Planta (2025) 261:138
https://doi.org/10.1007/s00425-025-04715-z

ORIGINAL ARTICLE

Insight into the genetic underpinnings of tobacco hairy root formation by variant-associated genes based on whole-genome resequencing

Xiaozong Wu[1] · Zhiwen Zhu[1] · Peilin Li[1] · Zhitao Qi[1] · Ruojie Zhu[2] · Chaonan Shi[1]

Received: 4 March 2025 / Accepted: 5 May 2025
© The Author(s) 2025

Abstract

Main conclusion **Our whole-genome resequencing of tobacco hairy roots reveals functionally relevant variations in secondary metabolism-related genes and NAC transcription factors, providing actionable targets for metabolic pathway optimization and bioreactor design.**

Abstract Tobacco hairy roots are a critical model system for studying plant root development and secondary metabolism. The in-depth analysis of their genetic background and molecular regulatory mechanisms is important for biotechnological applications. In this study, we performed a whole-genome resequencing of tobacco hairy roots to uncover their genomic variation characteristics and potential functional implications. Genes associated with stop-lost, stop-gained, start-lost, and premature-start-codon-gain variants were enriched in zeatin biosynthesis, flavonoid biosynthesis, and glycosyltransferase activities. The results of metabolite content determination showed that hairy roots possessed a low content of zeatin and flavonoid but a higher content of glycoside compounds. Among transcription factors associated with effective variants, NAC transcription factors constituted the largest proportion. Further characterization of NAC proteins revealed their functional domains and expression patterns. This study not only explores the molecular genetic underpinnings of tobacco hairy roots but also provides a critical dataset for metabolic engineering optimization, development of efficient bioreactors, and plant–microbe interaction mechanisms research.

Keywords Hairy roots · Insertion-deletion (InDel) · NAC family · *Nicotiana tabacum* · Single nucleotide polymorphism (SNP) · Whole-genome resequencing (WGRS)

Abbreviations

GO	Gene ontology
InDel	Insertion-deletion
KEGG	Kyoto encyclopedia of genes and genomes
SNP	Single nucleotide polymorphism
WGRS	Whole-genome resequencing

Communicated by Dorothea Bartels.

Xiaozong Wu and Zhiwen Zhu contributed equally to the article.

✉ Chaonan Shi
2023055@zzuli.edu.cn

Xiaozong Wu
wuxzong@126.com

Zhiwen Zhu
wenfzhu@yeah.net

Peilin Li
332417051004@zzuli.edu.cn

Zhitao Qi
332417051007@zzuli.edu.cn

Ruojie Zhu
ruojiez@163.com

[1] Key Laboratory of Biotechnology in Tobacco Industry, College of Tobacco Science and Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China

[2] Shiyan Branch of Hubei Tobacco Company, Shiyan 442000, China

Published online: 13 May 2025

Introduction

Hairy roots are highly branched adventitious roots that develop from callus after plant wounds are infected by *Agrobacterium rhizogenes*. Compared to normal roots, hairy roots exhibit significant morphologic differences and, due to a series of advantageous biologic characteristics, they have attracted considerable interest in plant biotechnology research (Zhu et al. 2024). Owing to their superior metabolic synthesis capabilities, plant hairy roots are used as bioreactors to produce functional metabolites, such as antibiotics, alkaloids, and pigments (Ji et al. 2023). Moreover, hairy roots exhibit rapid growth rates and short production cycles, which are conducive to large-scale industrial production and application. Owing to their hormone autotrophic nature, hairy roots possess autonomous hormone biosynthesis capability, and their genetic stability makes them well-suited for functional gene studies (Gerszberg and Wiktorek-smagur 2022; Aghaali and Naghavi 2023; Lin et al. 2023). Hairy root cultures are not only effective tools for producing plant secondary metabolites, recombinant proteins, and industrial products, but also serve as valuable resources for plant genetic engineering and environmental remediation.

Tobacco is one of the most important economic crops and is capable of synthesizing numerous valuable secondary metabolites and proteins. For example, a previous study identified more than 300 acyltransferases expressed in tobacco trichomes using high-throughput genetic screening and transcriptomic analysis (Schenck et al. 2022). As a model plant, tobacco has a high-quality, publicly available genome, along with extensive transcriptomic and metabolomic resources. It is also amenable to genetic transformation. Genetically modified tobacco has been widely used as a chassis organism in synthetic biology to produce pharmaceuticals, vaccines, hormones, and other bioactive compounds (Molina-hidalgo et al. 2021). In a recent study, the tobacco hairy root culture system was optimized by transforming regulatory genes involved in nicotine synthesis, establishing a high-throughput gene function identification platform (Qin et al. 2022). Hou et al. (2017) discovered that co-expression bHLH and MYB transcription factors in the tobacco hairy root system induced anthocyanin biosynthesis in both *Nicotiana tabacum* L. and *Ipomea tricolor*. These research findings highlight the potential of tobacco hairy roots as bioreactors for secondary metabolite production and as a genetic platform for investigating biosynthetic pathways.

Whole genome resequencing (WGRS) is a high-throughput sequencing technology widely used to identify genomic structural variations (SVs) among individuals. Compared to exon-capture or targeted sequencing approaches, WGRS provides comprehensive genomic coverage, enabling the detection of diverse genetic variants. These variations are typically categorized by size into short- and long-fragment variations, including single nucleotide polymorphisms (SNPs), small insertions and deletions (InDels), and structural variants (SVs) (Vashisht et al. 2024). Wang et al. (2024) employed WGRS to detect mutation sites in a *BBla*-overexpressing male sterile tobacco mutant. Through functional annotation of these variants, they identified 174 candidate genes potentially associated with male reproductive organs development. WGRS has also proven valuable in animal studies, elucidating the genetics basis of growth trait differences between Hu and Gamba sheep (Yang et al. 2024) and revealing population-specific genomic variations in pigs from China and Denmark (Wu et al. 2024).

While previous research on hairy roots has primarily focused on their biotechnological applications, studies investigating their genetic underpinnings remain limited. In this study, we performed WGRS of tobacco hairy roots and conducted a comparative genomic analysis to identify variation-enriched metabolic pathways.

Materials and methods

Experimental materials

Tobacco hairy roots were obtained by infecting *Nicotiana tabacum* L. K326 plants with *Agrobacterium rhizogenes* strain A4. *Agrobacterium rhizogenes* strain A4 and K326 seeds were stored in the laboratory of Zhengzhou University of Light Industry (Zhengzhou, China). The reference genome of regular K326 tobacco was obtained from: https://solgenomics.net/ftp/genomes/Nicotiana_tabacum/edwards_et_al_2017.

DNA was extracted from the tissue using the CTAB (cetyltrimethylammonium bromide) method. Two methods were used to determine DNA degradation and contamination: 1% agarose gel and ND-2000 (Shanghai, China) detection. Only high-quality DNA samples (OD260/280 = 1.8-2.0, OD260/230 ≥ 2.0) were used to construct the sequencing library.

WGRS of samples

The tobacco hairy root samples were submitted to Shanghai Majorbio Bio-Pharm Technology Co., Ltd. (Shanghai, China) for WGRS analysis. After the sample's genomic DNA passed the quality control, ultrasonic waves were used to break the DNA sequence into random fragments. These fragmented DNA segments were then subjected to end-repair, A-tailing at the 3' end, and ligation of sequencing

adapters. Magnetic beads were employed to enrich fragments with a length of approximately 350 bp, which were subsequently amplified by PCR to form a sequencing library. The prepared library underwent quality control. Only the high-quality libraries were sequenced using the Illumina NovaSeq X Plus™ platform with an Illumina PE150 sequencing strategy, yielding reads with a length of ~300 bp.

Alignment with the reference genome

After sequencing data (raw data) were downloaded from the Illumina NovaSeq X Plus™ platform, quality control was conducted to filter out low-quality sequences, resulting in high-quality data (clean data). The clean data were aligned to the reference genome sequence using BWA-MEME software (Youngmok and Dongsu 2022), and the positional assignment of the sequences was obtained (i.e., generating a BAM file). The BAM file was then processed through the Best Practices pipeline of GATK software (McKenna et al. 2010) for correction and detection of SNPs or InDels.

