Long *et al. J of Biol Res-Thessaloniki (2016) 23:6*
DOI 10.1186/s40709-016-0048-5

Journal of Biological Research-Thessaloniki

RESEARCH Open Access

Sequencing and characterization of leaf transcriptomes of six diploid *Nicotiana* species

Ni Long[1], Xueliang Ren[2], Zhidan Xiang[3], Wenting Wan[1] and Yang Dong[1,4*]

Abstract

Background: *Nicotiana* belongs to the Solanaceae family that includes important crops such as tomato, potato, eggplant, and pepper. *Nicotiana* species are of worldwide economic importance and are important model plants for scientific research. Here we present the comparative analysis of the transcriptomes of six wild diploid *Nicotiana* species. Wild relatives provide an excellent study system for the analysis of the genetic basis for various traits, especially disease resistance.

Results: Whole transcriptome sequencing (RNA-seq) was performed for leaves of six diploid *Nicotiana* species, i.e. *Nicotiana glauca*, *Nicotiana noctiflora*, *Nicotiana cordifolia*, *Nicotiana knightiana*, *Nicotiana setchellii* and *Nicotiana tomentosiformis*. For each species, 9.0–22.3 Gb high-quality clean data were generated, and 67,073–182,046 transcripts were assembled with lengths greater than 100 bp. Over 90 % of the ORFs in each species had significant similarity with proteins in the NCBI non-redundant protein sequence (NR) database. A total of 2491 homologs were identified and used to construct a phylogenetic tree from the respective transcriptomes in *Nicotiana*. Bioinformatic analysis identified resistance gene analogs, major transcription factor families, and alkaloid transporter genes linked to plant defense.

Conclusions: This is the first report on the leaf transcriptomes of six wild *Nicotiana* species by Illumina paired-end sequencing and de novo assembly without a reference genome. These sequence resources hopefully will provide an opportunity for identifying genes involved in plant defense and several important quality traits in wild *Nicotiana* and will accelerate functional genomic studies and genetic improvement efforts of *Nicotiana* or other important Solanaceae crops in the future.

Keywords: *Nicotiana*, Transcriptome, De novo assembly, Phylogenetic relationship, *Nicotiana setchellii*, *Nicotiana cordifolia*, *Nicotiana knightiana*, *Nicotiana tomentosiformis*, *Nicotiana noctiflora*, *Nicotiana glauca*

Background

The genus *Nicotiana* is a member of the Solanaceae or nightshade family, which includes many economically important crop plants such as tomato, potato, eggplant, and pepper. According to Goodspeed [1] and Goodspeed & Thompson [2], *Nicotiana* was initially divided into three subgenera and 14 sections. Recently, this genus was reclassified into 13 sections based on morphological, cytological, and DNA sequence data [3, 4]. *Nicotiana* includes over 75 naturally occurring species, almost half of which are allopolyploid [3]. The genus *Nicotiana* contains species of scientific and economic importance, with different evolutionary histories resulting to highly complex genomes [5]. Of all species, only *Nicotiana tabacum* (common tobacco) and *Nicotiana rustica* are cultivated worldwide, whereas the others are wild species. Moreover, *Nicotiana benthamiana* is used extensively as a model to study plant-pathogen interactions. Several other species, such as *Nicotiana alata* and *Nicotiana sylvestris,* are grown as ornamentals. In *N. tabacum* breeding programs, wild *Nicotiana* species are valuable sources for identifying genes involved in disease and pest resistance, important quality traits, and phytochemicals, which are not present in cultivated varieties [6].

*Correspondence: dongyang@dongyang-lab.org
[1] Faculty of Life Science and Technology, Kunming University of Science and Technology, South Jingming Road No.727, Kunming 650500, Yunnan, China
Full list of author information is available at the end of the article

Plants are constantly under the attack of bacteria, fungi, viruses, nematodes and insect pests. Some of them have successfully invaded crop plants, causing diseases and reducing crop quality and yield. To protect against pathogens, plants have evolved various defense mechanisms. Plant disease resistance (R) genes play a key role in defending plants from a range of pathogens. For instance, N genes from tobacco confer resistance to tobacco mosaic virus (TMV) [7]. In recent years, a set of 112 known and 104,310 putative R genes fighting against 122 different pathogens have been identified in 233 plant species [8]. Most of the characterized R genes share a few highly conserved domains, including nucleotide binding site (NBS), leucine-rich repeat (LRR), Toll/Interleukin-1 receptor (TIR) and coiled-coil (CC) domains [9–11]. These conservative domains provide convenient and reliable means for rapidly identifying and cloning R genes or resistance gene analogs (RGAs).

Identification of *Nicotiana* R genes and RGAs cannot only help elucidate the molecular mechanisms of host-pathogen interaction, but also benefit breeding programs for disease resistance in *Nicotiana* and other important Solanaceae crops. Transcriptomic sequences can be useful substitutes for gene discovery in species without sequenced genomes. In the past, a large RGA pool has been mined from transcriptomic sequences and expressed sequence tags (ESTs) of coffee [12], *Phaseolus vulgaris* [13], *Curcuma longa* [14] and *Cocos nucifera* [15]. Wild *Nicotiana* species are known to resist a variety of pathogens. For example, *N. glauca* has attractive potentials to resist black root rot (BRR), potato virus Y (PVY), tobacco etch virus (TEV), anthracnose (An), powdery mildew (PM), rattle virus (RV) and tobacco streak virus (TS) [16–18]. *Nicotiana noctiflora* is resistant to PM and PVY. *Nicotiana cordifolia* shows resistance to TS. *Nicotiana knightiana* manifests high resistance to An, PM, root knot nematodes (RK), PVY and TEV. *Nicotiana setchellii* shows resistance to RV and TEV. *Nicotiana tomentosiformi*s is resistant to cyst nematodes (CN), RK, RV and TEV [16, 17]. These observations suggest that wild *Nicotiana* species are excellent depositories of R genes and RGAs, but relevant analyses of these genes have been lacking.

In *Nicotiana* species, alkaloids (e.g. nicotine) are believed to function as a chemical defense mechanism against pathogens and herbivores. Nicotine and related pyridine alkaloids are synthesized in the tobacco root and then translocated to the aerial parts of the plant [19, 20]. Thus the translocation of nicotine from the root to the leaves is very important in tobacco defenses.

Comparative studies of closely related species can advance our understanding of the genetic architecture of adaptive traits. So far, such studies have been very limited for several crops including tobacco. This is mainly due to the lack of genomic resources hampering the development of genetic markers for investigating species divergence, adaptation and demographic processes in natural populations.

In the present study, we selected six wild *Nicotiana* species for analyses, which included *N. glauca, N. noctiflora, N. cordifolia, N. knightiana, N. setchellii,* and *N. tomentosiformis*. These diploid *Nicotiana* species (all with chromosome numbers of $2n = 24$) were chosen because they are repositories of pathogen resistant genes (Table 1). The six wild *Nicotiana* species belong to three sections: *Noctiflorae*, *Paniculatae* and *Tomentosae*. Trait introgression from wild relatives has been used to improve crop species. For example, characters from at least 13 different species have been transferred into tobacco [4]. With advances in next-generation sequencing (NGS) technologies, genomic data for several *Nicotiana* species have become available [21–25]. These data revealed that some *Nicotiana* genomes are large compared with other Solanaceae species such as the tomato [5]. For most wild *Nicotiana* species, very few genomic sequences are currently available. In this study, we performed transcriptome sequencing using the Illumina paired-end sequencing technique with the aim of identifying expressed RGAs, transcription factors important in plant defense, and alkaloid transporter genes by data mining. Our results will provide a useful basis for future identification and cloning of interest genes in wild *Nicotiana* and contribute to the improvement of cultivated tobacco and other important Solanaceae crops.

