Plant Mol Biol Rep
DOI 10.1007/s11105-009-0134-z

Identification of NBS-Type Resistance Gene Homologs in Tobacco Genome

Xiaodong Leng · Bingguang Xiao · Sheng Wang · Yijie Gui · Yu Wang · Xiuping Lu · Jiahua Xie · Yongping Li · Longjiang Fan

Abstract Tobacco (*Nicotiana tabacum*) is an important cash crop and an ideal experimental system for studies on plant–pathogen interaction. The sequenced tobacco genome provides an opportunity for examining resistance gene homologs (RGHs) in the tobacco genome. Thirty nucleotide-binding site-type RGHs were annotated from genomic data, and another 281 putative RGHs were identified via PCR amplification from wild and cultivated tobacco. The newly identified RGHs are similar to other known RGHs, and some were categorized into new groups or branches that are different from known *Nicotiana* R genes or RGHs. Of the 281RGHs, 146 were identified from a single tobacco genome. We did not find any polymorphism at the RGHs in cultivated accessions, implying that strong domestication selection and/or demographic effects might have caused a sharp reduction in nucleotide diversity. Three positive selection sites were found in several RGH groups, while purifying selection is pervasive in the RGH family. Our results provide a primary RGH pool and several positively selected sites for the further functional validation of resistance genes in tobacco.

Keywords *Nicotiana tabacum* · Resistance gene homolog (RGH) · Resistance gene analog (RGA) · Positive selection · Genetic diversity

Abbreviations

RGH	resistance gene homolog
RGA	resistance gene analog
NBS	nucleotide-binding site
R	resistance
LRR	leucine-rich repeat
TGI	Tobacco Genome Initiative
CC	coiled-coil
TIR	Toll-IL-1 resistance
EST	expressed sequence tag
BLAST	basic local alignment search tool
LRT	likelihood ratio test
BEB	Bayes empirical Bayes
SNP	single nucleotide polymorphism
RAPD	random amplified polymorphic DNA
AFLP	amplified fragment length polymorphism

Introduction

Tobacco (*Nicotiana tabacum*) products have been smoked or chewed by Native Americans for thousands of years for medical and recreational purposes (Winter 2000), and today, tobacco is a widespread cash crop that plays a significant role in the economies of many countries. The genus *Nicotiana* is a member of the family Solanaceae,

Electronic supplementary material The online version of this article (doi:10.1007/s11105-009-0134-z) contains supplementary material, which is available to authorized users.

X. Leng · S. Wang · Y. Gui · Y. Wang · L. Fan (✉)
Institute of Crop Science, Zhejiang University,
Hangzhou 310029, China
e-mail: fanlj@zju.edu.cn

X. Leng · B. Xiao (✉) · X. Lu · Y. Li
Joint Laboratory of Tobacco Bioinformatics,
Yunnan Institute of Tobacco Science,
Yuxi 653100, China
e-mail: xiaobg@263.net

J. Xie
Department of Pharmaceutical Sciences,
North Carolina Central University,
1801 Fayetteville Street,
Durham, NC 27707, USA

Published online: 19 August 2009

which includes *N. tabacum* and other well-known species such as tomato, potato, pepper, and eggplant. *Nicotiana tabacum* is an amphidiploid species with a center of origin in Tropical America (Goodspeed 1954). It was believed to have arisen by the hybridization of wild progenitor species (Gerstel 1996). One hypothesis based on current molecular evidence (such as Lim et al. 2004) is that tobacco arose from an allopolyploid interspecific hybrid between *Nicotiana sylvestris* (2n=24) and *Nicotiana tomentosiformis* (2n=24), and upon doubling its chromosomes, a relatively fertile and stable hybrid was formed, which has since evolved into the modern commercial species with 48 chromosomes (2n=4×=48) and a genome of approximately 4,500 Mb.

Plant resistance (R) genes, including several R genes from tobacco, such as the *N* gene, which confers resistance to tobacco mosaic virus (Whitham et al. 1994), enable the plant to recognize the presence of specific pathogens and initiate defense responses (Dangl and Jones 2001). Most of the R genes that have been characterized thus far share similar structural motifs or belong to an ancient family that encodes proteins with nucleotide-binding site (NBS) and leucine-rich repeat (LRR) domains (Bai et al. 2002; Cannon et al. 2002; Michelmore and Meyers 1998), and, therefore, can also be detected in silico (Kanazin et al. 1996). These sequences are called resistance gene homologs (RGHs) or resistance gene analogs (RGAs) due to their sequence similarity to known R genes. RGHs are abundant in plant genomes according to sequenced genomes and PCR amplification through degenerate primers based on conserved motifs (McDowell and Simon 2006). RGH studies have been performed in the family Solanaceae.

While the analysis of RGHs in tobacco is still in its infancy, a large RGH pool has been available for several solanaceous (potato: Leister et al. 1996; coffee: Noir et al. 2001; pepper: Pflieger et al. 1999; tomato: Seah et al. 2007; *Solanum caripense*: Trognitz and Trognitz 2005) and other plants for several years. The tobacco genome was recently sequenced by an American-based project, the Tobacco Genome Initiative (TGI) (www.tobaccogenome.org) (Opperman et al. 2007). This provided us an opportunity to examine RGHs in the tobacco genome. As a result, 30 NBS-type RGHs were annotated by genome data-mining and 281 other RGHs were identified via PCR amplification (flowchart in Fig. 1a). In this study, the phylogeny, genetic diversity and positive selection of the newly identified tobacco RGHs were investigated.