Variant detection and annotation

The alignment results were processed using the Best Practices pipeline of GATK, and the Haplotyper method of GATK was applied to detect SNPs/InDels. The filtering criteria adhered to the parameters recommended by GATK, which are detailed at https://software.broadinstitute.org/gatk/documentation/article.php?id=3225. Using the SnpEff program (Cingolani et al. 2012) and the tobacco genome annotation, we performed functional annotation of the detected SNPs/InDels. The default settings of SnpEff were used to examine the distribution of genes and functional regions within the genome, compile statistics on the location and function of each SNP and InDel, and summarize the functions of each variant type. In addition, variants with a large impact on gene function, such as missense variants and stop-retained variants, were considered effective variants for further in-depth analysis.

GO, KEGG and transcription factor families analysis

The Gene Ontology (GO) project aims to standardize the attributes of genes and gene products across species and databases (http://www.geneontology.org/). The functional identification of annotated single gene sequences was performed using the GO method. Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/) associates the activity of a single gene sequence with a biologic pathway. To further analyze the functions of genes involved in the effective variants, GO and KEGG enrichment analyses were performed. By conducting GO annotation and enrichment analysis on genes associated with effective SNPs and InDels variations, we selected GO terms at level 2 and counted the number of genes enriched in each term. We performed KEGG pathway enrichment analysis on genes related to effective SNPs and InDels (with q-value < 1). Enrichment analysis of transcription factor gene sets was also performed. To identify significantly enriched transcription factor families, enrichment analysis was performed on all associated transcription factors. The screening criteria were set as enrichment factor > 1 and p-value < 0.05.

Identification of metabolites

In this study, we selected samples of whole K326 plants (a mixed pool of five individual plants, including roots, stems, and leaves) and hairy roots (a mixed pool of hairy roots induced from five explants) for metabolite content determination, with three biologic replicates for each sample. Metabolite detection was conducted using the ultra-high performance liquid chromatography coupled with Fourier transform mass spectrometry (UHPLC-Exploris240) system. The metabolomics software Progenesis QI v3.0 (https://www.nonlinear.com/) was employed to analyze the off-machine data for peak extraction, alignment, identification, and other processes, ultimately obtaining a data matrix of identified metabolite information for subsequent differential comparative analysis. Chromatographic conditions: the column used was ACQUITY UPLC HSS T3 (100 mm × 2.1 mm i.d., 1.8 μm); mobile phase A consisted of 95% water + 5% acetonitrile (containing 0.1% formic acid), and mobile phase B consisted of 47.5% acetonitrile + 47.5% isopropanol + 5% water (containing 0.1% formic acid). The injection volume was 3 μL, and the column temperature was maintained at 40°C.

Bioinformatics approaches for gene family studies

The chromosomal locations of 56 NtNAC genes were mapped using tobacco genome annotations and visualized with TBtools gene location visualization tool (Chen et al. 2023). Gene structure information (exon-intron organization) of the NAC family members was extracted directly from the tobacco genome annotation file. Protein conserved motifs of NtNAC sequences were predicted via the MEME suite (http://meme-suite.org) (Bailey and Elkan 1994), with the motif number set to 10. The results were visualized using TBtools.

A phylogenetic tree of tobacco NAC proteins was constructed with MAGE11 using the neighbor-joining (NJ) method (Tamura et al. 2021). The tree topology was subsequently visualized on the iTOL platform (https://itol.embl.de/) (Letunic and Bork 2021). Gene duplication events among NAC family members were analyzed and visualized through TBtools. Protein–protein interaction (PPI) networks of the NAC family members were predicted

using the STRING database (https://string-db.org/) and visualized with Cytoscape (v3.10.3) (Shannon et al. 2003; Szklarczyk et al. 2023). GO enrichment analysis of NAC family members was performed using TBtools, and the results were graphically represented via the Bioinformatics.com.cn platform (https://www.bioinformatics.com.cn/) (Chen et al. 2023; Tang et al. 2023).

Results

WGRS data statistics and analysis results

The WGRS of *Agrobacterium rhizogenes* strain A4 induced hairy roots from *Nicotiana tabacum* cv. K326 generated 171,494,470 high-quality clean reads. The mapped reads accounted for 79.66% of the total clean reads. The sequencing depth was greater than 20×, with a coverage of 91.16%. The proportion of bases with quality values greater than or equal to 30 in the raw sequencing data were 95.32% (Table 1). The distribution of the insert fragment sizes in the samples conformed to a normal distribution with a central value of around 350 bp, indicating no abnormalities in the construction of the sequencing data library (Fig. 1a). Both the coverage and sequencing depth reflected high uniformity of the sequencing data and homology to the reference sequence (Fig. 1b).The genome is covered relatively evenly, indicating good sequencing randomness. Uneven depth on the plot may be caused by repetitive sequences, PCR bias, or centromeric regions (Fig. 1c).

Table 1 Quality control results of WGRS data and genome alignment results for tobacco hairy roots

Index	Value
Raw_Reads	173,527,976
Clean_Reads	171,494,470
Raw_Bases (bp)	2,620,2724,376 (26.20 G)
Clean_Bases (bp)	25,765,631,649 (25.77 G)
Clean_Bases_Percent (%)	98.33
Clean_GC (%)	45.02
Clean_Q30 (%)	95.32
Mapped_Ratio (%)	79.66
Properly_Mapped (%)	63.9
Real_Depth	6.13
Insert_Size	300
Coverage (%) (≥ 1×)	91.16
Coverage (%) (≥ 4×)	60.08

Variations exhibit a biased distribution in the genome

To further elucidate the genetic characteristics of tobacco hairy roots, a comprehensive analysis was conducted on detected SNPs and InDels across the 24 chromosomes. A total of 3,059,775 SNPs and 242,577 InDels were identified. Among these variants, 2,950,939 were located in intergenic regions, while 351,413 were found within coding regions (Table 2). These results reveal a rich spectrum of genetic variation and a distinct bias in genomic distribution.

We further annotated and analyzed the variants in non-intergenic regions, identifying a total of 23,923 genes. Notably, chromosome 17 contained the highest number of genes (2267), whereas chromosome 11 had the fewest (414) (Fig. 2). This distribution highlights the uneven genetic complexity across different chromosomes, providing valuable insights into the genomic architecture and potential regulatory mechanisms underlying tobacco hairy root development.

Genome-wide SNPs/InDels identification and functional annotation

Following genome alignment and screening, a total of 3,059,775 SNPs were identified across the chromosomes of tobacco hairy roots, with SNP density ranging from the highest on chromosome 17 to the lowest on chromosome 21. Genome-wide SNP mapping revealed a dense distribution across all 24 chromosomes (Fig. S1). After annotating and filtering out intergenic SNPs, intronic variants predominated (55.14%) among the effective SNPs (in this study, non-synonymous variants occurring in the coding regions of genes). Functional annotation classified 33,498 effective SNPs into distinct categories: missense variants constituted the largest category (28,205 SNPs; 84.2%), followed by termination-altering variants (2311 SNPs; 6.9%), highlighting their potential impacts on protein structure and transcriptional regulation (Fig. 3a).

Compared with the reference genome, the results showed that a total of 242,577 InDels were identified in the tobacco hairy roots (Table 2). Consistent with the distribution pattern of SNP variations, chromosome 17 exhibited the highest frequency of InDels (22,052), whereas chromosome 21 showed the lowest count (6605). Notably, chromosome 17 demonstrated the highest density for both SNP and InDel variations (Figs. S1, S2). Further analysis of the InDel distribution revealed that 31,808 variations (54.12% of non-intergenic InDels) were located in intronic regions. Among these, 2485 (4.23%) were identified as frameshift variants, while 869 were classified as non-frameshift variants (Fig. 3b). This distribution pattern profoundly reveals the genomic variation characteristics of tobacco hairy roots.