Results and discussion

Assembly of RNA-seq reads and evaluation

The Illumina paired-end sequencing yielded 100 bp paired-end independent reads from each insert of cDNA. After stringent quality assessment and data filtering, reads with Q20 bases (those with a base quality greater

Table 1 Summary of the six wild *Nicotiana* species investigated in this study

Species	Sections	Subgenus	Resistance to diseases
N. glauca	*Noctiflorae*	*Petunioides*	BRR, An, PM, RV, TEV, TS, PVY
N. noctiflora	*Noctiflorae*	*Petunioides*	PM, PVY
N. cordifolia	*Paniculatae*	*Rustica*	TS
N. knightiana	*Paniculatae*	*Rustica*	An, PM, RK, TEV, PVY
N. setchellii	*Tomentosae*	*Tabacum*	RV, TEV
N. tomentosiformis	*Tomentosae*	*Tabacum*	CN, RK, RV, TEV

BRR black root rot, *An* anthracnose, *CN* cyst nematodes, *PM* powdery mildew, *RK* root-knot nematodes, *RV* rattle virus, *TS* tobacco streak virus, *PVY* potato virus Y, *TEV* tobacco etch virus

than 20) were selected as high quality reads for further analysis. In this study, 9.0–22.3 Gb of clean data were generated for each sample (Additional file 1). Due to the lack of reference genome information, Trinity was used for de novo assembly of the six wild *Nicotiana* species [26]. We ultimately obtained 182,046, 146,188, 134,519, 67,073, 102,935 and 117,640 transcripts with length >100 bp for *N. glauca, N. noctiflora, N. cordifolia, N. knightiana, N. setchellii* and *N. tomentosiformis,* respectively (Additional file 1). Subsequently, open reading frames (ORFs) were predicted and the transcripts were translated into peptides culled at a minimum length of 100 amino acids. Only ORFs longer than 300 bp were considered to be possible protein-encoding transcripts and 33,995–79,449 ORFs were obtained through this process for the studied species (see Additional file 2). Although the ORFs of the six wild *Nicotiana* species varied within a large range, from 33,995 to 79,449, after removing redundancy due to alternative splicing isoforms, the ORFs ranged from 22,168 to 29,356 (*N. glauca* 22,934*, N. noctiflora* 26,788*, N. cordifolia* 29,356*, N. knightiana* 22,168*, N. setchellii* 26,579 and *N. tomentosiformis* 24,213).

In the absence of a reference genome, evaluating the quality of the de novo assembled transcriptomes becomes a tedious job. To resolve it, we marked *N. tomentosiformis* as a reference. A total of 53,753 reported peptide sequences (ftp://solgenomics.net/genomes/Nicotiana_tomentosiformis/annotation/, Accessed 27th Apr 2015) were blasted [27] against our predicted ORFs of *N. tomentosiformis* using BLASTp with a cut-off e-value of 10^{-5}. A total of 50,390 (93.74 %) *N. tomentosiformis* proteins had a BLAST hit in our ORFs and 32,761 (60.95 %) proteins showed ≥90 % identity with more than 50 % matched length of the corresponding proteins, which suggests our assembly should be largely complete. Moreover, ORFs were compared to the core eukaryote gene (CEG) set of 248 proteins from six reference species [28] to assess the quality of each transcriptome. The CEGs were well-represented in the assembled transcriptomes of the *N. glauca*, *N. noctiflora*, *N. cordifolia*, *N. knightiana*, *N. setchellii*, *N. tomentosiformis*, with significant matches (alignment length ≥50 % CEG length and e-value <10^{-5}) to 87.10, 92.34, 91.94, 89.92, 90.73 and 91.53 % of the CEGs, respectively. This indicated that the quality and completeness of our transcriptome assemblies were high enough for subsequent analyses. These transcriptome sequences may greatly enrich the *Nicotiana* sequence database, and will be useful in trait-related gene mining, such as the identification of plant defense genes.

Transcriptome annotation and expression analysis

To obtain the most informative and complete annotation, ORFs from six species of *Nicotiana* were annotated separately. Sequence similarity searches were conducted against the NCBI NR and Swiss-Prot databases using the BLASTp algorithm with a cutoff e-value of 10^{-5}. Using this approach, 94.68–97.43 % ORFs showed homology with sequences in the NR database (Table 2) and 71.06–77.76 % ORFs returned significant matches in the Swiss-Prot database (Table 2). The e-value distribution of the top hits in the Swiss-Prot database showed that 60.78 % of the mapped sequences had a strong homology (smaller than 10^{-5}, Additional file 3). The remaining un-annotated ORFs appeared to be either *Nicotiana*-specific genes or homologous genes with unknown functions in other species.

Besides, a higher (>90 %) match rate in the NR database was shown by ORFs with >200 aa in length, whereas ORFs shorter than 200 aa exhibited a lower match rate (Fig. 1). An almost similar match rate pattern was observed in the annotation for Swiss-Prot database (Additional file 4).

The expression level of each ORF from six wild *Nicotiana* species was normalized and quantified by the FPKM (fragments per kilobase per million sequenced reads)

Table 2 Summary of functional annotation of predicted ORFs

	N. glauca	*N. noctiflora*	*N. cordifolia*	*N. knightiana*	*N. setchellii*	*N. tomentosiformis*
ORF	65,423	63,930	79,449	33,995	52,916	53,121
NR	63,133	63,930	75,226	33,121	50,490	51,207
Percentage	96.50 %	95.88 %	94.68 %	97.43 %	95.42 %	96.40 %
SwissProt	47,743	47,197	56,453	26,434	38,954	40,138
Percentage	72.98 %	73.83 %	71.06 %	77.76 %	73.61 %	75.56 %
KEGG	37,385	37,007	43,884	20,439	29,817	30,858
Percentage	57.14 %	57.89 %	55.24 %	60.12 %	56.35 %	58.09 %
GO	36,508	35,846	43,279	20,370	30,079	27,592
Percentage	55.80 %	56.07 %	54.47 %	59.92 %	56.84 %	51.94 %