Materials and Methods

Plant Materials

Twenty-five *Nicotiana* accessions were selected from a wide range of geographical locations and groups (burley, flue-cured, oriental, cigar and dark) to represent a broad genetic diversity within both cultivated tobacco (*N. tabacum*) and wild species such as *N. sylvestris*. Detailed information on the 25 accessions is shown in Table 1.

Sequence Data Sources

Tobacco genomic data (total of about 870 Mb of unassembled raw reads) from TGI (www.tobaccogenome.org, version of 04-Jan-2008) were used for our RGH annotation. Other R or RGH sequences were downloaded from GenBank. Tobacco EST sequences were retrieved from TGI, ESTobacco (European Sequencing of Tobacco Project, http://www.estobacco.info/), and the TAB project (Transcriptome Analysis of BY-2, http://mrg.psc.riken.go.jp/strc/).

PCR and DNA Sequencing

Based on the genomic sequences of 30 annotated RGHs obtained from genomic data using the Primer3 program (Koressaar and Remm 2007), 30 primer pairs, each with a genomic fragment of over 800 bp that covered the main motifs of the NBS domain (Fig. 1b), were designed (Table 2). Genomic DNA was extracted from fresh tobacco leaves using a cetyltrimethylammonium bromide (CTAB) protocol (Sambrook and Russell 1989). PCR reactions were carried out on a thermocycler (Eppendorf) under the following conditions: 95°C for 5 min, followed by 35 cycles of denaturation at 95°C for 30 s, annealing at 53°C for 30 s and extending at 72°C for 90 s, with a final extension at 72°C for 10 min. PCR products were visualized on 1.0% agarose gel. The amplified products were purified using GLASSMILK® PCR purification kits (BioDev-Tech, China). Purified PCR products were first directly sequenced from both ends using forward and reverse primers. For pairs that consisted of heterozygous members, PCR products were cloned into pGEM-T Easy Vector (TaKaRa), and at least 15 independent clones were sequenced for each primer pair. All sequences have been deposited into GenBank with accession numbers EU713645–EU713835 and EU783974–EU784068.

RGH Annotation

The protein sequences of known NBS-coding R genes (Table S1) were collected as queries to search the TGI genomic reads using the TBLASTN program (Altschul et al. 1990) (E-value threshold <1e-5). The sequences of 2,644 hits were submitted to Egassembler (http://egassembler.hgc.jp) (Masoudi-Nejad et al. 2006) for sequence cleaning and assembly. FGENESH was used to perform gene prediction (http://www.softberry.com) (Salamov and Solovyev 2000). HMMER (Eddy 1998) was used to search protein sequences

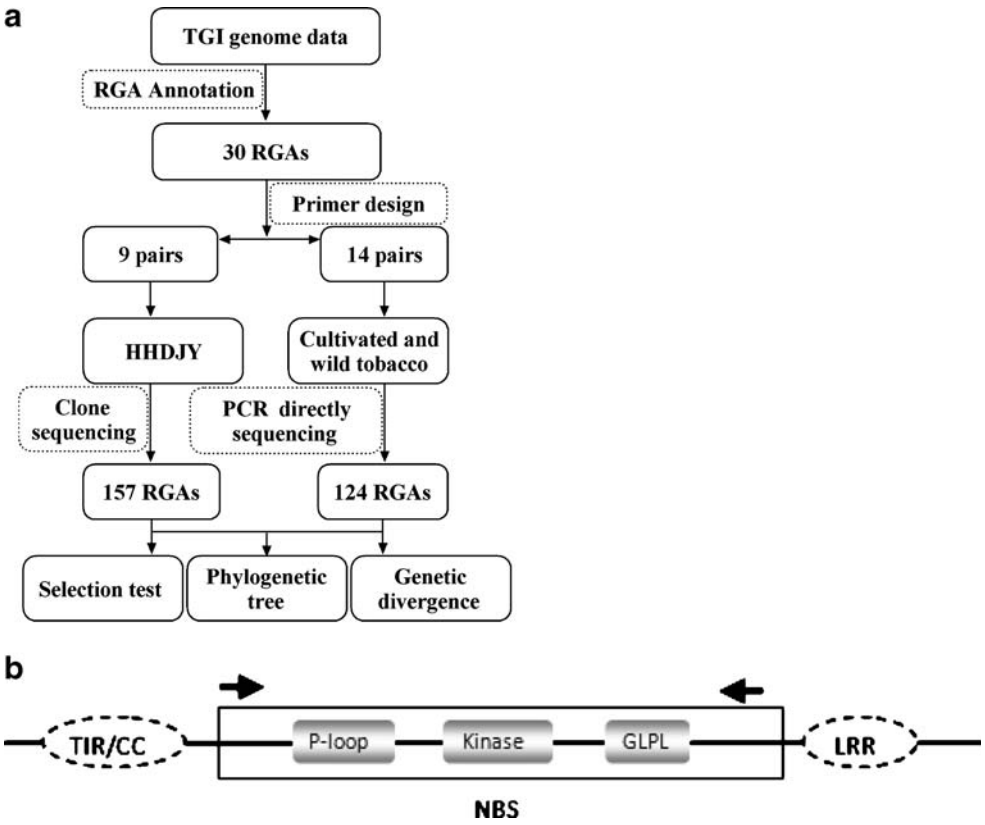

Fig. 1 Identification and analysis of resistance gene homologs (RGHs) in tobacco. **a** Flowchart for this study. **b** The structure of the plant resistance gene and the sites used for primer design. The common domain in C- and N-terminals and the three main motifs of NBS are shown. The *arrows* show the sites for primer design