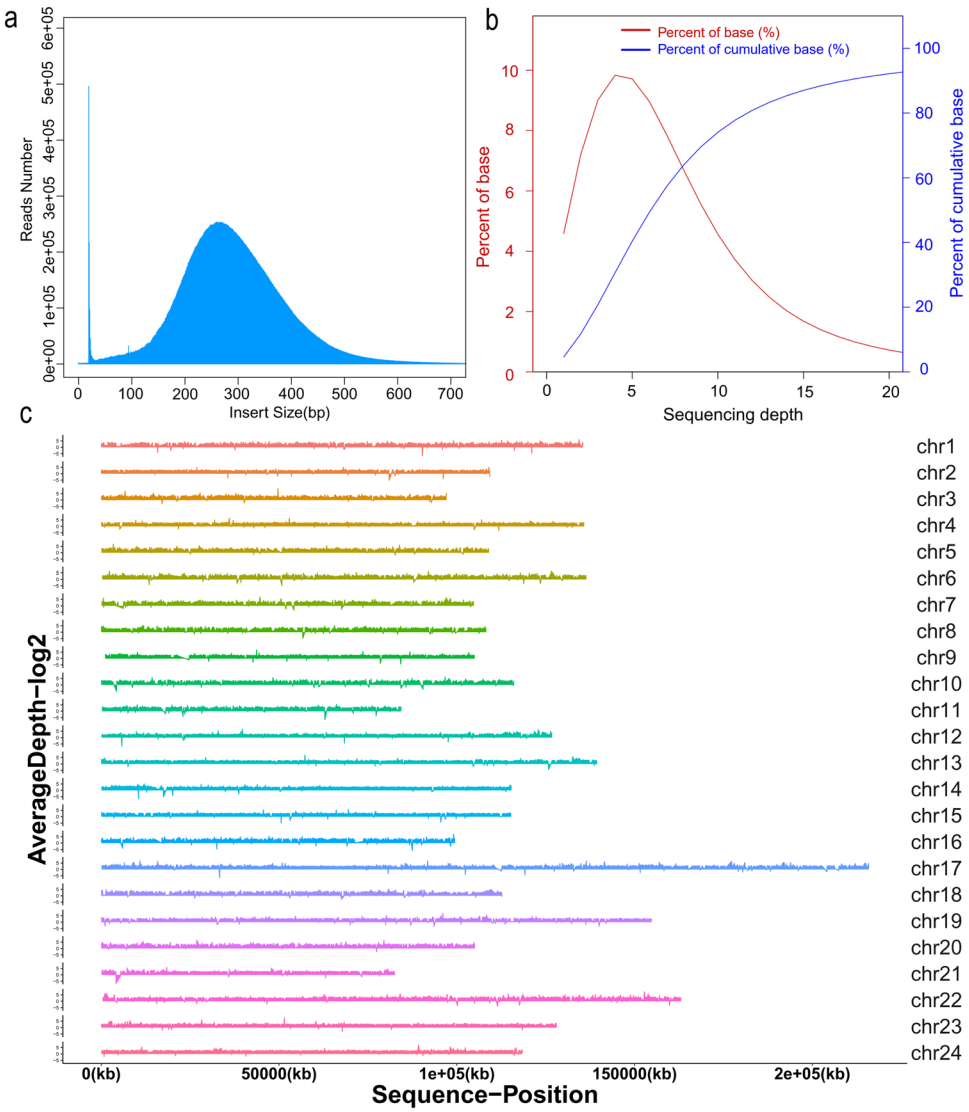

Fig. 1 Summary of WGRS results. **a** Insertion fragment length distribution plot of the sample. **b** Depth distribution plot of the sample. **c** Chromosome coverage depth distribution plot of the sample

Functional enrichment analysis of variant-associated genes

To gain functional insights, we conducted GO and KEGG enrichment analyses on genes associated with effective SNPs and InDels. The GO functional analysis of variant genes revealed significant enrichment in molecular function (MF), cellular component (CC), and biological process (BP) categories. Specifically, variant genes associated with SNPs/InDels were predominantly enriched in binding and catalytic activity (MF), metabolic/cellular/single-organism processes (BP), and membrane/cell/cell part components (CC) (Fig. 4a, b). Overall, the GO enrichment analysis demonstrated that the majority of variant-related genes for both SNPs and InDels were concentrated in the MF and BP categories. This suggests that genomic alterations induced by *Agrobacterium rhizogenes* infection in tobacco hairy roots primarily affect biologic processes and molecular functions, with comparatively fewer effects on cellular components.

KEGG pathway enrichment analysis was performed on genes associated with effective SNPs and InDels, and the top 20 most enriched pathways were visualized in a bubble chart (Fig. 5a, b). Secondary metabolite biosynthesis was the most significantly enriched pathway for both SNP-related and InDel-related genes, while the notch signaling pathway, sesquiterpenoid/triterpenoid biosynthesis, and histidine metabolism were notably enriched for SNPs, and the protein export pathway emerged as the top pathway for InDels (Fig. 5a, b). The KEGG enrichment analysis highlighted that the secondary metabolite biosynthesis pathway was the most prominently modulated pathway following *Agrobacterium rhizogenes* infection, underscoring its potential role in response to the infection.

Transcription factors, as pivotal regulators of gene expression, play an essential role in plant growth and

Table 2 Statistics of variations in each chromosome

Chromosome	Number of total variations	Variation type		Variation feature	
		SNPs	InDels	Intergenic regions	Transcripts
Chr1	153,274	143,557	9717	142,356	10,918
Chr2	112,405	103,266	9139	97,500	14,905
Chr3	110,625	103,038	7587	102,861	7764
Chr4	140,458	128,875	11,583	118,439	22,019
Chr5	131,062	122,627	8435	122,604	8458
Chr6	170,204	159,005	11,199	160,259	9945
Chr7	122,485	113,977	8508	112,906	9579
Chr8	135,660	126,791	8869	127,199	8461
Chr9	97,948	90,406	7542	85,357	12,591
Chr10	133,983	124,579	9404	125,401	8582
Chr11	104,613	97,656	6957	99,209	5404
Chr12	127,458	117,386	10,072	109,335	18,123
Chr13	155,471	143,070	12,401	134,937	20,534
Chr14	120,025	111,027	8998	104,437	15,588
Chr15	122,935	113,122	9813	102,482	20,453
Chr16	110,588	103,262	7326	103,292	7296
Chr17	289,324	267,272	22,052	247,342	41,982
Chr18	118,789	110,167	8622	108,505	10,284
Chr19	166,408	152,059	14,349	139,119	27,289
Chr20	126,507	118,112	8395	117,719	8788
Chr21	91,856	85,251	6605	82,928	8928
Chr22	194,742	181,061	13,681	177,101	17,641
Chr23	138,022	127,112	10,910	119,838	18,184
Chr24	127,510	117,097	10,413	109,813	17,697
Total	3,302,352	3,059,775	242,577	2,950,939	351,413

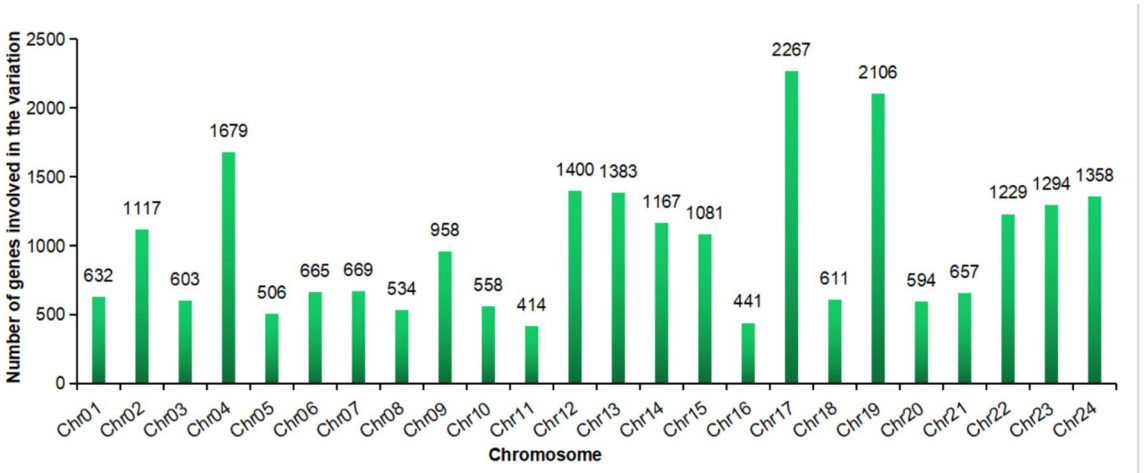

Fig. 2 Chromosome distribution of genes harboring variations

development. We performed an enrichment analysis on transcription factor gene sets to identify potential regulatory roles of these genes. A total of 776 genes possessed effective variations were associated with 73 transcription factor families, including NAC (56 genes), MYB (62 genes), and MADS (44 genes). Among these families, 41 transcription factor families exhibited significant enrichment. Notably, nine families (NAC, MADS, C2H2, FAR1, C3H, WRKY, bZIP, ABI3VP1, and SET) contained more than 20 genes each (Fig. 6). These findings reveal the potential regulatory