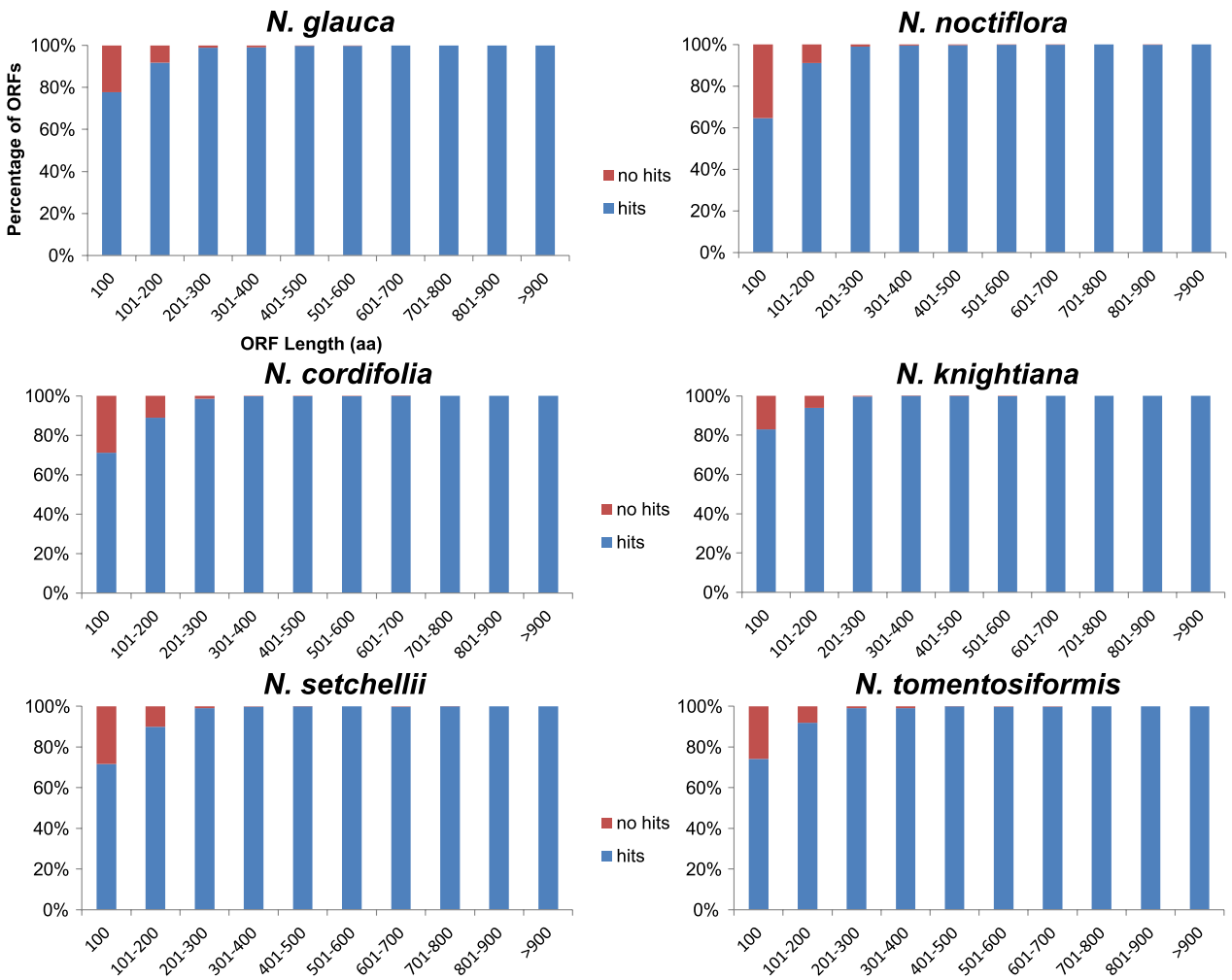

Fig. 1 Comparison of ORF length between hit and no hit proteins in NR database. For *N. glauca*, *N. noctiflora*, *N. cordifolia*, *N. knightiana*, *N. setchellii*, *N. tomentosiformis*, longer ORFs were more likely to have BLASTp homologs in protein database

method (Additional file 5). The ORFs with FPKM <1 were considered to be unexpressed, ORFs with FPKM values between 1 and 3 were considered lowly expressed, those between 3 and 15 were considered expressed at medium levels, and those with FPKM values >60 were considered highly expressed (Table 3). The top 20 ORFs with highest FPKM values for each species can be seen in Additional file 6. These ORFs either encode chloroplast proteins or play role in photosynthesis. These results are consistent with the fact that leaves are the plant's main photosynthetic organs.

Phylogenetic analysis

Large-scale transcriptome data are a potential source of information for multigene phylogenetic analysis (the phylogenomic approach). In this study, 2491 single copy orthologs were identified and two phylogenetic trees were constructed by the neighbor-joining (NJ) method in Phylip [29] (Fig. 2) and maximum likelihood (ML) method in PhyML [30] (Additional file 7). The two phylogenies showed identical topologies. Earlier, Goodspeed placed *N. glauca* in the section *Paniculatae* based on evidence from morphology, cytology, biogeography, and crossing experiments [31]. Later, *N. glauca* was placed in the section *Noctiflorae* based on analysis of sequences from internal transcribed spacer (ITS) of nuclear ribosomal DNA (nrDNA) [3, 32]. Current phylogenetic analysis of the transcriptomes from six diploid species of *Nicotiana* with *Solanum lycopersicum* (tomato) as an outgroup supported the *Noctiflorae*, *Paniculatae,* and *Tomentoase* clades. The phylogenetic trees obtained in the current study placed *N. glauca* in the section *Noctiflorae*, supporting the results of previous works by Chase et al. [32]. and Knapp et al. [3].

Table 3 Distribution of ORF expressions in six wild *Nicotiana* species

FPKM interval	*N. glauca*	*N. noctiflora*	*N. cordifolia*	*N. knightiana*	*N. setchellii*	*N. tomentosiformis*
0–1	31,857 (48.69 %)	28,985 (45.34 %)	33,698 (42.41 %)	6245 (18.37 %)	16,346 (30.89 %)	18,670 (35.15 %)
1–3	11,832 (18.09 %)	11,105 (17.37 %)	16,828 (21.18 %)	2936 (8.63 %)	11,577 (21.88 %)	11,338 (21.34 %)
3–15	11,696 (17.88 %)	12,299 (19.24 %)	16,812 (21.16 %)	14,986 (44.08 %)	13,912 (26.29 %)	12,948 (24.37 %)
15–60	6710 (10.26 %)	7708 (12.06 %)	8232 (10.36 %)	6782 (19.95 %)	7495 (14.16 %)	6836 (12.87 %)
>60	3328 (5.09 %)	3833 (5.10 %)	3879 (4.88 %)	3046 (8.96 %)	3586 (6.78 %)	3329 (6.27 %)

Ratios of ORF number to total ORF number are presented in parentheses

FPKM fragments per kilobase per million sequenced reads

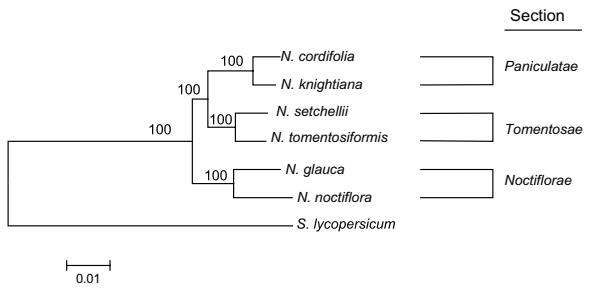

Fig. 2 Phylogenetic tree based on the transcriptomes of the six wild *Nicotiana* species and *S. lycopersicum*. Phylogenetic tree was constructed using the neighbor-joining method with 1000 bootstraps. Bootstrap support is shown at the nodes

Functional classification by KEGG

ORFs of six wild *Nicotiana* species were compared with KEGG (Kyoto Encyclopedia of Genes and Genomes) database using BLASTp with an e-value less than 10^{-5}, and the corresponding pathways were established. For the six species, 55.24–60.12 % of ORFs were successfully annotated to KEGG pathways (Table 2). Genes within the same pathway usually cooperate with each other to exercise their biological function, and hence pathway-based analysis contributes to the exploration of biological functions and interactions of genes [33]. The sequence annotation in KEGG largely contained metabolic pathways of major biomolecules such as carbohydrates, amino acids, lipids, nucleotides, etc. (Fig. 3a). The metabolic pathways with most representation by proteins were those of carbohydrate metabolism and amino acid metabolism. In the secondary metabolism, for *N. glauca, N. noctiflora, N. cordifolia, N. knightiana, N. setchellii, N. tomentosiformis,* 687, 817, 1064, 518, 779 and 677 proteins were classified into 14 subcategories, respectively (Fig. 3b). Among them, the cluster for "Phenylpropanoid biosynthesis" represents the largest group followed by "Stilbenoid, diarylheptanoid and gingerol biosynthesis". The phenylpropanoid pathway is often considered to be involved in plant resistance [34]. Flavonoids and glucosinolates are secondary metabolites that play important roles in protecting plants against pathogens. We also found unigenes involved in the biosynthesis of flavonoid and glucosinolate. The results will facilitate the discovery of novel genes involved in the specific metabolic pathways and secondary metabolic pathways and will provide a valuable resource for investigating the defense-related pathways in *Nicotiana* and other Solanaceae species.

