv. Zhongyan 100	Tabacum	Genuinae	Cultivated

SNP detection

Sequence alignment was conducted in a pairwise manner in order to find similar sequences that contained SNPs in BLAT (Kent, 2002), based on the ESTs of the 10 *Nicotiana* species. A series of Perl scripts were then used to detect SNPs in the aligned BLAT results, and a detailed annotation of SNPs, included replacement and transversion rates and the detection frequency of each SNP locus, was conducted.

Primer design

In order to verify the accuracy of the SNPs, sequences of SNP-rich regions were obtained using BioPerl scripts (http://www.bioperl.org/wiki/Main_Page). Primers were designed based on the ESTs of *N. benthamiana* by Eprimer3 (Untergasser et al., 2012), and flanked SNP-rich sequences between *N. benthamiana* and *N. tabacum*.

Polymerase chain reaction (PCR) amplification

PCRs were performed in 45-µL volumes that contained 60-75 ng DNA, 4.5 µL 10X PCR buffer, 7.5 mM $MgCl_2$, 0.75 mM dNTPs, 1.08 µM forward primers, 1.08 µM reverse primers, and 3 U *Taq* DNA polymerase. The thermocycling conditions were as follows: an extension for 5 min initial denaturation at 94°C, followed by 35 cycles of 1 min at 94°C, 45 s at 55°C, 1 min at 72°C, and a final extension at 72°C for 10 min. A 1% agarose configuration (220 V, 30 min) was used to separate the PCR products.

Sequence and phylogenetic analysis

After the PCR products were verified by agarose film, they were sequenced by Sangon Biotech (Shanghai, China). Alignments were conducted in order to ascertain the SNP posi-

tions, and sequence quality was checked by ClustalX (Larkin et al., 2007). A maximum-likelihood phylogenetic tree was constructed using MEGA (Tamura et al., 2013), with a 1000-replication bootstrap procedure.

Database architecture and implementation

The NSNP database is a SNP information database in MySQL (www.mysql.com), the data in which are handled by Perl Script. The web interface is implemented in HTML and runs on an Apache web server. NSNP allows internet access with a web client, and all of the data can be searched and used non-commercially.

RESULTS

SNP analysis

In total, 20,607,973 SNPs were predicted (Table 2). The abundance ranged between 67 SNPs/kb (*N. tabacum*) to 256 SNPs/kb (*Nicotiana suaveolens* x *tabacum*), with an average abundance of 87.57 SNPs/kb.Twelve type mutations were present among the SNPs. Four types of mutation (A to G, T to C, G to A, and C to T) accounted for more than 10% of the SNPs, and they were all replacements (Table 3). Four types of replacement accounted for 59.3% of the SNPs; the remaining eight types were transversions, and accounted for only 40.7%.

Table 2. Single-nucleotide polymorphism (SNP) abundance in *Nicotiana*.

Species	Size (kb)	SNPs	Abundance (/kb)
N. attenuata	109.975	14,490	131.76
N. benthamiana	32,530.44	5,615,017	172.61
N. glauca	2.666	505	189.42
N. glauca x *langsdorffii*	2.285	572	250.33
N. langsdorffii	5.741	1,070	186.38
N. langsdorffii x *sanderae*	6125.83	1,456,187	237.71
N. megalosiphon	93.44	9,043	96.78
N. suaveolens x *tabacum*	70.16	17,984	256.32
N. sylvestris	3274.202	390,207	119.18
N. tabacum	193,129.97	13,102,898	67.84

Table 3. Characteristics of single-nucleotide polymorphisms in *Nicotiana.*

Mutation type	Number	Percentage (%)
a to g	108,452,644	13.14
t to a	57,215,598	6.93
a to c	44,937,884	5.44
g to c	32,120,822	3.89
t to c	136,538,823	16.54
c to a	43,356,269	5.25
a to t	57,779,519	7.00
g to a	106,013,995	12.84
c to g	30,293,712	3.67
c to t	138,623,596	16.79
t to g	34,360,071	4.16
g to t	35,940,032	4.35

Verification *in vitro*

After designing the primers, one special primer (CCAAAGCCATCTCCTCACTC, ACTCCATTGCATTCCCAAAC) was selected to process the verification.