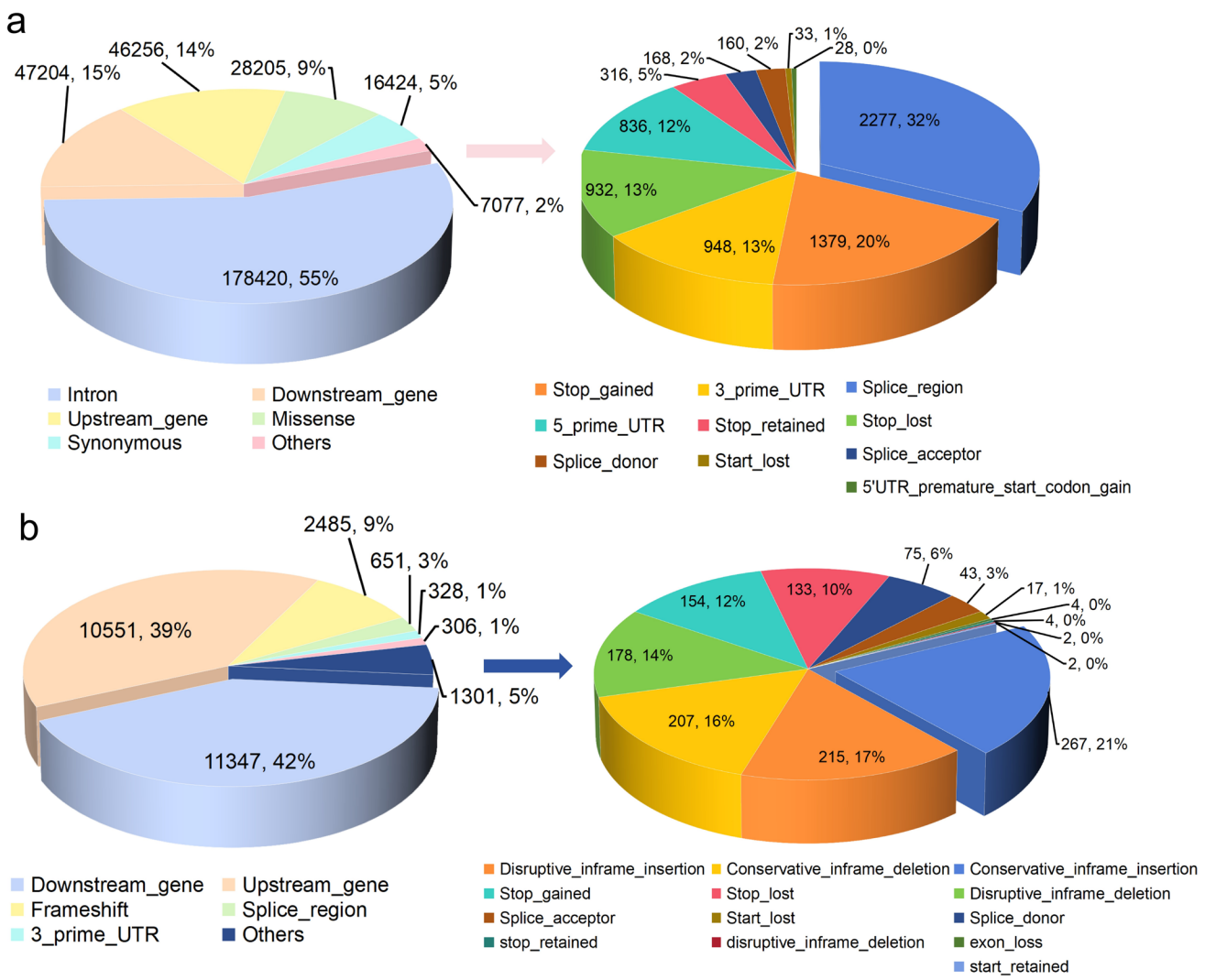

Fig. 3 **a** Proportion of SNP variations of different functional domains. **b** Proportion of InDels with different functional domains

significance of these transcription factor families in response to genomic variations.

Functional analysis of genes involved in start codon and premature termination variants

The loss of a start codon disrupts normal translation initiation, whereas premature termination variants cause early translational termination, producing truncated and functionally impaired proteins. Both variant types represent critical genetic mutations that profoundly impact protein function (Fig. 7a). Functional enrichment analysis of genes with stop-lost, stop-gained, start-lost, and premature-start-codon-gain variants in hairy roots revealed significant enrichment in key metabolic pathways, including zeatin biosynthesis, flavonoid biosynthesis, ascorbate and aldarate metabolism, and glycosyltransferases (Fig. 7a, b). Notably, the MAPK signaling pathway also showed significant enrichment (Fig. 7b). Further quantification analysis of zeatin and flavonoid metabolite in normal plants versus hairy roots revealed a significant reduction in both zeatin derivatives (DL-dihydrozeatin and trans-zeatin) and multiple flavonoids, including bavachin, cianidanol, corymbosin, guaijaverin, hyperoside, isoquercitrin, isorhamnetin 3,4′-diglucoside, kaempferol 7-neohesperidoside, myricetin-3-galactoside, naringenin-6-C-glucoside, oroxylin A, panasenoside, and procyanidin B2 (Fig. 7c). In contrast, comparative metabolomic profiling identified elevated levels of glycoside compounds in hairy roots, notably 2-phenylethyl beta-D-glucopyranoside, 4-hydroxybenzoic acid 4-O-glucoside, aurantio-obtusin beta-D-glucoside, kaempferol 3,7-diglucoside, pinoresinol 4-O-glucoside, and salicylic acid glucoside (Fig. 7d). These findings provide critical insights into the metabolic reprogramming associated with hairy root development,

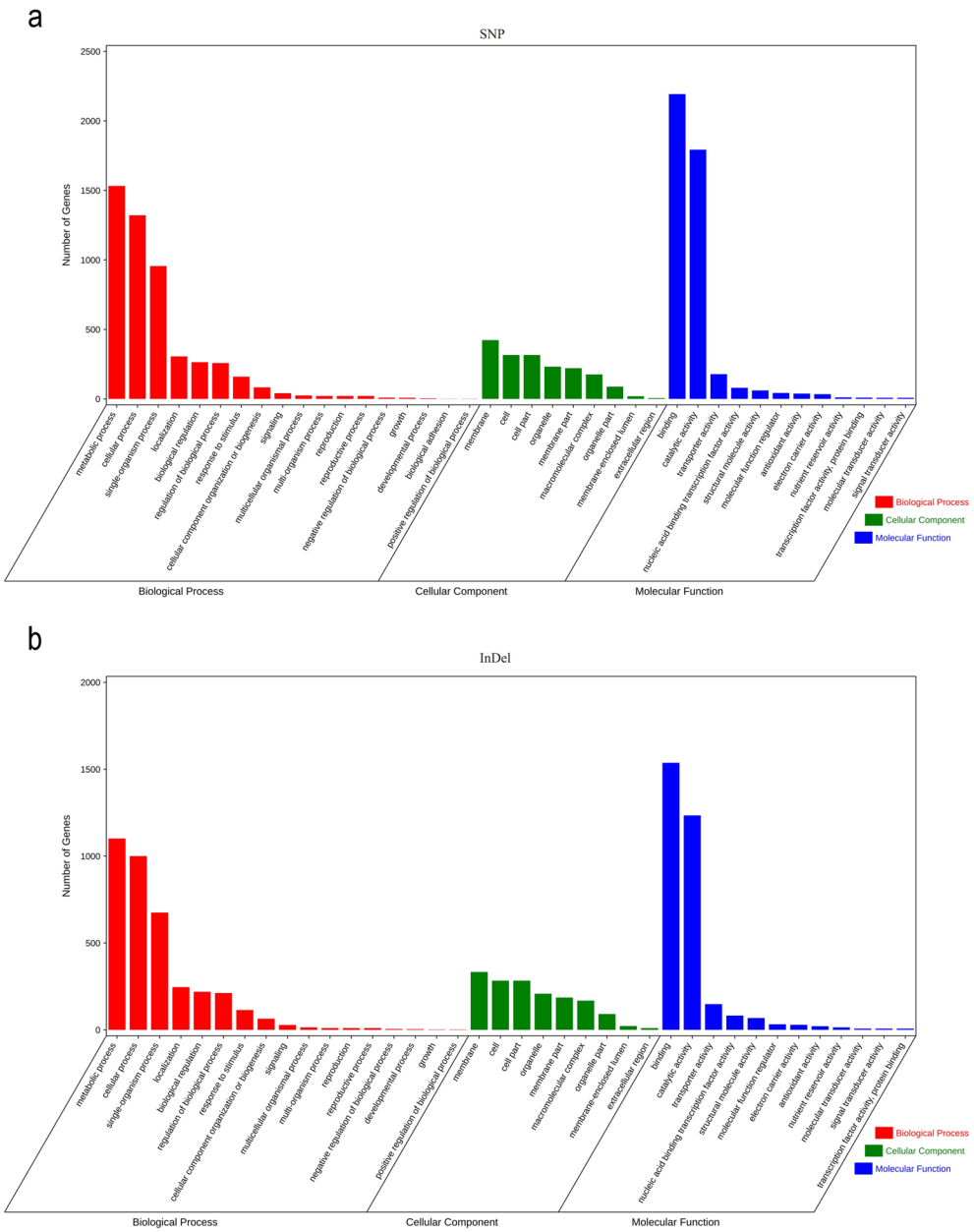

Fig. 4 **a** GO annotation classification of genes involved in effective SNPs. **b** GO annotation classification of genes involved in effective InDels

particularly the shift from flavonoid/zeatin biosynthesis toward glycoside accumulation.