Functional classification by GO

Gene ontology (GO) [35] provides ontologies of defined terms representing gene product properties and describes gene products in terms of their associated biological processes, cellular components, and molecular functions. In this study, 36,508, 35,846, 43,279, 20,370, 30,079 and 27,592 annotated ORFs corresponding to *N. glauca*, *N. noctiflora*, *N. cordifolia*, *N. knightiana*, *N. setchellii*, and *N. tomentosiformis*, respectively, were assigned to one or more sub-categories of GO terms. The GO terms of the subcategories are presented in Fig. 4. For the six wild *Nicotiana* species, among these groups, genes involved in "metabolic process" and "cellular process" were the most highly represented in the biological process category. Genes involved in other important biological processes such as biological regulation, response to stimulus, and anatomical structure formation process were also identified. Furthermore, a relatively large number of sequences were found to be involved in the metabolism of pigmentation. Within the cellular components category, "cell" and "cell parts" were the most highly represented groups. The molecular function category comprised proteins involved in "binding" and "catalytic activity". These six wild *Nicotiana* transcriptomes shared broad similarities in the three main categories and many subcategories except viral reproduction.

Identification of NBS encoding genes and defense response associated transcription factors

The majority of disease resistance genes in plants contain a nucleotide-binding site and leucine-rich repeat (NBS-LRR) domain [36, 37], which confers resistance to

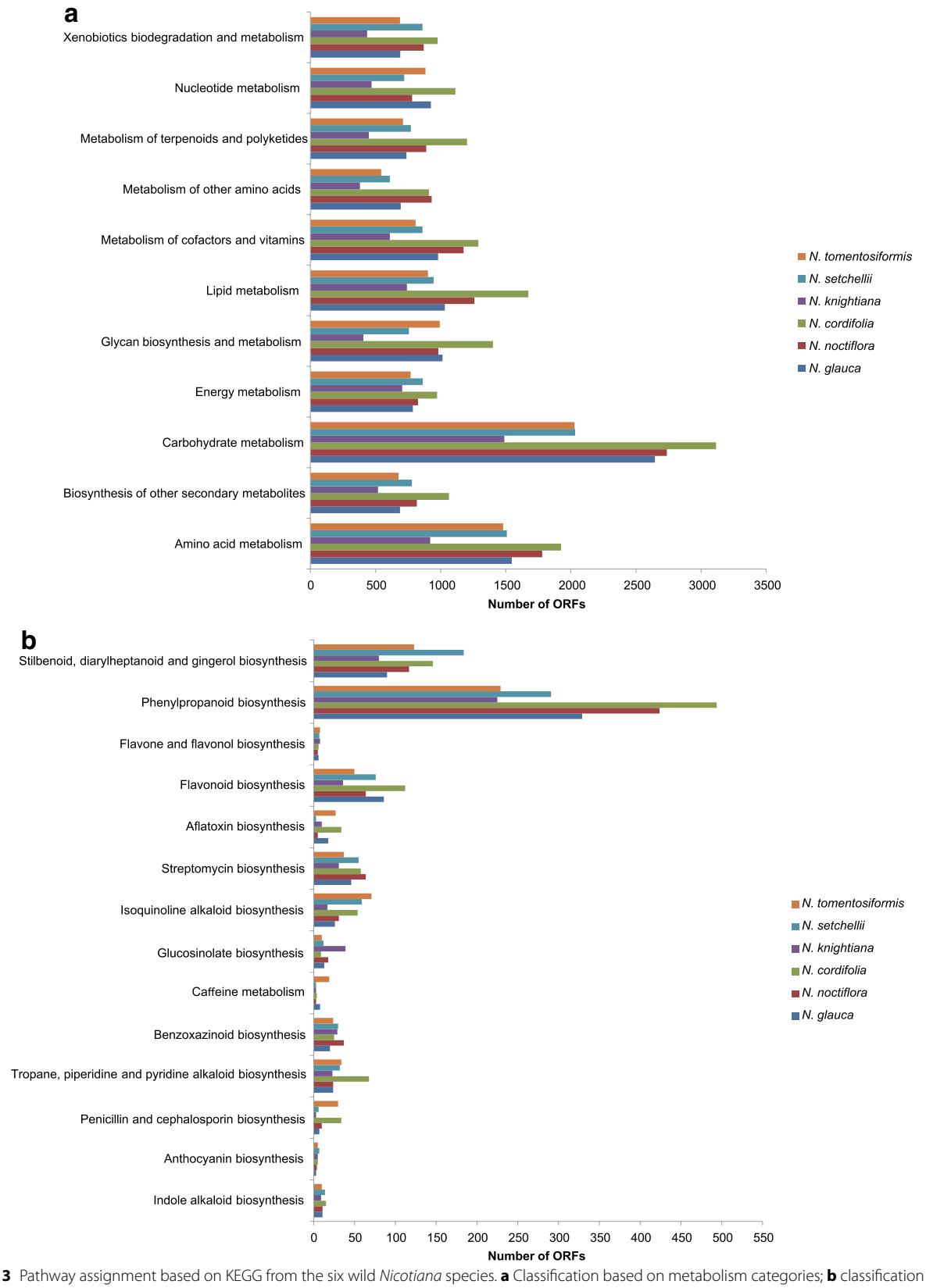

Fig. 3 Pathway assignment based on KEGG from the six wild *Nicotiana* species. **a** Classification based on metabolism categories; **b** classification based on secondary metabolism categories