Reference sequence and quality analysis

The two *N. benthamiana* were the same as the reference sequence (Figure 1), indicating that the sequences were of high quality and accurate. On the final agarose film, all of the bands were distinct and approximately 400 bp long, indicating that there were no significant differences between the species.

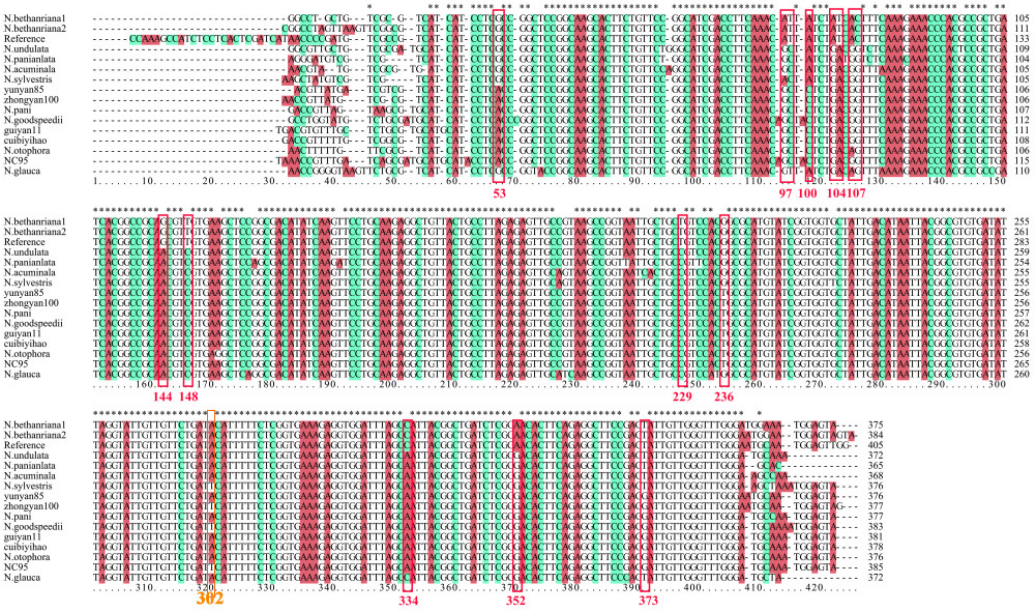

Figure 1. Alignment analysis of 15 *Nicotiana* species. The reference sequence belongs to *Nicotiana benthamiana*, and single nucleotide polymorphisms were obtained by comparison with this sequence; the marked positions are also according to the reference sequence.

SNP positions

The results of the alignment revealed that the sequencing results were in accordance with the predictions (Figure 1). Based on these sequences, the 15 species could be distinguished, including closely related species. There were five varieties of *N. tabacum* L., and three of them (Cuibiyihao, Yunyan 85, and Guiyan 11) matched the predictions. NC 95 was similar to *N. benthamiana*, and the seven predicted SNPs (97, 98, 100, 104, 105, 107, and 108 bp) did not exist, but there was an SNP at 302 bp that was not predicted. In addition to these five varieties, some predicted SNPs also existed in the wild species, and the predictions could also be applied to different species belonging to the kindred plant.

Phylogenetic analysis

Based on these sequences, the evolutionary relationships between these varieties can be roughly understood (Figure 2). The five main varieties are all popular cultivars in China, and are very closely related. These five varieties could be separated into three groups by their SNPs (Figure 2), and SNPs also classified the wild varieties into several groups. With only one sequence smaller than 400 bp, the phylogenetic analysis may not have accurately elucidated the relationships, but in future studies that include more sequences and varieties, relationships among the *Nicotiana* will be clearer and germplasm identification and genetic diversity analysis will be much easier.