for the NBS-ARC, LRR, and TIR domains from the Pfam database (www.sanger.ac.uk/Software/Pfam/) (Sonnhammer et al. 1998) and coiled-coil (CC) structure prediction was performed with NCOILS (Koretke et al. 1999), with −min_P set to 0.1 and −win set to 21, using the MTIDK matrix.

Phylogenetic and Selection Analysis

Multiple sequence alignment was performed by CLUSTALW (Higgins 1994). The phylogenetic tree was constructed by MEGA3 (Kumar et al. 2004) using the neighbor-joining algorithm. Positive selection was detected by the likelihood ratio test (LRT) using the program CODEML in the PAML packet (Yang et al. 2005).

Results

Annotation of RGHs

Only raw sequencing reads of tobacco genome were provided by TGI. Reads with high similarity to known R genes (Table S1) in the TGI dataset were determined and assembled first in our annotation efforts. In this way, a total of 2,644 reads were returned under the threshold E-value 1e-5 in our TBLASTN search. RGH genes were then predicted from the contigs using gene-finding tools, and 30 RGHs containing an intact NBS domain and an average of 2,288 bp of coding sequences were predicted (Table 3). CC-NBS-LRR-type RGHs were dominant in our annotation. Over half of the 30 RGHs had expression evidence in current tobacco EST databases (BLASTN with E-value <1e-10 and greater than 80% identity to the corresponding EST).

The 30 RGHs present a wide range of genetic diversity for the tobacco RGH family (see "Materials and Methods") and provide a good starting point for our study of RGH identification and genetic diversity in the tobacco genome (Fig. 1a).

Identification of New RGHs Via PCR Amplification

To investigate polymorphism and to identify additional RGHs in tobacco, primers covering the main motifs of the NBS domain (Fig. 1b) were designed based on the 30 annotated genes (Table 1). Nine primer pairs presented heterozygous PCR products in direct sequencing, though only one expected band was obtained (Fig. S1), suggesting that they appeared to be members of more complex RGH groups of sequences. The PCR products of the nine pairs were then cloned, and at least 15 clones were sequenced for each pair (Fig. 1a). In particular, 36 and 48 clones from two genomes were sequenced using two pairs of primers (NBS226 and NBS271). In this way, some candidate

Table 1 Plant materials used in this study

Species	Accession	Type	Origin
N. tabacum	Honghuadajinyuan (HHDJY)	Flue-cured	China
	Yunyan No.87	Flue-cured	China
	Xiaohuangjin 1025	Flue-cured	China
	Dabaijin 599	Flue-cured	China
	Changbohuang	Flue-cured	China
	Jingyehuang	Flue-cured	China
	C151	Flue-cured	China
	Hicks Broad Leaf	Flue-cured	USA
	K326	Flue-cured	USA
	Speight G-28	Flue-cured	USA
	Coker 176	Flue-cured	USA
	T.I.245	Flue-cured	USA
	Coker 371-Gold	Flue-cured	USA
	NC82	Flue-cured	USA
	Coker 86	Flue-cured	USA
	Samsun	Burley	Albania
	Burley 21	Burley	Germany
	TN86	Burley	USA
	Turkey Basma	Oriental	Turkey
	Florida 301	Cigar	USA
	Beinhart 1000-1	Cigar	USA
	TI 448A	Dark	Columbia
N. undulata	YT1215-1	Wild	Peru
N. glutinosa	YT1208-1	Wild	Peru
N. sylvestris	Y1T214-1	Wild	Argentina

sequences from the nine pairs were determined in the tobacco genomes (Table 4). The sequencing quality was manually checked for each read based on sequencing-quality files. All of the candidate RGH sequences were found to have at least one base substitution or one insertion/deletion (indel) with other sequences.

The result of phylogeny analysis indicated that the newly identified Nicotiana RGHs in this study are homologous to previously identified NBS-LRRs from other Solanaceous species (Fig. 2). In a non-TIR clade, some I2 homologs (red) have previously been investigated (Couch et al. 2006), and over 30 new homologs (such as the NBS334 and NBS267 groups) were identified by our team. NBS359 was grouped with a Prf-like gene (AAL85346) from *N. benthamiana*, while NBS226 was clustered with an R gene (Rx2) from tomato. In the TIR clade, where the R gene *N* from *N. tabacum* is located, many RGHs (e.g., NBS329, NBS414, and NBS141) were determined in this study. Moreover, some newly identified RGHs were grouped into new subfamilies or branches that can be clearly distinguished from those of known *Nicotiana* R genes or RGHs.