NAC transcription factor family are closely associated with hairy root formation

Previous studies have demonstrated that NAC transcription factors play a pivotal role in root growth, development, and hairy roots formation (Xi et al. 2019; Yang et al. 2019; Xin et al. 2025). In this study, we observed that the NAC gene family exhibited the highest number of effective variations among the transcription factors (Fig. 6). Therefore, we systematically analyzed the functional domains and expression patterns of NAC transcription factor members harboring effective variants in hairy roots.

The chromosome localization results of NAC transcription factors showed that the 56 NAC transcription factors were unevenly distributed on 20 chromosomes, with the highest number (12) located on chromosome 4, followed by chromosome 17 (Fig. S3). The conserved motif analysis of the NAC family revealed that the family was mainly divided into seven subfamilies (G1-G7) (Figs. 8, 9a), containing a total of 20 motifs. Among them, subfamilies G1 and G2 shared similar motifs (e.g., motif 2, motif 3, and

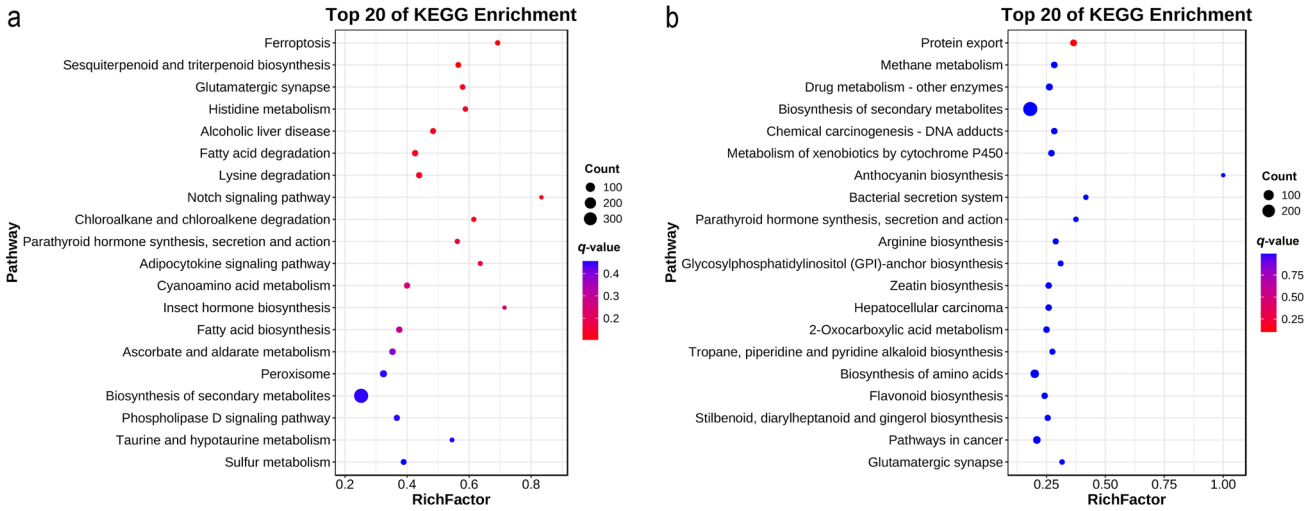

Fig. 5 **a** KEGG enrichment analysis of genes involved in effective SNPs. **b** KEGG enrichment analysis of genes involved in effective InDels