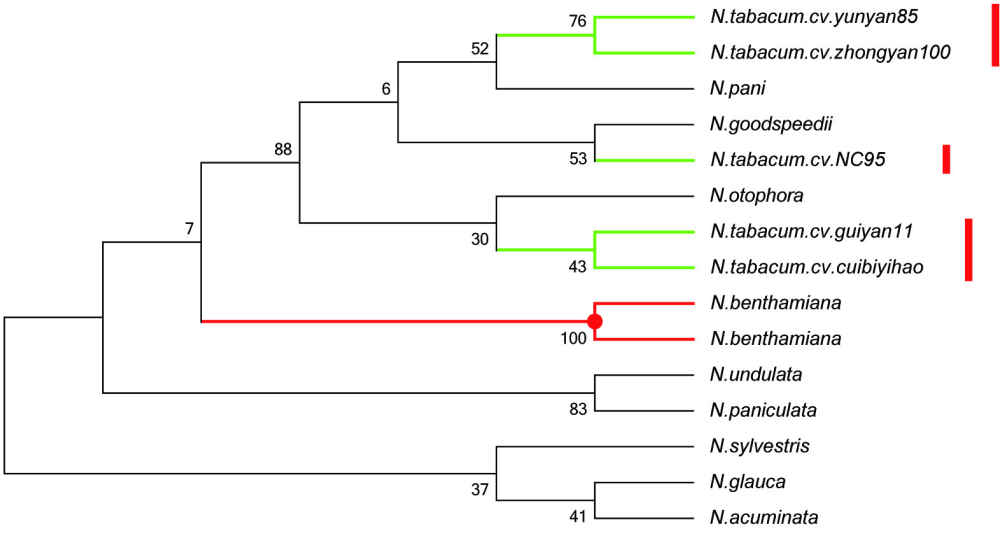

Figure 2. Evolutionary relationships between 15 varieties of *Nicotiana* based on their sequences. Varieties with vertical bars behind them are major varieties in China; *Nicotiana benthamiana* is the reference sequence and the others are wild species.

NSNP database

All of the data from this study are stored in the NSNP database (http://biodb.sdau.edu.cn/nsnp/), which consists of three main parts: "search", "tool", and "data". The "search" part provides information on potential SNPs in *Nicotiana*, including the species, the unique id, the segment, the SNP loci, and the frequency of the prediction. Users should not choose the same species, because no result will be forthcoming; some groups may not have any SNPs because of a lack of similar sequences.

The "tool" part provides a tool with which to detect SNPs. Users who have their own *Nicotiana* sequences can use this to predict SNPs in their sequences. The results are provided in a file that contains database information, the query, and the SNPs.

The "data" part is an introduction to the species involved, and an analysis of the SNPs. "Data" displays the number of SNPs, the frequency of each type of SNP, and the abundance of SNPs. Users can also determine whether the replacement rate is higher than the transversion rate.

DISCUSSION

Mutation rate and existence of the SNPs

Above all the SNPs, the lowest number of mutations were G to C and C to G, possibly because the C-G pair is more stable than the others. The results indicate that during the evolution of *Nicotiana*, the selection pressure of transversion was higher than that of replacement, or transversion had more influence than replacement during the development of the plant.

SNPs existed in all of the species at a relatively high density. This indicates that SNPs are widespread in the *Nicotiana* genome, and can be used to study the genetic diversity and evolutionary relationships of the taxon.

Analysis of the SNP at the 302-bp locus

N. tabacum cv. Zhongyan 100 contained all of the predicted SNPs and a particular SNP on 302 bp (A to T), possibly because *N. tabacum* cv. Zhongyan 100 was cross-fertilized by *N. tabacum* cv. NC 82 and *N. tabacum* cv. 9201. *N. tabacum* cv. NC 82 is an American germplasm, and is one of the selection varieties of *N. tabacum* cv. NC 95. *N. tabacum* cv. Cuibiyihao, *N. tabacum* cv. Yunyan 85, *N. tabacum* cv. Guiyan 11, and *N. tabacum* cv. 9201 are Chinese germplasms, so the SNP at the 302 bp locus may belong to the American germplasm, and the other predicted SNPs may belong to Chinese germplasms. All of these SNPs were assembled in *N. tabacum* cv. Zhongyan 100 by recombination, so the products of Primer 1 were able to distinguish between Chinese and American germplasms, and the 302-bp locus was specific to the American germplasm.