Evolutionary Selection

The LRT can discern positive selection that affects single amino acid residues by comparing a null model that does not allow ω (d_N/d_S, nonsynonymous/synonymous substitution rate ratio) >1 to an alternative model that does (Nielsen and Yang 1998; Suzuki and Gojobori 1999). Two LRTs were used in regular checks for selection patterns. The first test involved a null model M1a, which assumes two site classes with proportions p_0 and $p_1=1-p_0$ and a value range of $0<\omega_0<1$ and $\omega_1=1$, respectively, and an alternative model M2a, to which was added a proportion p_2 of sites with $\omega_2>1$. The second model is comparable to the null model M7 (beta), which assumes a beta distribution for ω (in the interval $0<\omega<1$), vs the alternative model M8 (beta and ω), which features an extra class of sites with positive selection ($\omega_s>1$). A Bayes empirical Bayes (BEB) procedure was used to calculate the posterior probability of a site from a particular site class, and sites with high posterior probabilities ($P>95\%$) from the class with $\omega>1$ were considered to be under positive selection (Yang et al. 2005).

Table 2 The 30 pairs of primers used for RGH amplification

Primer ID	Forward	Reverse	Expected size (bp)
NBS133	TCTTCTTACTTCCCAATACCGTAG	CCTTCATCACTTCCTCTTCATCC	858
NBS141	GATCGTTGACCAAATTTCCAAG	CAACTCTTGAGAACTTGCATGATG	858
NBS147	GCAATCCACAACTGAATGACTTG	TGGAAAAGCAAATTGGTAGGC	931
NBS148	GGCAAATGCTTATAGCTATGCTC	CAGAGGTACCAAAGATCCAGC	845
NBS155	TGGTAAACAAGAAACTAGAGGACC	GACCATTAGCAATCCACAACTG	894
NBS162	GACATGTTGAAGATCACCCGG	GCCTTTGTGACATGGTTTGAC	917
NBS173	ATCAAGCACAGTCCGAACAAC	CTCCTGAGCATTGTTGATAAGC	811
NBS178	CGATTCAACTATATCCATGCAACC	CCCATTGCTTATCCTTCTAGTTTC	845
NBS182	AAGGCGCATGCAATATCGAGG	TCATCCCACTGTGTTATGTCTTGG	978
NBS191	CTGAAACCAATTATAGACATGCCC	CACTGTGGTACAAAGTTCATTGG	897
NBS209	GGTGATCCGATAAATGGTCATAAC	TTCTTGTCAGAGATCATGGTCTC	906
NBS224	AGTGGTCGAGGTTGACTTAGTATC	GGCTTAAGATGACATGGTAAATCC	837
NBS226	GGCAATTCAAAACAAGCTCAAG	TGCATTGGACATTAACCTCTG	809
NBS228	GGAGAAGTTGGAAGGCACCAC	CGGGAAGATCATTGTAGCTCAAC	890
NBS265	CATGAAGATTCAACACAAAGGC	CCCAAGTACATGCAAACATTG	814
NBS267	CAGCAAATTGGTTGTCTTGGC	AACATCAACGCTGGTAATATGC	815
NBS270	CTCCTAAGCATTTATCTGGATGAC	AGTACCCATGATGTTGAGGATG	829
NBS271	GTCCTCTCTGGTGCCTTGAG	CACACGGTCCTCTAAAAATGC	959
NBS329	GCTGAGCTGCTTAACAATTTCTTC	CATCGACAGGATTTCTGGTTAC	927
NBS334	AAGGAGAAGTTGGAAGGCACC	GTAGCTCAACATTAACGCCGG	870
NBS345	TCGAGGAACAGATTCTGTACATC	CAATCCAATTTTCAGGTGTGAGTC	861
NBS358	CTTCTTGTTGTGAACTCTCCCAG	GGCTTGAGGTGATTAGATAGATGG	824
NBS359	GAATCCAATCACAGACATGGTC	CACTCTGGTACAAAGTTCTTTGG	903
NBS374	CTCTATCATAGCCTTTGCGTCAC	GAAGAAAACACCTGCAGTTGAC	805
NBS383	GCAATGCAATGAACACTGACATAG	TCCACAACCATGTCAACTTTG	887
NBS390	CAGCAACAACTTATTTACCTCCC	CCACCCAAATTCTAATCAACG	891
NBS391	TCAACAACTGACATGGACGTAG	TACCTTCATGCTCTGCTCTCC	803
NBS410	CTCTTTTGTATCGATTAGTCCCTC	GGATCTGCCTATGTGTTAAACTTC	966
NBS414	GCAATCTTCTCTCCCTTTCCC	TGGGCTCTAATCCATCATAACTG	923
NBS500	TGAACAAGTTGGGCCATGTA	GGTTCTCTTCGTTTCCCCTTA	935

An overview of the evolution of the nine newly identified RGH groups (Table 4) was obtained by plotting their neighbor-joining trees (Fig. S2). This figure reveals substantial variation in tree shape and depth, indicative of a wide range of evolutionary histories. This provides an opportunity to investigate evolutionary selection in these tobacco RGH groups, and particularly positive selection acting on specific amino acids, which is important for inferring gene function.

With regard to representation and the quality of sequences, the number of members in each RGH group

Table 3 Summary of 30 annotated RGHs containing an intact NBS domain in the tobacco genome

Type	Number	Average coding length (bp)	No. with EST hits[a]
CC-NBS-LRR	11	2,779	8
TIR-NBS-LRR	1	2,604	1
TIR-NBS	2	1,536	1
NBS-LRR	7	2,554	6
NBS	9	1,612	1
Total	30	2,288	16

[a] The hits were identified by BLASTN with e-value <1e-10 and identity greater than 80% of the corresponding EST.