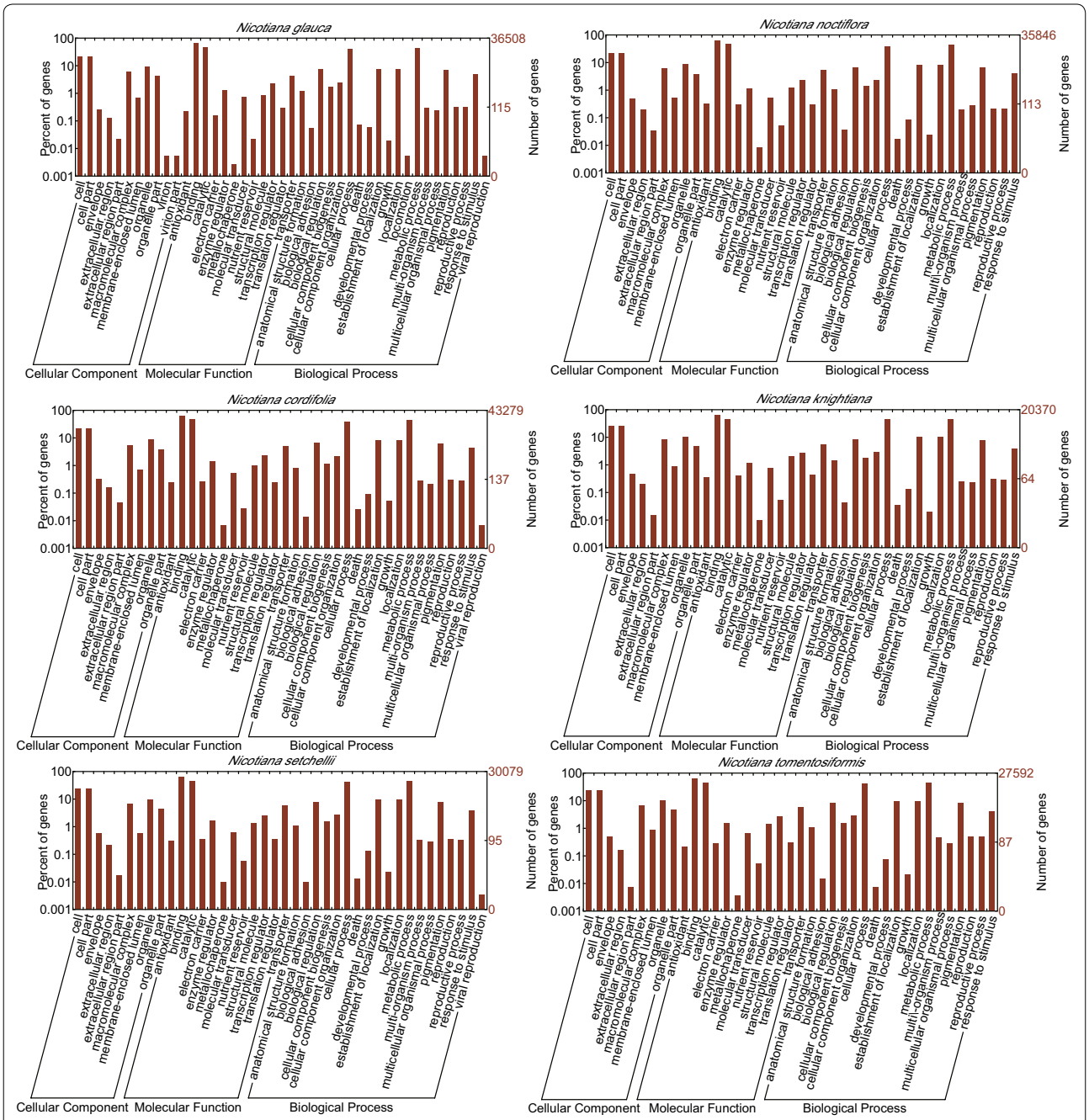

Fig. 4 Histogram presentation of GO classification. The GO annotation results from the ORFs of six *Nicotiana* species are summarized in three main categories: biological process, cellular component and molecular function. The right *y-axis* indicates the number of genes in a category. The left *y-axis* indicates the percentage of a specific category of genes in that main category

fungi, bacteria, viruses, and nematodes. In plants, based on the presence or absence of a TIR homology region at the N-terminus, the NBS-LRR genes can be subdivided into two main groups: TIR-NBS-LRR and non-TIR-NBS-LRR. The latter may have a coiled-coil (CC) motif in the N-terminal region and can be called as CC-NBS-LRR.

To control diseases in certain agriculturally important plants, the identification of resistance genes from their less susceptible relatives has been the top priority in crop breeding programs. In the case of Solanaceae species, the pepper *Bs2* gene with NBS-LRR domain was introduced into tomato lines to develop resistance against bacterial

spot disease [38]. In tobacco, the TIR-NBS-LRR encoding N gene was introduced into *N. benthamiana*, which resulted in the acquirement of hypersensitivity response to tobacco mosaic virus (TMV) [39].

In this study, after going through a filtering process, 87–173 unigenes encoding NBS domains were identified from the six wild *Nicotiana* species. These NBS-encoding genes were classified into six classes on the presence or absence of CC domain, TIR domain, and/or LRR domain. These six classes include CC-NBS-LRR, CC-NBS, TIR-NBS-LRR, TIR-NBS, NBS-LRR, and NBS (Table 4). The NBS class was the most represented class (61–113 unigenes) for all six species in the present study. The TIR-NBS class had 3–8 unigenes for each species, and the NBS-LRR class had 5–14 unigenes (0 for *N. tomentosiformis*). Additionally, 2–11 unigenes (0 for *N. cordifolia*) were predicted to encode TIR-NBS-LRR, 11–31 unigenes were identified as CC-NBS, and 3–7 unigenes contained CC-NBS-LRR. The candidate R genes will enhance our knowledge about the mechanisms of disease resistance in Solanaceae species and help breed novel disease resistant varieties.

Transcription factors (TFs) are also important in disease resistance. They bind to the promoters of resistance genes and regulate their expression. The TFs related to defense or disease resistance mainly belong to the MYB [40], WRKY [41], bZIP [42] and Whirly [43] families. Overexpression of the defense-related TFs has improved disease resistance in many transgenic crops [44]. By using Pfam annotations, we identified 439–618 candidate unigenes matching the defense-related TFs in the six wild species of *Nicotiana* (Table 5). These candidate TFs will be potential targets for developing the resistant lines of tobacco and other Solanaceae crops.

Identification of alkaloid transporter genes

Alkaloids are mainly produced in the root and then translocated via xylem transport towards the aerial parts. These toxic chemicals function as part of the chemical defense against invaders [19, 20]. To date, the plant alkaloid transporters are mainly characterized into the ATP-binding cassette (ABC) protein, multidrug and toxic compound extrusion (MATE), and purine permease (PUP) families. Some transporters were found to be required for the efficient biosynthesis of alkaloids in plants [45]. In tobacco, several alkaloid transporter genes have been identified, such as tobacco jasmonate-inducible alkaloid tranporter1 (*Nt-JAT1*), *Nt-JAT2*, tobacco nicotine uptake permease1 (*Nt-NUP1*), *NtMATE1* and *NtMATE2* [46–50].

In the present study, we began our investigation by searching the assembled transcriptome for orthologous genes to known alkaloid transporter genes in the tobacco. *Nt-JAT1* transports nicotine and other alkaloids in a proton gradient-dependent manner. *Nt-JAT1* mRNA is expressed in leaves, stems, and roots. In leaf cells, *Nt-JAT1* localizes to the tonoplast and might play a role in the vacuolar sequestration of nicotine [46]. We found one orthologous gene for *Nt-JAT1* in the assembled *N. cordifolia*, *N. setchellii* and *N. tomentosiformis* transcriptomes with high confidence, respectively. For *Nt-JAT2* and *NtNUP1* genes, we found orthologous genes in *N. noctiflora*, *N. cordifolia*, *N. knightiana*, *N. setchellii* and *N. tomentosiformis*. According to previous reports, *Nt-JAT2* is specifically expressed in leaves. *Nt-JAT2* contributes to the transportation of nicotine into the vacuole of leaves [49]. *Nt-NUP1* is a plasma membrane-localized nicotine transporter of the PUP family. It is involved in the movement of apoplastic nicotine into the cytoplasm of tobacco root cells, which affects nicotine metabolism and root growth. *NUP1* transcripts are less abundant in the leaves, but are abundant in root tips where nicotine is actively synthesized [48]. Two homologous MATE transporters, *NtMATE1* and *NtMATE2*, were reported to be responsible for the vacuolar accumulation of nicotine in the root. We found one orthologous gene corresponding to the tobacco *MATE1/2* genes in the leaf of *N. noctiflora*. This gene may have a different function.

We did not identify any orthologous genes for *Nt-JAT2* and *Nt-NUP1* in *N. glauca*. Since *N. glauca* can grow to a tree of several meters tall, it is possible that the acropetal transport of defensive alkaloid is relatively inefficient.