NSNP database update

The database will be regularly updated. In the future, additional ESTs and other sequences, including the whole genome sequence, will be included in the database, and other services related to SNPs will be provided.

ACKNOWLEDGMENTS

Research supported by the Open Project Program of Key Laboratory of Tobacco Pest Monitoring Controlling and Integrated Management, China, and the Agricultural Big-Data Foundation of Shandong Agricultural University, China.

REFERENCES

Bindler G, Plieske J, Bakaher N, Gunduz I, et al. (2011). A high density genetic map of tobacco (*Nicotiana tabacum* L.) obtained from large scale microsatellite marker development. *Theor. Appl. Genet*. 123: 219-230.

Bombarely A, Rosli HG, Vrebalov J, Moffett P, et al. (2012). A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. *Mol. Plant Microbe Interact*. 25: 1523-1530.

Cao H, Wang Y, Xie Z, Huang L, et al. (2013). TGB: the tobacco genetics and breeding database. *Mol. Breeding* 31: 655-663.

Delourme R, Falentin C, Fomeju BF, Boillot M, et al. (2013). High-density SNP-based genetic map development and linkage disequilibrium assessment in *Brassica napus* L. *BMC Genomics* 14: 120.

Dong Q, Schlueter SD and Brendel V (2004). PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res*. 32: D354-D359.

Goodspeed TH (1945). Cytotaxonomy of *Nicotiana* I. *Bot. Rev*. 11: 533-592.

Goodspeed TH and Thompson MC (1959). Cytotaxonomy of *Nicotiana* II. *Bot. Rev*. 25: 385-415.

Guimaraes CT, Simoes CC, Pastina MM, Maron LG, et al. (2014). Genetic dissection of Al tolerance QTLs in the maize genome by high density SNP scan. *BMC Genomics* 15: 153.

Hirakawa H, Shirasawa K, Ohyama A, Fukuoka H, et al. (2013). Genome-wide SNP genotyping to infer the effects on gene functions in tomato. *DNA Res*. 20: 221-233.

Kent WJ (2002). BLAT- the BLAST-like alignment tool. *Genome Res*. 12: 656-664.

Knapp S, Chase MW and Clarkson JJ (2004). Nomenclatural changes and a new sectional classification in *Nicotiana* (Solanaceae). *Taxon* 53: 73-82.

Larkin MA, Blackshields G, Brown NP, Chenna R, et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947-2948.

Lewis RS (2011). *Nicotiana*. In: Wild crop relatives: genomic and breeding resources. Springer, Berlin and Heidelberg, 185-208.

Li R, Li Y, Fang X, Yang H, et al. (2009). SNP detection for massively parallel whole-genome resequencing. *Genome Res*. 19: 1124-1132.

Lu F, Lipka AE, Glaubitz J, Elshire R, et al. (2013). Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genetics* 9: e1003215.

Martínez-García PJ, Parfitt DE, Ogundiwin EA, Fass J, et al. (2013). High density SNP mapping and QTL analysis for fruit quality characteristics in peach (*Prunus persica* L.). *Tree Genet. Genomes* 9: 19-36.

Qi J, Liu X, Shen D, Miao H, et al. (2013). A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet*. 45: 1510-1515.

Sierro N, Battey JN, Ouadi S, Bovet L, et al. (2013). Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biol*. 14: R60.

Tamura K, Stecher G, Peterson D, Filipski A, et al. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol*. 30: 2725-2729.

Tong Z, Yang Z, Chen X, Jiao F, et al. (2012). Large-scale development of microsatellite markers in *Nicotiana tabacum* and construction of a genetic map of flue-cured tobacco. *Plant Breed. 131*: 674-680.

Untergasser A, Cutcutache I, Koressaar T, Ye J, et al. (2012). Primer3- new capabilities and interfaces. *Nucleic Acids Res*. 40: e115-e115.

Wang S, Wong D, Forrest K, Allen A, et al. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnol. J*. 12: 787-796.