Table 4 Tobacco RGH groups identified in this study

Group	Type	Genome	No. clones sequenced	Candidate sequences
NBS226	CC-NBS-LRR	HHDJY	21	16
		Coker176	15	14
NBS271	CC-NBS-LRR	HHDJY	33	18
		Samsun	15	11
NBS267	CC-NBS-LRR	HHDJY	15	15
NBS334	CC-NBS-LRR	HHDJY	15	13
NBS359	CC-NBS-LRR	HHDJY	15	13
NBS173	NBS-LRR	HHDJY	25	14
NBS374	NBS-LRR	HHDJY	23	13
NBS182	TIR-NBS	HHDJY	15	15
NBS329	NBS	HHDJY	15	15
Total			207	157

was reduced to guarantee at least a 1% difference between sequences among a group (Couch et al. 2006), and all members were collected from a single genome (HHDJY). In this way, paralogous sequences from eight RGH groups were used for our positive selection test (for sequence information, see Table S2). Among these eight groups, models (M2a and M8), which assumed a site class of proportion p_0 with $0<\omega_0<1$ combined with an extra class of positive sites of proportion $1-p_0$ with $\omega>1$ were statistically acceptable in three groups (NBS182/267/374) and were rejected in five other groups, indicating that the five groups belong to an alternative evolutionary model or a nearly neutral model (M1a and M7) (Table 5). Since the proportions of sites with $0<\omega_0<1$ ranged from 76.9% to 100% (average 83.8%) in the eight groups, it appears that purifying selections dominate the evolution of the eight RGH groups. Three sites in two groups were detected with significantly high probabilities from the class with $\omega>1$ as estimated by the Bayes approach (Yang et al. 2005). All three sites are located at nearby motifs (such as P-loop, Kinase2, and GLPL) of the NBS domain. For example, the two positively selected sites of NBS182, 78R and 220R, sit close to (<10 amino acids) P-loop and GLPL, respectively. In both sites, most of the members contain arginine (R), while the rest contain glutamine (Q), threonine (T), or histidine (H), implying that these mutations may play key roles in functional divergence among members of this group.

Genetic Diversity

Since unique PCR products were amplified from 14 of the 30 RGH primer pairs, they should correspond to a particular RGH locus. For the 14 loci observed, their sequence variations in cultivated and wild tobacco were investigated (Fig. 1a). Our sequencing efforts in the first two loci (NBS133 and NBS148) of the cultivated population, which includes 21 cultivated accessions from a wide range of groups (burley, flue-cured, oriental, cigar, and dark), faced a considerable challenge: no nucleotide mutations or indels could be found among them (for example, the alignment of NBS133 orthologs is shown in Fig. S3). Thus, we continued our efforts for other RGHs. Three to nine accessions were sequenced for each gene and the same result was seen again: no changes were found among these cultivated accessions. This eventually made us give up further efforts to investigate the polymorphism of tobacco RGHs. Thus, the cultivated tobacco genome is highly homozygous, as has been suggested by many previous studies based on molecular markers (see "Discussion" section).

The high genetic similarity in tobacco RGHs encouraged us to develop RGH-linked molecular markers. Forty three pairs of SSR primers were designed based on TGI genomic reads or contigs with high sequence similarity to R genes (TBALSTN, E-value <1e-5). A total of eight pairs (Table S3) that showed polymorphism among 18 cultivated accessions showed a success rate of close to 20%, suggesting that SSR markers seem to be a good choice for the development of RGH molecular markers in tobacco.

Divergence between tobacco and its wild ancestors was apparent (Fig. S3). Based on four RGH genes, an average density of 65.3 SNPs and 4.2 indels per Kb was found between tobacco and its wild ancestor *N. sylvestris* (Table 6).

Discussion

In this study, 281 putative RGHs from the tobacco genome were identified by our team, and these provide a primary RGH pool of putative R genes for further functional validation. We identified 132 candidate RGH sequences in the HHDJY genome by clone sequencing and 14 other candidate RGHs by direct PCR product sequencing. Thus, a total of 146 candidate RGH sequences were identified from