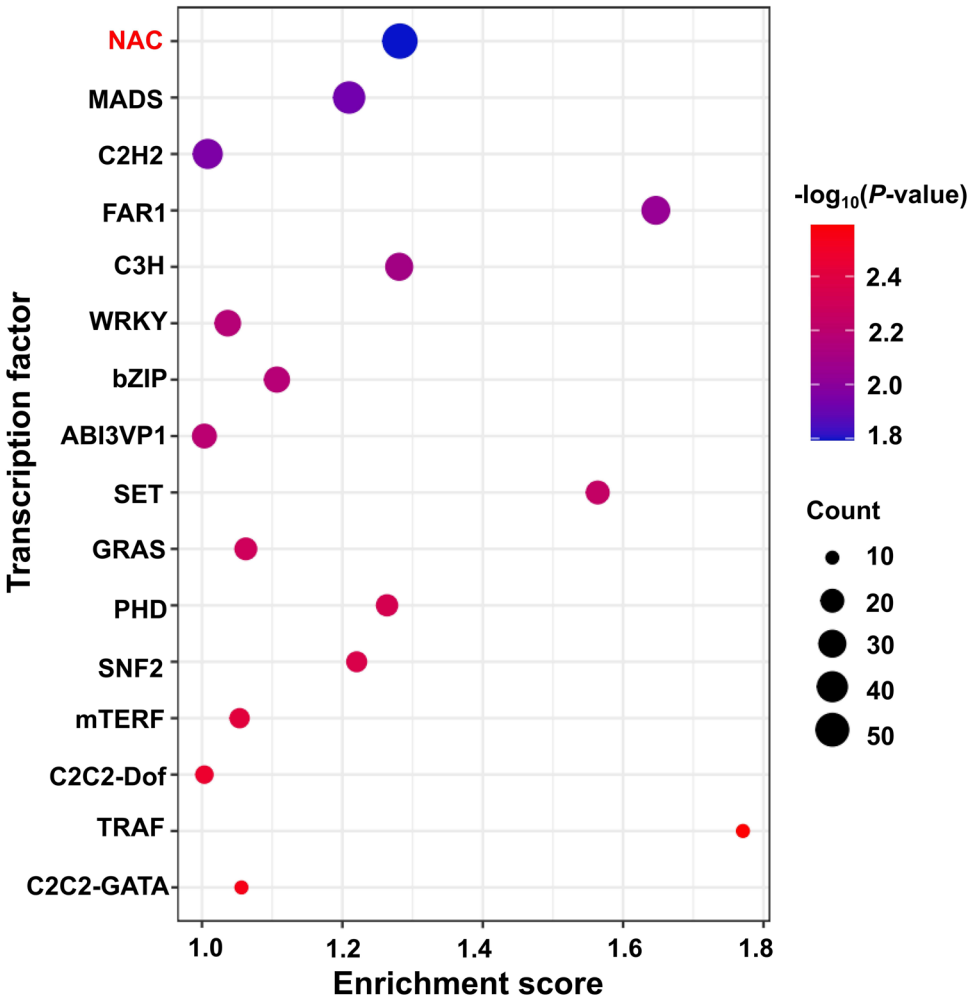

Fig. 6 Enrichment analysis diagram of transcription factor families

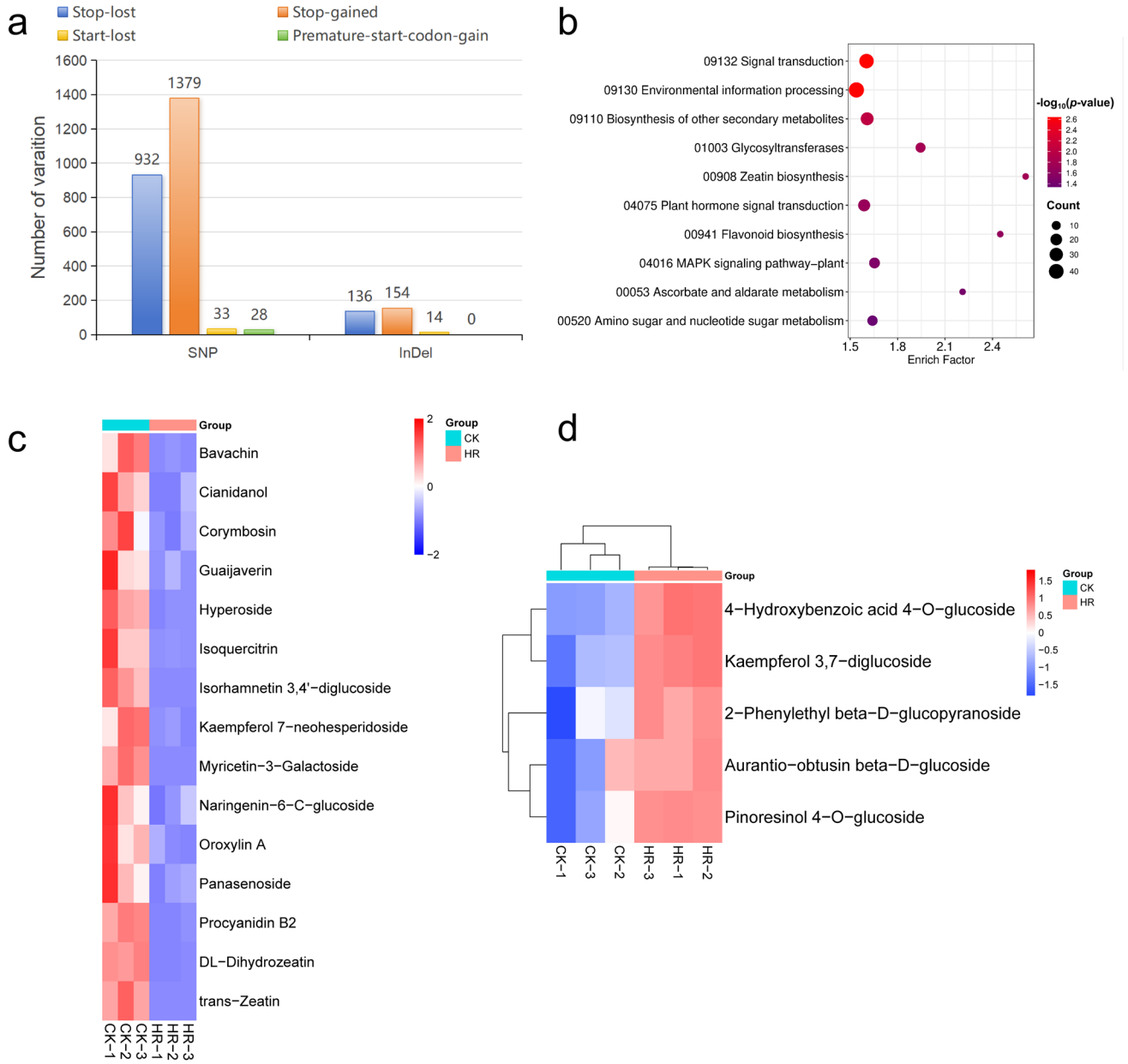

Fig. 7 **a** The number distribution of four types of variations (stop-lost, stop-gained, start-lost, and premature-start-codon-gain) among all SNPs and InDels. **b** KEGG enrichment analysis of genes involved in above four variations (stop-lost, stop-gained, start-lost, and premature-start-codon-gain). **c** Determination of zeatin and flavonoid pathway substances in normal plants and hairy roots. **d** Determination of some glycoside compounds in normal plants and hairy roots

motif 5), whereas subfamily G7 displayed a distinct motif composition, including motif 8, motif 9, and motif 11 (Fig. 8). Gene structure analysis indicated that, except for Nitab4.5_0000483g0290 which had 16 exons, the number of exons in other genes ranged from 1 to 6, and only two NAC genes had UTR regions (Fig. 8).

Through the collinearity analysis by MCScanX, 9 pairs of collinear genes were found to be distributed on different chromosomes (Fig. 9b), suggesting that tandem duplication and segmental duplication may participate in the expansion of the NAC genes. Using the STRING database to predict the potential interactions among NAC transcription factors, we found that Nitab4.5_0000033g0090 and Nitab4.5_0004318g0060 are key genes in the interaction network (Fig. 9c). The results of the GO enrichment analysis revealed that these NAC transcription factors were highly enriched in biological processes such as DNA binding and nucleic acid metabolism (Fig. 9d), which is consistent with the characteristics of NAC as transcription

Fig. 8 Phylogenetic tree, conserved motifs, and gene structure of NAC family members

factors. Based on the transcriptomic data of plants under drought stress treated with melatonin, we found that the expression levels of Nitab4.5_0000662g0120 (homologous gene of *ANAC72* in Arabidopsis), Nitab4.5_0000021g0030 (homologous gene of *ANAC062* in Arabidopsis), and Nitab4.5_0000913g0020 (homologous gene of *ANAC083* in Arabidopsis) changed under both drought stress and salicylic acid induction. While the expression of Nitab4.5_0001204g0100 increased under drought induction, and the expression of Nitab4.5_0001461g0040 increased 12 hours after salicylic acid treatment (Fig. S4a, b). But these genes did not respond to melatonin treatment (Fig. S4a) (Chen et al. 2021).

Discussion

Hairy root culture systems serve as powerful biotechnological platforms, functioning both as efficient production factories for plant-derived metabolites, recombinant proteins, and industrial compounds, and as versatile tools for plant genetic engineering and phytoremediation applications (Zhu et al. 2024). Tobacco, a well-established model plant, is abundant in secondary metabolites and possesses a wealth of publicly available genomic, transcriptomic, and metabolomic data. The Ri plasmid carried by *Agrobacterium rhizogenes* can infect wounded plant tissues, prompting the synthesis of specific phenolic compounds. This process activates genes in the Vir region of the Ri plasmid and integrates its T-DNA