Table 4 Classification of NBS encoding genes based on the predicted domains from six wild *Nicotiana* transcriptomes

	N. glauca	*N. noctiflora*	*N. cordifolia*	*N. knightiana*	*N. setchellii*	*N. tomentosiformis*
CC-NBS	28	31	11	12	21	24
CC-NBS-LRR	3	7	2	2	7	5
TIR-NBS	4	4	8	3	5	4
TIR-NBS-LRR	9	11	0	4	2	6
NBS-LRR	14	8	13	5	8	0
NBS	113	112	84	61	84	79
Total	171	173	118	87	127	118

Table 5 Summary of ten transcription factors involved in plant defense in six wild *Nicotiana* transcriptomes

	N. glauca	*N. noctiflora*	*N. cordifolia*	*N. knightiana*	*N. setchellii*	*N. tomentosiformis*
MYB	200	221	217	180	225	152
WRKY	48	57	57	50	54	35
ERF-type/AP2-EREBP	78	87	95	67	103	79
CBF	16	15	22	18	20	16
bZIP	55	63	57	45	52	39
SBP/SPL6	20	23	24	20	20	20
NAC domain/NAM	55	65	69	45	58	45
TFIIA	1	2	2	2	2	1
Whirly	3	2	2	2	2	2
Homeo-domain	60	69	73	63	70	50
Total	536	604	618	492	606	439

Conclusions

It is well-known that genetic diversity is essential for the continuous genetic modification and improvement of cultivated crops, as well as for many basic studies in plant biology. As an economic crop, cultivated tobacco has received a fair number of desired genes from wild *Nicotiana* relatives [51–57]. Although wild *Nicotiana* species have played important roles in many research areas of plant biology, their genomic resources have been slowly developed relative to most other major crop species. With this study, we provide the reference transcriptome sequences of six wild *Nicotiana* species for public use. By constructing phylogenetic trees, we confirmed the classification of *Nicotiana* into three sections and the placement of these wild species in each section. Our study will provide a better understanding of the genomic architecture of wild *Nicotiana* and help elucidate genes involved in plant defense. It is likely that these *Nicotiana* species will be used as model systems for investigating many aspects of general plant biology in future. These sequences will be an important resource for evolutionary and developmental genetics in the genus *Nicotiana* and will contribute significantly to the improvement of cultivated tobacco and other important Solanaceae crops.

Methods

Plant materials and RNA extraction

Six wild species of *Nicotiana*, including *N. setchellii*, *N. cordifolia*, *N. knightiana*, *N. tomentosiformis*, *N. noctiflora*, and *N. glauca*, were grown in a greenhouse in Guizhou Province under the same cultivation conditions. Fresh leaves from 30-day-old flowerless plants were collected, snap frozen in liquid nitrogen, and stored at −70 °C. RNA was purified using TRIzol (Invitrogen, CA, USA) from the frozen materials according to the manufacturer's instructions. RNA degradation and contamination was monitored on 1 % agarose gels. RNA integrity was confirmed using the 2100 Bioanalyzer (Agilent Technologies) with a minimum RNA integrated number value of 8 after checking the RNA purity and concentration.

Library preparation and Illumina sequencing

RNA sequencing libraries were constructed in parallel from the six species using TruSeq RNA Sample Prep Kits (Illumina, SanDiego, USA). Briefly, first strand cDNA synthesis was performed with oligo-dT primer and Superscript II reverse transcriptase (Invitrogen, CA, USA). The second strand was synthesized with *Escherichia coli* DNA Pol I (Invitrogen, CA, USA). Double-stranded cDNA was purified with a Qiaquick PCR purification kit (Qiagen), and sheared with a nebulizer (Invitrogen, CA, USA) into 200–250 bp fragments. After the end repair and addition of a 3′-dA overhang, the cDNA was ligated to Illumina PE adapter oligo mix (Illumina). The products were then purified and enriched with PCR to create the final sequencing cDNA library. Both ends of the library were sequenced on the Illumina HiSeq 2000 platform.

De novo transcriptome assembly

Before performing the assembly, raw reads (FASTQ format) were cleaned by removing reads containing adaptor sequences, reads containing poly-N, and low-quality reads. For each species, de novo transcriptome assembly was performed using Trinity [26] (version: trinityrnaseq_r2012-06-08) with default settings except min_kmer_cov set to 2, which is a method for the efficient and robust de novo reconstruction of transcriptomes. Afterwards, transcripts with length <200 bp in each species were removed. The protein-coding region prediction program in the Trinity software suite (transcripts_to_best_scoring_ORFs.pl) was used to identify putative open reading frames (ORFs) consisting of at least 100 amino acids on the basis of nucleotide composition.

Functional annotation and expression level analysis

Peptide annotation for the ORFs obtained in this study was performed by BLASTp (version 2.2.29 +) [27] searching in the NCBI NR database (24th June 2014) and Swiss-Prot database (28th April 2015). An e-value cut-off of 10^{-5} was used and only one best hit was retained for each sequence query. To assign preliminary GO terms to the ORFs, InterProScan (version 5.4) [58] was used to screen the annotated peptide sequences against all the default databases. GO classification of the ORFs was conducted based on biological processes, molecular function, and cellular component and subsequently was visualized by the WEGO online tool [59]. Pathway assignments were carried out based on the KEGG database (8th March 2011) [60]. The ORFs from each species were first compared with KEGG database using BLASTp (version 2.2.29 +) with an e-value less than 10^{-5}. An in-house Perl script (https://github.com/NiLong/kegg_stat/) was developed to retrieve KO (KEGG Orthology) information from BLASTp results and correlation between peptides and database pathway was established. The RSEM software (version 1.2.13) [61] was used to quantify the expression level of the ORFs of six wild *Nicotiana* species measured as FPKM values.

Gene family and phylogenetic analysis

The initial set of annotation contained a high level of redundancy as more than half of the annotated transcripts were alternative splicing isoforms [62]. To avoid this redundancy in subsequent analyses, ORFs from six *Nicotiana* species were clustered using cd-hit-est command in CD-HIT v4.6.1-2012-08-27 [63] with >95 % similarity cutoff and only the representative ORFs in each cluster were retained. This yielded non-redundant sequence datasets for *N. glauca* (22,934 genes), *N. noctiflora* (26,788 genes), *N. cordifolia* (29,356 genes), *N. knightiana* (22,168 genes), *N. setchellii* (26,579 genes) and *N. tomentosiformis* (24,213 genes). These non-redundant sequences were defined as unigenes. Similarly, tomato coding sequences (CDSs) obtained from ITAG2.4 were also clustered and 33,721 genes were obtained. OrthoMCL v1.4 [64] was used to identify ortholog relationships between the six wild *Nicotiana* species and tomato.

Phylogenetic trees were constructed using sequences of 2491 single copy orthologs from six *Nicotiana* species with *S. lycopersicum* (tomato) as an outgroup. Multiple sequence alignments were performed by MUSCLE v3.8.31 [65]. Two methods were used to reconstruct phylogenetic trees: (1) the neighbor-joining method in Phylip v3.696 [29] and bootstrapped with 1000 replicates, and (2) the maximum likelihood method in PhyML v3.0 [30].

Identification of NBS containing genes and transcription factors related to disease resistance

For the identification of the NBS-encoding genes in this study, we followed the method described for diploid cotton *Gossypium raimondii* [66]. Firstly, the protein sequences of 112 manually curated reference disease resistance genes were collected from the plant resistance gene database (http://www.prgdb.org) [8]. Non-redundant protein sequences (unigenes) from the six wild *Nicotiana* species were subsequently checked for sequence homology with at least one resistance protein contained in the reference dataset using BLASTp (version 2.2.29+) (scores $\geq$100 and e-values $\leq 10^{-5}$). In a second step, all the BLAST hits were used for further analysis and were screened for protein domains by InterProScan version 5.4 [58]. In the third step, the genes with NBS domain were filtered out according to the NBS domain annotation (PF00931) given by the Pfam database (v27.0) [67]. Subsequently, the Pfam database, SMART protein motif analyses (Simple Modular Architecture Research Tool) (v6.2) [68], and the ncoils program (version 2.2) [69] were used to classify the NBS genes based on TIR, NBS, LRR and CC motifs. The program ncoils was used by InterProScan version 5.4 with default settings to predict coiled-coils domains. Pfam database (v27.0) and SMART protein motif analysis (v6.2) were used to detect the TIR (PF01582) and LRR domains (PF00560, PF07723, PF07725, PF12799, PF13306, PF13504, PF13516, PF13855, and PF14580).