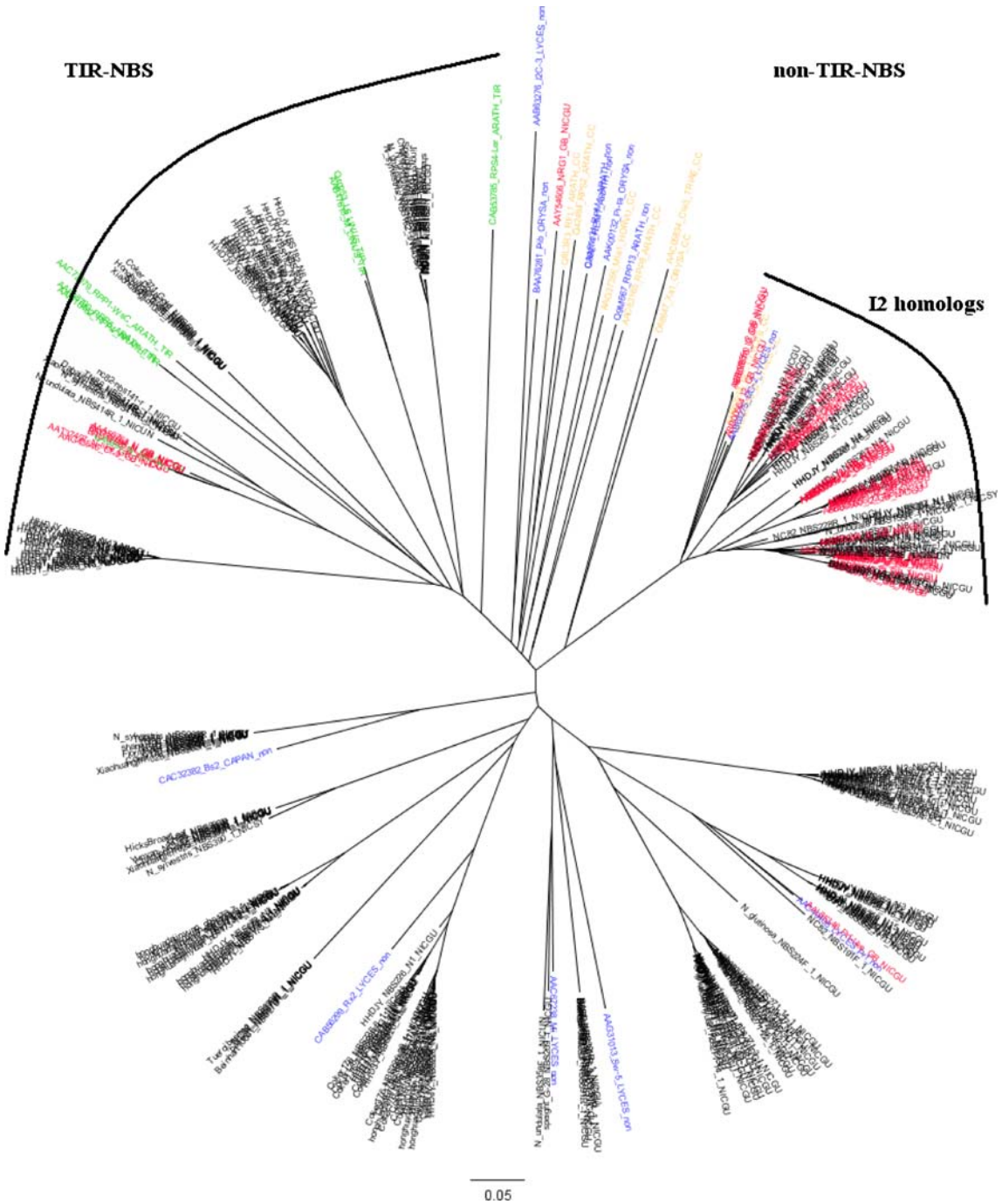

Fig. 2 A neighbor-joining phylogenetic tree of tobacco RGHs based on protein sequences. Different colors correspond to a different sequence source. *Black*: tobacco RGHs identified by this study; *red*: *Nicotiana* R or RGHs from GenBank; *others*: known R genes from other plants (see Table S1)

the tobacco genome. About 150, 250, and 400 RGHs have been annotated in Arabidopsis, rice, and Populus genomes (with genome sizes of 125, 400, and 550 Mb, respectively) (Arabidopsis Genome Initiative 2000; International Rice Genome Sequencing Project 2005; Tuskan et al. 2006). Considering that the tobacco genome has a size of 4,500 Mb, its RGH pool might be rather large. However, it must be noted that our PCR method cannot actually exclude the possibility of error of template switching between highly related RGH sequences in the nine complex RGH groups and, therefore, cannot determine the precise number of RGHs in a given group. The number of independent (not unique) RGH sequences in each group or genome still provides the fact of diversity of RGHs within each group/genome.

Table 5 Results of the LRT

Group	n[a]	P value M1a:M2a	P value M7:M8	M8 estimates[b]	Positively selected sites[c]
NBS182	8	0.026	0.026	$\hat{p}_0$=0.769, $\hat{\omega}$=2.138	78R 220R
NBS267	7	0.038	0.018	$\hat{p}_0$=0.844, $\hat{\omega}$=1.776	
NBS374	3	0.022	0.022	$\hat{p}_0$=0.992, $\hat{\omega}$=123.840	265Q
NBS334	7	>0.05	>0.05	$\hat{p}_0$=0.811, $\hat{\omega}$=1.098	
NBS226	5	>0.05	>0.05	$\hat{p}_0$=0.563, $\hat{\omega}$=1	
NBS271	3	>0.05	>0.05	$\hat{p}_0$=1, $\hat{\omega}$=138.01	
NBS329	2	>0.05	>0.05	$\hat{p}_0$=0.905, $\hat{\omega}$=5.714	
NBS173	2	>0.05	>0.05	$\hat{p}_0$=0.821, $\hat{\omega}$=1.299	

[a] Number of sequences in the group. There are at least 1% differences between two sequences from a group. For detailed information on sequences, see Table S2.

[b] $\hat{\omega}$ is d_N:d_S estimated under M8; $1-\hat{p}_0$ is the inferred proportion of positively selected sites.

[c] Position locations are based on alignments with gaps.