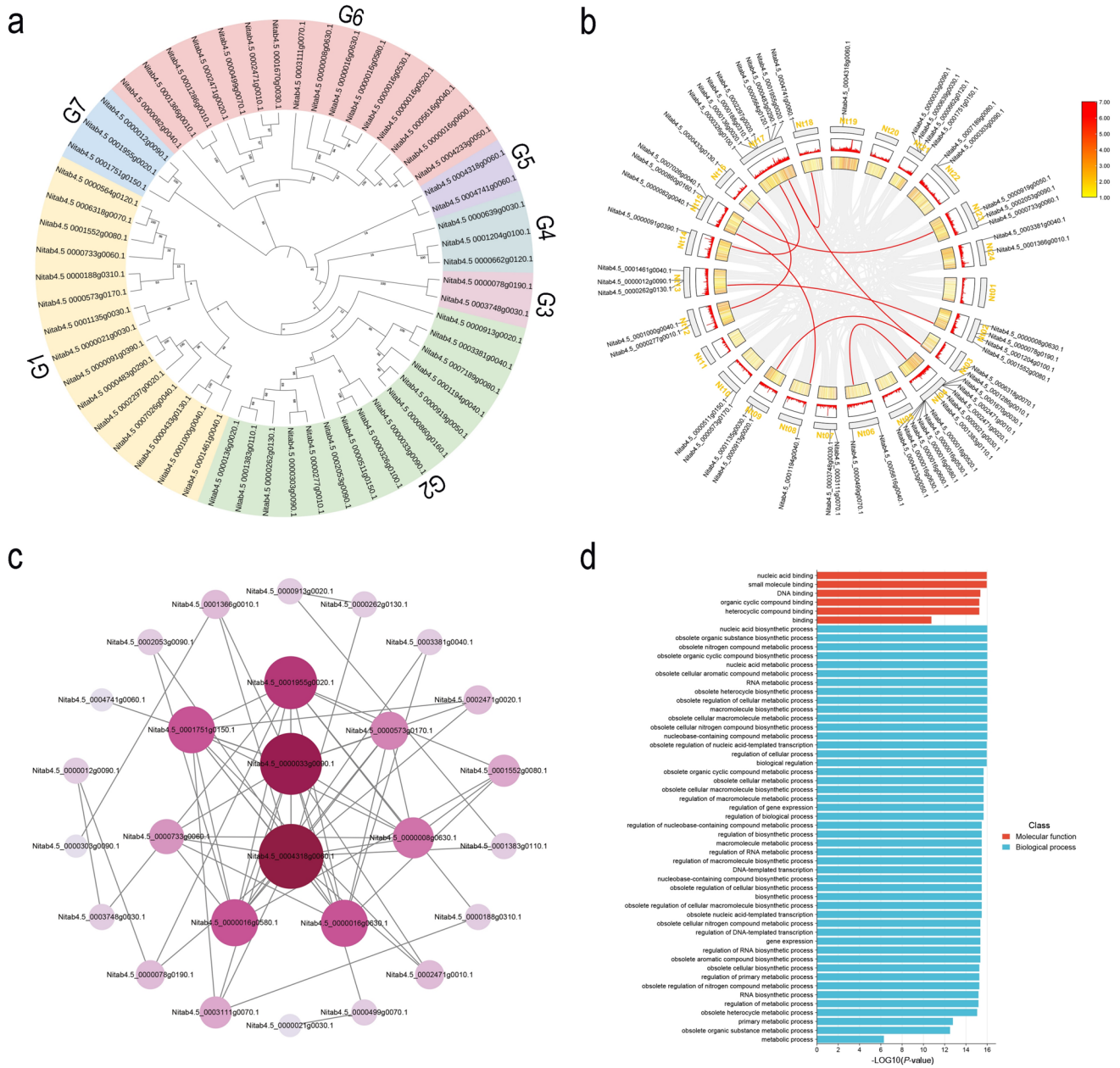

Fig. 9 **a** Phylogenetic tree of 56 NAC transcription factors. A phylogenetic tree was constructed by the neighbor-joining method based on MEGA11 with 1000 bootstrap replicates. The tree branched the NAC transcription factors into seven subgroups, which were represented by outer rings of different colors. **b** Distribution and collinearity of the NAC transcription factor family members. The two middle rings represent the gene density of each chromosome. The gray background lines indicate a collinear background, while the red lines represent the collinear relationship among NAC family members. **c** Protein–protein interaction network of key NAC family genes. Nodes represent proteins, black lines indicate interactions between nodes, and the darker the color, the more important the protein is in the interaction network. **d** GO enrichment analysis of the NAC family members.

into the plant genome, leading to hairy root formation (Wang et al. 2024). Research efforts have focused on the biologic and industrial applications of hairy roots.

In this study, we conducted WGRS-based comparative genomic analysis to explore the genetic differences between tobacco hairy roots and normal tobacco plants. We identified a total of 28,205 missense variants, which represented the largest proportion of effective SNPs. The identified InDels included 2485 frameshift variants and 869 non-frameshift variants, and total of 11,428 genes were effectively associated with variants. In addition, SVs have gained increasing attention for their potential significance (Campbell et al. 2016; Ma et al. 2024). Although the focus of our analysis in this study is on SNPs and small insertions/deletions

(InDels), SVs may have profound impacts on the formation of hairy roots and the accumulation of secondary metabolites at the genomic level. Future research should employ higher coverage whole-genome resequencing to conduct an in-depth analysis of gene SVs, in combination with SNP and InDel variants, to jointly elucidate the genetic mechanisms underlying hairy root formation.

GO enrichment analysis of the genes associated with effective variants showed that these variant-related genes were mainly enriched in MF related to molecular binding and catalytic activity, indicating that they may be involved in the biosynthesis of some secondary metabolites or enzymes in plant hormone metabolic pathways. In BP, hairy root-specific variant genes were mainly enriched in metabolic and cellular processes, which might be related to the strong anabolic and catabolic abilities of hairy roots and their hormone autotrophism. KEGG enrichment analysis identified secondary metabolite biosynthesis as the most significantly enriched pathway. The genes associated with SNPs were mainly significantly enriched in the Notch signaling, sesquiterpenoid and triterpenoid biosynthesis, and histidine metabolism pathways. Genes associated with InDels were mainly enriched in protein export pathways. Enrichment analysis of transcription factors associated with valid variants showed significant associations with NAC (56 genes), MYB (62 genes), and MADS (44 genes). These transcription factors exert important genetic regulatory functions in substance synthesis and metabolism and may be involved in metabolite synthesis and degradation in tobacco hairy roots (Gao et al. 2022; Li et al. 2024).

Hairy roots serve as bioreactors for enhancing the production of secondary metabolites, and genetic variations within them have significant impacts on metabolic processes (Bagal et al. 2023). The analysis of genetic variants, particularly affecting translation initiation and termination (such as stop-lost, stop-gained, start-lost, and premature-start-codon-gain), offers valuable insights into the molecular mechanisms underlying metabolic regulation in hairy roots (Cirulli et al. 2011). Our findings show that genes associated with these variants are enriched in key metabolic pathways, including zeatin biosynthesis, flavonoid biosynthesis, and glycosyltransferase activities, highlighting their role in regulating secondary metabolite production. The significant reduction in zeatin and flavonoid metabolites in hairy roots suggests that these genetic variants may impair the synthesis or regulation of these compounds. This is further supported by the enrichment of the MAPK signaling pathway, which is known for its role in stress response and developmental regulation (Zhang et al. 2018; Zhu et al. 2019). The observed accumulation of specific glycoside compounds, such as 2-phenylethyl beta-D-glucopyranoside and salicylic acid glucoside, indicates a potential shift in metabolic priorities, possibly as a compensatory mechanism or due to altered regulatory networks. These results highlight the complexity of metabolic regulation in hairy roots, and underscores for future studies on functional validation of these variations through targeted gene editing and metabolic engineering approaches to further elucidate their roles in hairy root development and metabolite accumulation.

Among the genes with effective variations in hairy roots, several encode NAC transcription factors were enriched. Previous study found that NAC transcription factors such as ANAC092 negatively regulate root development by binding to the promoters of *ARF8* and *PIN4* (Xi et al. 2019). In addition overexpression of *GmNAC109* in Arabidopsis has been shown to significantly enhance the drought and salt tolerance of transgenic plants (Yang et al. 2019). The interaction between LpNAC48 and the B-box family transcription factor LpBBX28 is crucial for regulating hairy root formation in *Lilium pumilum*, with gene silencing leading to defective hairy root development (Xin et al. 2025). These findings further support the importance of NAC transcription factors in the formation and genetic regulation of hairy roots, and provides a theoretical basis for improving plant root traits through genetic engineering.

In summary, this study provides insights into the molecular mechanisms underlying the formation of tobacco hairy roots and the biologic foundations of tobacco hairy roots various applications. In future, research may focus on the genes that in this study were found to be associated with secondary metabolite synthesis; a deep understanding of their precise functions will likely facilitate the development of hairy root systems for industrial applications. These findings lay the groundwork for further exploration of the genetic underpinnings of hairy root formation and their potential applications in biotechnology and agriculture.

Conclusions

In this study, we explored the genetic underpinnings of hairy root formation in tobacco through whole-genome resequencing (WGRS). A total of 3,059,775 SNPs and 242,577 InDels were identified, involving 23,923 genes, of which 11,428 exhibited effective variations. These genes were significantly enriched in multiple metabolic pathways, such as sesquiterpenoid and triterpenoid biosynthesis, as well as in various transcription factor gene families, including MADS, FAR1, C3H, and PHD. Notably, genes associated with stop-lost, stop-gained, start-lost, and premature-start-codon-gain are enriched in zeatin biosynthesis, flavonoid biosynthesis, and glycosyltransferase activities. In addition, hairy roots possessed lower levels of zeatin and flavonoid compounds, but higher levels of glycoside compounds. We further conducted functional domain and expression analyses of the NAC transcription factor family, which had the highest number of

enriched genes, and identified three key candidate genes potentially involved in hairy root formation in tobacco. These findings provide valuable insights into the genetic distinctions between hairy roots and normal tobacco plants and highlight potential key genes associated with hairy root formation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00425-025-04715-z.

Acknowledgements The authors thanks for LetPub (www.letpub.com.cn) for linguistic assistance and pre-submission expert review.

Author contributions Xiaozong Wu and Chaonan Shi designed this study. Zhiwen Zhu wrote the original draft preparation; Xiaozong Wu, Zhiwen Zhu, Peilin Li and Zhitao Qi participated in the analysis of resequencing data. Ruojie Zhu made some of the pictures in the article. Chaonan Shi and Xiaozong Wu revised the manuscript. All authors read and approved the final manuscript.

Funding This research was funded by the Doctoral fund project of Zhengzhou University of Light Industry, China (Project No.: 2024BSJJ011).

Data availability The authors confirm that the data supporting the findings of this study are available within the article.