Ten different TF families involved in disease resistance were retrieved from the literature by Sood et al. [70]. Then the unigenes were searched against the domains of ten TFs, including MYB (PF00249, PF13921, PF14379), WRKY (PF03106), ERF-type/AP2-EREBP (PF00847), CBF (PF02312, PF00808, PF03914), bZIP (PF00170, PF03131, PF07716, PF12498), SBP/SPL6 (PF03110), NAC domain/NAM (PF02365, PF14303), TFIIA (PF03153, PF02268, PF02751), Homeo-domain (PF00046, PF05920, PF00157, PF13384, PF13565) and Whirly (PF08536) using the Pfam database from InterProScan version 5.4, respectively. Finally, proteins matching with the Pfam IDs were selected as TFs associated with disease resistance.

Identification of alkaloid transporter genes

The coding sequences of *Nt-JAT1* (accession number AM991692), *Nt-JAT2* (accession number AB922128), *NtMATE1* (accession number AB286961), *NtMATE2* (accession number AB286962) and *NtNUP1* (accession number GU174267) of *N. tabacum* were retrieved from NCBI. To identify orthologs of the alkaloid transporter genes in each species, we first performed a bidirectional BLAST search of the alkaloid transporter genes and

unigenes from the six species against each other (identity ≥80 %, e-values $\leq 10^{-5}$, query coverage ≥50 %).

Availability of supporting data

RNA-seq data have been deposited in the NCBI sequence Read Archive under the accession numbers SRR2106216, SRR2106514, SRR2106516, SRR2106517, SRR2106530 and SRR2106531. The data set supporting the results of this article is included within the additional files.

Additional files

Additional file 1. Characteristics of raw data and assembled transcripts in six wild *Nicotiana* species.

Additional file 2. Length distribution of the ORFs predicted from the transcripts of six wild *Nicotiana* species. Length values are represented in amino acids.

Additional file 3. E-value distribution of the top BLAST hits for each ORF in the Swiss-Prot database. The cutoff e-value was set to 10^{-5}.

Additional file 4. Comparison of ORF length between hit and no hit proteins in Swiss-Prot database. For *N. glauca, N. noctiflora, N. cordifolia, N. knightiana, N. setchellii, N. tomentosiformis*, longer ORFs were more likely to have BLASTp homologs in protein database.

Additional file 5. The expression levels of ORFs (FPKM) in the six *Nicotiana* species.

Additional file 6. The top 20 abundant ORFs in the six *Nicotiana* species.

Additional file 7. Phylogenetic tree based on the transcriptomes of six *Nicotiana* and *S. lycopersicum* using ML method.

Authors' contributions

NL analyzed the data and drafted the manuscript. XLR provided the plants and revised the manuscript. WTW performed the experiment, prepared the mRNA and performed sequencing. ZDX performed data preprocess. YD designed the project and revised the manuscript. All authors read and approved the final manuscript.

Author details

[1] Faculty of Life Science and Technology, Kunming University of Science and Technology, South Jingming Road No.727, Kunming 650500, Yunnan, China. [2] Guizhou Tobacco Research Institute, North Yuntan Road, Jinyang District, Guiyang 550003, Guizhou, China. [3] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 32 East Jiaochang Road, Kunming 650223, Yunnan, China. [4] Biological Big Data College, Yunnan Agricultural University, Kunming 650201, Yunnan, China.

Acknowledgements

We thank Guizhou Tobacco Research Institute for their assistance in providing wild *Nicotiana* materials.

Competing interests

The authors declare that they have no competing interests.

Received: 9 November 2015 Accepted: 5 April 2016
Published online: 18 April 2016

References

1. Goodspeed TH. Cytotaxonomy of *Nicotiana*. Bot Rev. 1945;11:533–92.
2. Goodspeed TH, Thompson MC. Cytotaxonomy of *Nicotiana*. II. Bot Rev. 1959;25:385–415.
3. Knapp S, Chase MW, Clarkson JJ. Nomenclatural changes and a new sectional classification in *Nicotiana* (Solanaceae). Taxon. 2004;53:73–82.
4. Lewis RS. *Nicotiana*. In: Kole C, editor. Wild crop relatives: genomic and breeding resources, plantation and ornamental crops. Berlin: Springer; 2011. p. 185–208.
5. Battey JN, Sierro N, Bakaher N, Ivanov NV. Advances in *Nicotiana* genetic and "omics" resources. In: Tuberosa R, Graner A, Frison E, editors. Genomics of plant genetic resources. Dordrecht: Springer; 2014. p. 511–32.
6. Lewis R, Linger L, Wolff M, Wernsman E. The negative influence of N-mediated TMV resistance on yield in tobacco: linkage drag versus pleiotropy. Theor Appl Genet. 2007;115:169–78.
7. Whitham S, Dinesh-Kumar SP, Choi D, Hehl R, Corr C, Baker B. The product of the tobacco mosaic virus resistance gene N: similarity to toll and the interleukin-1 receptor. Cell. 1994;78:1101–15.
8. Sanseverino W, Hermoso A, D'Alessandro R, Vlasova A, Andolfo G, Frusciante L, et al. PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants. Nucleic Acids Res. 2013;41:D1167–71.
9. Michelmore RW, Meyers BC. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. 1998;8:1113–30.
10. Bai J, Pennill LA, Ning J, Lee SW, Ramalingam J, Webb CA, et al. Diversity in nucleotide binding site-leucine-rich repeat genes in cereals. Genome Res. 2002;12:1871–84.
11. Cannon SB, Zhu H, Baumgarten AM, Spangler R, May G, Cook DR, et al. Diversity, distribution, and ancient taxonomic relationships within the TIR and non-TIR NBS-LRR resistance gene subfamilies. J Mol Evol. 2002;54:548–62.
12. Alvarenga SM, Caixeta ET, Hufnagel B, Thiebaut F, Maciel-Zambolim E, Zambolim L, et al. In silico identification of coffee genome expressed sequences potentially associated with resistance to diseases. Genet Mol Biol. 2010;33:795–806.
13. Liu Z, Crampton M, Todd A, Kalavacharla V. Identification of expressed resistance gene-like sequences by data mining in 454-derived transcriptomic sequences of common bean (*Phaseolus vulgaris* L.). BMC Plant Biol. 2012;12:42.
14. Joshi RK, Kar B, Nayak S. Survey and characterization of NBS-LRR (R) genes in *Curcuma longa* transcriptome. Bioinformation. 2011;6:360–3.
15. Rajesh MK, Rachana KE, Naganeeswaran SA, Shafeeq R, Thomas RJ, Shareefa M, et al. Identification of expressed resistance gene analog sequences in coconut leaf transcriptome and their evolutionary analysis. Turk J Agric For. 2015;39:489–502.
16. Burk L, Heggestad H. The genus *Nicotiana*: a source of resistance to diseases of cultivated tobacco. Econ Bot. 1966;20:76–88.
17. Durbin RD. *Nicotiana*: procedures for experimental use. Washington: Technical Bulletins; 1979.
18. Trojak-Goluch A, Berbeć A. Potential of *Nicotiana glauca* (Grah.) as a source of resistance to black root rot *Thielaviopsis basicola* (Berk. and Broome) Ferr. in tobacco improvement. Plant Breed. 2005;124:507–10.
19. Baldwin IT. Mechanism of damage-induced alkaloid production in wild tobacco. J Chem Ecol. 1989;15:1661–80.
20. Steppuhn A, Gase K, Krock B, Halitschke R, Baldwin IT. Nicotine's defensive function in nature. PLoS Biol. 2004;2:E217.
21. Bombarely A, Rosli HG, Vrebalov J, Moffett P, Mueller LA, Martin GB. A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. Mol Plant Microbe Interact. 2012;25:1523–30.
22. Nakasugi K, Crowhurst RN, Bally J, Wood CC, Hellens RP, Waterhouse PM. De novo transcriptome sequence assembly and analysis of RNA silencing genes of *Nicotiana benthamiana*. PLoS ONE. 2013;8:e59534.
23. Sierro N, Battey JN, Ouadi S, Bovet L, Goepfert S, Bakaher N, et al. Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. Genome Biol. 2013;14:R60.
24. Nakasugi K, Crowhurst R, Bally J, Waterhouse P. Combining transcriptome assemblies from multiple *de novo* assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. PLoS ONE. 2014;9:e91776.
25. Sierro N, Battey JND, Ouadi S, Bakaher N, Bovet L, Willig A, et al. The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun. 2014; 5.
26. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.