Our results on RGH sequences indicate that tobacco is a highly homozygous plant. In the 14 RGH loci (cumulative genomic sequence of 9.8 Kb; 700 bp per RGH), no SNP was observed. These results are consistent with previous investigations based on molecular markers (such as RAPD and AFLP), which concluded that most of the markers failed to show polymorphism within the species *N. tabacum* (e.g., Julio et al. 2006; Nishi et al. 2003; Ren and Timko 2001). In Arabidopsis, about 50.5 SNPs per Kb were observed, based on a survey of 27 R genes (Bakker et al. 2006). In selfing rice and outcrossing maize, 3.7–6.6 and 20.0–36.2 SNPs per Kb were reported based on other protein-coding genes (Tenaillon et al. 2001; Yamasaki et al. 2005; Zhu et al. 2007). Tobacco, like rice, has evolved under a highly self-pollinating reproductive mode. Many crops have experienced a severe bottleneck of domestication and strong recent positive selection and demographic effects. Both of these effects could cause a reduction in levels of nucleotide diversity during domestication (Wright and Gaut 2005). The apparent lack of molecular diversity could be related to the strong selection of tobacco during domestication. Only a few genotypes became the progenitors of most modern cultivars during tobacco domestication (Wernsman 1999).

It has been shown that positive selection could be used as an evolutionary profile that identifies NBS-coding genes that contribute to disease resistance (Bergelson et al. 2001; Mondragon-Palomino et al. 2002). In an investigation of the complete NBS-LRR gene family of *Arabidopsis thaliana*, significant positive selection was detected in NBS or LRR domains in several groups of the family (Mondragon-Palomino et al. 2002). Positive selection on LRR domains of R genes from other plants, including several positive selection sites in the I2 homologs of RGHs in the family Solanaceae, has also been revealed (Bergelson et al. 2001; Couch et al. 2006). Our results on RGHs from the I2 subfamilies (NBS334 and NBS267) are consistent with the selective situation suggested by Couch et al. (2006). Both our test results and previous observations (Couch et al. 2006; Mondragon-Palomino et al. 2002) on positive selection on NBS domains of tobacco RGHs revealed that pervasive purifying selection that characterized evolution of the RGH family, as well as positive selection, could be detected in a few scattered sites. These positive selection sites are usually the source of functional diversity for resistance. Moreover, our selective investigation based on RGH paralogs is distinct from the genetic diversity investigation of RGHs between different varieties. The

Table 6 Molecular diversity between *N. tabacum* and *N. sylvestris*

Locus	No. cultivars	Length (bp)[a]	Number		Density (per Kb)	
			SNP	Indel[b]	SNP	Indel[b]
NBS133	21	730	36	8 (75)	49.3	11.0 (102.7)
NBS162	8	879	52	3 (6)	59.2	3.4 (6.8)
NBS390	6	597	23	0	38.5	0
NBS155	4	824	94	2 (7)	114.1	2.4 (8.5)
Average	9.8	757.5	51.3	3.3 (22)	65.3	4.2 (29.5)

[a] Multiple alignment with gaps

[b] Cumulative lengths (bp) of indel are shown in parentheses

former demonstrated the evolutionary history of the RGHs within a genome under natural selection, while the latter present current diversity after recent strong domestication and genetic improvement of tobacco.

Acknowledgements This work was supported by the Yunnan Tobacco Company, which is affiliated with the China Tobacco Company (06A03). We thank Prof. Mingwei Gao (Zhejiang University) for his critical reading of the manuscript.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815. doi:10.1038/35048692

Bai J, Pennill LA, Ning J, Lee SW, Ramalingam J, Webb CA, Zhao B, Sun Q, Nelson JC, Leach JE, Hulbert SH (2002) Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res 12:1871–1884. doi:10.1101/gr.454902

Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A genome-wide survey of R gene polymorphisms in Arabidopsis. Plant Cell 18:1803–1818. doi:10.1105/tpc.106.042614

Bergelson J, Kreitman M, Stahl EA, Tian D (2001) Evolutionary dynamics of plant R-genes. Science 292:2281–2285. doi:10.1126/science.1061337

Cannon SB, Zhu H, Baumgarten AM, Spangler R, May G, Cook DR, Young ND (2002) Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. J Mol Evol 54:548–562. doi:10.1007/s00239-001-0057-2

Couch BC, Spangler R, Ramos C, May G (2006) Pervasive purifying selection characterizes the evolution of I2 homologs. Mol Plant Microbe Interact 19:288–303. doi:10.1094/MPMI-19-0288

Dangl JL, Jones JD (2001) Plant pathogens and integrated defence responses to infection. Nature 411:826–833. doi:10.1038/35081161

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763. doi:10.1093/bioinformatics/14.9.755

Gerstel DU (1996) Segregation in new allopolyploids of Nicotiana. I. Comparison of 6×(N. Tabacum×Tomentosiformis) and 6×(N. Tabacum×Otophora). Genetics 45:1723–1734

Goodspeed TH (1954) The genus nicotiana. Chronica Botanica, Waltham

Higgins DG (1994) CLUSTAL V: multiple alignment of DNA and protein sequences. Methods Mol Biol 25:307–318

Julio E, Verrier JL, Dorlhac de Borne F (2006) Development of SCAR markers linked to three disease resistances based on AFLP within Nicotiana tabacum L. Theor Appl Genet 112:335–346. doi:10.1007/s00122-005-0132-y