Declarations

Conflicts of interest The authors declare no conflicts of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

Aghaali Z, Naghavi MR (2023) Biotechnological approaches for enhancing polyhydroxyalkanoates (PHAs) production: current and future perspectives. Curr Microbiol 80:345–358. https://doi.org/10.1007/s00284-023-03452-4

Bagal D, Chowdhary AA, Mehrotra S, Mishra S, Rathore S, Srivastava V (2023) Metabolic engineering in hairy roots: an outlook on production of plant secondary metabolites. Plant Physiol Biochem 201:e107847. https://doi.org/10.1016/j.plaphy.2023.107847

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36

Campbell BW, Hofstad AN, Sreekanta S, Fu F, Kono TJ, O'Rourke JA, Vance CP, Muehlbauer GJ, Stupar RM (2016) Fast neutron-induced structural rearrangements at a soybean *NAP1* locus result in gnarled trichomes. Theor Appl Genet 129:1725–1738. https://doi.org/10.1007/s00122-016-2735-x

Chen Z, Jia W, Li SW, Xu JY, Xu ZC (2021) Enhancement of *Nicotiana tabacum* resistance against dehydration-induced leaf senescence via metabolite/phytohormone-gene regulatory networks modulated by melatonin. Front Plant Sci 12:e686062. https://doi.org/10.3389/fpls.2021.686062

Chen CJ, Wu Y, Li JW, Wang X, Zeng ZH, Xu J, Liu YL, Feng JT, Chen H, He YH, Xia R (2023) TBtools-II: a "one for all, all for one" bioinformatics platform for biological big-data mining. Mol Plant 16:1733–1742. https://doi.org/10.1016/j.molp.2023.09.010

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu XY, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2*; *iso-3*. Fly 6:80–92. https://doi.org/10.4161/fly.19695

Cirulli ET, Heinzen EL, Dietrich FS, Shianna KV, Singh A, Maia JM, Goedert JJ, Goldstein DB et al (2011) A whole-genome analysis of premature termination codons. Genomics 98:337–342. https://doi.org/10.1016/j.ygeno.2011.07.001

Gao Y, Lin YJ, Xu M, Bian HX et al (2022) The role and interaction between transcription factor NAC-NOR and DNA demethylase SlDML2 in the biosynthesis of tomato fruit flavor volatiles. New Phytol 235:1913–1926. https://doi.org/10.1111/nph.18301

Gerszberg A, Wiktorek-Smagur A (2022) Hairy root cultures as a multitask platform for green biotechnology. Plant Cell Tiss Organ Cult 150:493–509. https://doi.org/10.1007/s11240-022-02316-2

Hou XJ, Li JM, Liu BL, Wei L (2017) Co-expression of basic helix-loop-helix protein (bHLH) and transcriptional activator-Myb genes induced anthocyanin biosynthesis in hairy root culture of *Nicotiana tabacum L* and *Ipomea tricolor*. Acta Physiol Plant 39:e59. https://doi.org/10.1007/s11738-017-2362-4

Ji HY, Yang BY, Jing YY, Luo Y, Li B, Yan YY, Zhang G, Zhao F, Wang BQ, Peng L, Hu BX (2023) Trehalose and brassinolide enhance the signature ingredient accumulation and anti-oxidant activity in the hairy root cultures of *Polygala tenuifolia* Willd. Ind Crop Prod 196:e116521. https://doi.org/10.1016/j.indcrop.2023.116521

Letunic I, Bork P (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 9:W293–W296. https://doi.org/10.1093/nar/gkab301

Li S, Dong Y, Li D, Shi S, Zhao N, Liao J, Liu Y, Chen H (2024) Eggplant transcription factor SmMYB5 integrates jasmonate and light signaling during anthocyanin biosynthesis. Plant Physiol 194:1139–1165. https://doi.org/10.1093/plphys/kiad531

Lin J, Yin X, Zeng YR, Hong XY et al (2023) Progress and prospect: biosynthesis of plant natural products based on plant chassis. Biotechnol Adv 69:108266–108284. https://doi.org/10.1016/j.biotechadv.2023.108266

Ma YY, Wang Y, Zhou ZQ, Zhang RQ et al (2024) A large presence/absence variation in the promotor of the *ClLOG* gene determines trichome elongation in watermelon. Theor Appl Genet 137:e98. https://doi.org/10.1007/s00122-024-04601-4

McKenna A, Hanna M, Branks E, Sivachenko A et al (2010) The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://doi.org/10.1101/gr.107524.110

Molina-Hidalgo FJ, Vazquez-Vilar M, D'Andrea L, Demurtas OC, Fraser P, Giuliano G, Bock R, Orzaez D, Goossens A (2021) Engineering metabolism in *Nicotiana* species: a promising future. Trends Biotechnol 39:901–913. https://doi.org/10.1016/j.tibtech.2020.11.012

Qin SQ, Liu YR, Yan JP, Lin SW, Zhang WJ, Wang BW (2022) An optimized tobacco hairy root induction system for functional analysis of nicotine biosynthesis-related genes. Agronomy 12:e348. https://doi.org/10.3390/agronomy12020348

Schenck CA, Anthony TM, Jacobs M, Jones AD, Last RL (2022) Natural variation meets synthetic biology: promiscuous trichome-expressed acyltransferases from *Nicotiana*. Plant Physiol 190:146–164. https://doi.org/10.1093/plphys/kiac192

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303

Szklarczyk D, Kirsch R, Koutrouli M, Nastou K et al (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 51:D638–D646. https://doi.org/10.1093/nar/gkac1000

Tamura K, Stecher G, Kumar S (2021) MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol 38:3022–3027. https://doi.org/10.1093/molbev/msab120

Tang DD, Chen MJ, Huang XH, Zhang GC, Zeng L, Zhang GS, Wu SJ, Wang YL (2023) SRplot: a free online platform for data visualization and graphing. PLoS ONE 18:e0294236. https://doi.org/10.1371/journal.pone.0294236

Vashisht A, Mondal AK, Woodall J, Kolhe R (2024) From genomic exploration to personalized treatment: next-generation sequencing in oncology. Curr Issues Mol Biol 46:12527–12549. https://doi.org/10.3390/cimb46110744

Wang QL, Luo X, Wan K, Shi YS, Cui YZL, Liu Y (2024) Analysis of male sterility of *BBLa* overexpressed tobacco based on re-sequencing technology. Seed 43:37–43. https://doi.org/10.16590/j.cnki.1001-4705.2024.02.037

Wu XD, Xiang DC, Zhang W, Zhao GY, Yin ZJ (2024) Identification of breed-specific SNPs of danish large white pig in comparison with four chinese local pig breed genomes. Genes 15:e623. https://doi.org/10.3390/genes15050623

Xi DD, Chen X, Wang YX, Zhong RL, He JM, Shen JB, Ming F (2019) *Arabidopsis* ANAC092 regulates auxin-mediated root development by binding to the *ARF8* and *PIN4* promoters. Integr Plant Biol 9:1015–1031. https://doi.org/10.1111/jipb.12735

Xin Y, Pan WQ, Zhao YJ, Yang CL et al (2025) The NAC transcription factor LpNAC48 promotes trichome formation in *Lilium pumilum*. Plant Physiol 197:kiaf001. https://doi.org/10.1093/plphys/kiaf001

Yang XF, Kim MY, Ha JM, Lee SH (2019) Overexpression of the soybean NAC gene *GmNAC109* increases lateral root formation and abiotic stress tolerance in transgenic *Arabidopsis* plants. Front Plant Sci 10:e1036. https://doi.org/10.3389/fpls.2019.01036

Yang PF, Shang MY, Bao JJ, Liu TY, Xiong JK, Huang JP, Sun JH, Zhang L (2024) Whole-genome resequencing revealed selective signatures for growth traits in Hu and Gangba sheep. Genes 15:e551. https://doi.org/10.3390/genes15050551

Youngmok J, Dongsu H (2022) BWA-MEME: BWA-MEM emulated with a machine learning approah. Bioinformatics 38:2404–2413. https://doi.org/10.1093/bioinformatics/btac137

Zhang MM, Su JB, Zhang Y, Xu J, Zhang SQ (2018) Conveying endogenous and exogenous signals: MAPK cascades in plant growth and defense. Curr Opin Plant Biol 45:1–10. https://doi.org/10.1016/j.pbi.2018.04.012

Zhu QK, Shao YM, Ge ST, Zhang MM et al (2019) A MAPK cascade downstream of IDA-HAE/HSL2 ligand-receptor pair in lateral root emergence. Nat Plants 5:414–423. https://doi.org/10.1038/s41477-019-0396-x

Zhu YT, Zhu X, Wang LH, Wang YL et al (2024) Plant hairy roots: Induction, applications, limitations and prospects. Ind Crop Prod 219:e119104. https://doi.org/10.1016/j.indcrop.2024.119104

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.