27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
28. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.
29. Felsenstein J. PHYLIP-phylogeny inference package (version 3.2). Cladistics. 1989;5:164–6.
30. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21.
31. Goodspeed TH. The genus *Nicatiana*: origins, relationships and evolution of its species in light of their distribution, morphology and cytogenetics. Waltham: Chronica Botanica Co; 1954.
32. Chase MW, Knapp S, Cox AV, Clarkson JJ, Butsko Y, Joseph J, et al. Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). Ann Bot. 2003;92:107–27.
33. Wenping H, Yuan Z, Jie S, Lijun Z, Zhezhi W. De novo transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. Genomics. 2011;98:272–9.
34. La Camera S, Gouzerh G, Dhondt S, Hoffmann L, Fritig B, Legrand M, et al. Metabolic reprogramming in plant innate immunity: the contributions of phenylpropanoid and oxylipin pathways. Immunol Rev. 2004;198:267–84.
35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
36. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell. 2003;15:809–34.
37. Raju KS, Sheshumadhav M, Murthy T. Molecular diversity in the genus *Nicotiana* as revealed by randomly amplified polymorphic DNA. Physiol Mol Biol Plants. 2008;14:377–82.
38. Tai TH, Dahlbeck D, Clark ET, Gajiwala P, Pasion R, Whalen MC, et al. Expression of the Bs2 pepper gene confers resistance to bacterial spot disease in tomato. Proc Natl Acad Sci USA. 1999;96:14153–8.
39. Liu Y, Schiff M, Marathe R, Dinesh-Kumar SP. Tobacco Rar1, EDS1 and NPR1/NIM1 like genes are required for N-mediated resistance to tobacco mosaic virus. Plant J. 2002;30:415–29.
40. Katiyar A, Smita S, Lenka SK, Rajwanshi R, Chinnusamy V, Bansal KC. Genome-wide classification and expression analysis of MYB transcription factor families in rice and *Arabidopsis*. BMC Genom. 2012;13:544.
41. Pandey SP, Somssich IE. The role of WRKY transcription factors in plant immunity. Plant Physiol. 2009;150:1648–55.
42. Alves MS, Dadalto SP, Goncalves AB, De Souza GB, Barros VA, Fietto LG. Plant bZIP transcription factors responsive to pathogens: a review. Int J Mol Sci. 2013;14:7815–28.
43. Desveaux D, Subramaniam R, Després C, Mess J-N, Lévesque C, Fobert PR, et al. A "Whirly" transcription factor is required for salicylic acid-dependent disease resistance in *Arabidopsis*. Dev Cell. 2004;6:229–40.
44. Shin R, Han J-H, Lee G-J, Peak K-H. The potential use of a viral coat protein gene as a transgene screening marker and multiple virus resistance of pepper plants coexpressing coat proteins of cucumber mosaic virus and tomato mosaic virus. Transgenic Res. 2002;11:215–9.
45. Shitan N, Kato K, Shoji T. Alkaloid transporters in plants. Plant Biotechnology. 2014;31:453–63.
46. Morita M, Shitan N, Sawada K, Van Montagu MC, Inzé D, Rischer H, et al. Vacuolar transport of nicotine is mediated by a multidrug and toxic compound extrusion (MATE) transporter in *Nicotiana tabacum*. Proc Natl Acad Sci USA. 2009;106:2447–52.
47. Shoji T, Inai K, Yazaki Y, Sato Y, Takase H, Shitan N, et al. Multidrug and toxic compound extrusion-type transporters implicated in vacuolar sequestration of nicotine in tobacco roots. Plant Physiol. 2009;149:708–18.
48. Hildreth SB, Gehman EA, Yang H, Lu RH, Ritesh KC, Harich KC, et al. Tobacco nicotine uptake permease (NUP1) affects alkaloid metabolism. Proc Natl Acad Sci USA. 2011;108:18179–84.
49. Shitan N, Minami S, Morita M, Hayashida M, Ito S, Takanashi K, et al. Involvement of the leaf-specific multidrug and toxic compound extrusion (MATE) transporter Nt-JAT2 in vacuolar sequestration of nicotine in *Nicotiana tabacum*. PLoS ONE. 2014;9:e108789.
50. Kato K, Shitan N, Shoji T, Hashimoto T. Tobacco NUP1 transports both tobacco alkaloids and vitamin B6. Phytochemistry. 2015;113:33–40.
51. Yi Y, Rufty R. RAPD markers elucidate the origin of the root-knot nematode resistance gene (Rk) in tobacco. Tob Sci. 1998;42:58–63.
52. Johnson E, Wolff M, Wernsman E, Atchley W, Shew H. Origin of the black shank resistance gene, Ph, in tobacco cultivar Coker 371-Gold. Plant Dis. 2002;86:1080–4.
53. Johnson E, Wolff M, Wernsman E, Rufty R. Marker-assisted selection for resistance to black shank disease in tobacco. Plant Dis. 2002;86:1303–9.
54. Lewis RS. Transfer of resistance to potato virus Y (PVY) from *Nicotiana africana* to *Nicotiana tabacum*: possible influence of tissue culture on the rate of introgression. Theor Appl Genet. 2005;110:678–87.
55. Milla S, Lewin J, Lewis R, Rufty R. RAPD and SCAR markers linked to an introgressed gene conditioning resistance to *Peronospora tabacina* DB Adam in tobacco. Crop science. 2005;45:2346–54.
56. Lewis RS, Milla SR, Kernodle SP. Analysis of an introgressed *Nicotiana tomentosa* genomic region affecting leaf number and correlated traits in *Nicotiana tabacum*. Theor Appl Genet. 2007;114:841–54.
57. Moon H, Nicholson J. AFLP and SCAR markers linked to resistance in tobacco. Crop Sci. 2007;47:1887–94.
58. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
59. Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, et al. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 2006;34:W293–7.
60. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277–80.
61. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:1.
62. Zhou X, Rinker DC, Pitts RJ, Rokas A, Zwiebel LJ. Divergent and conserved elements comprise the chemoreceptive repertoire of the nonblood-feeding mosquito *Toxorhynchites amboinensis*. Genome Biol Evol. 2014;6:2883–96.
63. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
64. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
65. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
66. Wei H, Li W, Sun X, Zhu S, Zhu J. Systematic analysis and comparison of nucleotide-binding site disease resistance genes in a diploid cotton *Gossypium raimondii*. PLoS ONE. 2013;8:e68435.
67. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997;28:405–20.
68. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA. 1998;95:5857–64.
69. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science. 1991;252:1162–4.
70. Sood A, Jaiswal V, Chanumolu SK, Malhotra N, Pal T, Chauhan RS. Mining whole genomes and transcriptomes of Jatropha (*Jatropha curcas*) and Castor bean (*Ricinus communis*) for NBS-LRR genes and defense response associated transcription factors. Mol Biol Rep. 2014;41:7683–95.