Kanazin V, Marek LF, Shoemaker RC (1996) Resistance gene analogs are conserved and clustered in soybean. Proc Natl Acad Sci USA 93:11746–11750. doi:10.1073/pnas.93.21.11746

Koressaar T, Remm M (2007) Enhancements and modifications of primer design program Primer3. Bioinformatics 23:1289–1291. doi:10.1093/bioinformatics/btm091

Koretke KK, Russell RB, Copley RR, Lupas AN (1999) Fold recognition using sequence and secondary structure information. Proteins 37(Suppl 3):141–148. doi:10.1002/(SICI)1097-0134(1999)37:3+<141::AID-PROT19>3.0.CO;2-F

Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5:150–163. doi:10.1093/bib/5.2.150

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800. doi:10.1038/nature03895

Leister D, Ballvora A, Salamini F, Gebhardt C (1996) A PCR-based approach for isolating pathogen resistance genes from potato with potential for wide application in plants. Nat Genet 14:421–429. doi:10.1038/ng1296-421

Lim K, Matyasek R, Kovarik A, Leitch AR (2004) Genome evolution in allotetraploid Nicotiana. Biol J Linn Soc Lond 82:599–606. doi:10.1111/j.1095-8312.2004.00344.x

Masoudi-Nejad A, Tonomura K, Kawashima S, Moriya Y, Suzuki M, Itoh M, Kanehisa M, Endo T, Goto S (2006) EGassembler: online bioinformatics service for large-scale processing, clustering and assembling ESTs and genomic DNA fragments. Nucleic Acids Res 34:W459–W462. doi:10.1093/nar/gkl066

McDowell JM, Simon SA (2006) Recent insights into R gene evolution. Mol Plant Pathol 7:437–448. doi:10.1111/j.1364-3703.2006.00342.x

Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res 8:1113–1130

Mondragon-Palomino M, Meyers BC, Michelmore RW, Gaut BS (2002) Patterns of positive selection in the complete NBS-LRR gene family of Arabidopsis thaliana. Genome Res 12:1305–1315. doi:10.1101/gr.159402

Nielsen R, Yang Z (1998) Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. Genetics 148:929–936

Nishi T, Tajima T, Noguchi S, Ajisaka H, Negishi H (2003) Identification of DNA markers of tobacco linked to bacterial wilt resistance. Theor Appl Genet 106:765–770

Noir S, Combes MC, Anthony F, Lashermes P (2001) Origin, diversity and evolution of NBS-type disease-resistance gene homologues in coffee trees (Coffea L.). Mol Genet Genomics 265:654–662. doi:10.1007/s004380100459

Opperman CH, Lommel S (2007) The Tobacco Genome Initiative: Gene discovery and data mining in Nicotiana tabacum. In: Plant & Animal Genomes XIV Conference, San Diego

Pflieger S, Lefebvre V, Caranta C, Blattes A, Goffinet B, Palloix A (1999) Disease resistance gene analogs as candidates for QTLs involved in pepper–pathogen interactions. Genome 42:1100–1110. doi:10.1139/gen-42-6-1100

Ren N, Timko MP (2001) AFLP analysis of genetic polymorphism and evolutionary relationships among cultivated and wild Nicotiana species. Genome 44:559–571. doi:10.1139/gen-44-4-559

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res 10:516–522. doi:10.1101/gr.10.4.516

Sambrook J, Russell DW (1989) Molecular cloning: A laboratory manual, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

Seah S, Telleen AC, Williamson VM (2007) Introgressed and endogenous Mi-1 gene clusters in tomato differ by complex rearrangements in flanking sequences and show sequence exchange and diversifying selection among homologues. Theor Appl Genet 114:1289–1302. doi:10.1007/s00122-007-0519-z

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R (1998) Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res 26:320–322. doi:10.1093/nar/26.1.320

Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16:1315–1328

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (Zea mays ssp. mays L.). Proc Natl Acad Sci USA 98:9161–9166. doi:10.1073/pnas.151244298

Trognitz F, Trognitz BR (2005) Survey of resistance gene analogs in Solanum caripense, a relative of potato and tomato, and update on R gene genealogy. Mol Genet Genomics 274:595–605. doi:10.1007/s00438-005-0038-z

Tuskan GA et al (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313:1596–1604. doi:10.1126/science.1128691

Wernsman E (1999) An overview of tobacco breeding-past. Present, and future. Recent Adv Tob Sci 25:12–15

Whitham S, Dinesh-Kumar SP, Choi D, Hehl R, Corr C, Baker B (1994) The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor. Cell 78:1101–1115. doi:10.1016/0092-8674(94)90283-6

Winter JC (2000) Tobacco use by native North Americans—sacred smoke and silent killer. University of Oklahoma Press, Norman

Wright SI, Gaut BS (2005) Molecular Population Genetics and the Search for Adaptive Evolution in Plants. Mol Biol Evol 22:506–519. doi:10.1093/molbev/msi035

Yamasaki M, Tenaillon MI, Bi IV, Schroeder SG, Sanchez-Villeda H, Doebley JF, Gaut BS, McMullen MD (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17:2859–2872. doi:10.1105/tpc.105.037242

Yang Z, Wong WSW, Nielsen R (2005) Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol Biol Evol 22:1107–1118. doi:10.1093/molbev/msi097

Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol 24:875–888. doi:10.1093/molbev/msm005