frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 21 May 2024
DOI 10.3389/fgene.2024.1395224

OPEN ACCESS

EDITED BY
Diego Hojsgaard,
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Germany

REVIEWED BY
Diaga Diouf,
Cheikh Anta Diop University, Senegal
Apoorv Tiwari,
National Agri-Food Biotechnology Institute, India

*CORRESPONDENCE
Weina Ge,
gwn-06@163.com
Zhenyi Wang,
zhenyiwang0301@163.com

†These authors have contributed equally to this work

RECEIVED 03 March 2024
ACCEPTED 02 May 2024
PUBLISHED 21 May 2024

CITATION
Chen H, Liu F, Chen J, Ji K, Cui Y, Ge W and Wang Z (2024), Identification, molecular evolution, codon bias, and expansion analysis of NLP transcription factor family in foxtail millet (*Setaria italica* L.) and closely related crops. *Front. Genet.* 15:1395224. doi: 10.3389/fgene.2024.1395224

Identification, molecular evolution, codon bias, and expansion analysis of NLP transcription factor family in foxtail millet (*Setaria italica* L.) and closely related crops

Huilong Chen[1†], Fang Liu[1†], Jing Chen[1], Kexin Ji[1], Yutong Cui[2], Weina Ge[1*] and Zhenyi Wang[1*]

[1]College of Life Sciences, North China University of Science and Technology, Tangshan, Hebei, China, [2]College of Management, North China University of Science and Technology, Tangshan, Hebei, China

The NODULE-INCEPTION-like protein (NLP) family is a plant-specific transcription factor (TF) family involved in nitrate transport and assimilation in plants, which are essential for improving plant nitrogen use efficiency. Currently, the molecular nature and evolutionary trajectory of NLP genes in the C4 model crop foxtail millet are unknown. Therefore, we performed a comprehensive analysis of NLP and molecular evolution in foxtail millet by scanning the genomes of foxtail millet and representative species of the plant kingdom. We identified seven NLP genes in the foxtail millet genome, all of which are individually and separately distributed on different chromosomes. They were not structurally identical to each other and were mainly expressed on root tissues. We unearthed two key genes (*Si5G004100.1* and *Si6G248300.1*) with a variety of excellent characteristics. Regarding its molecular evolution, we found that NLP genes in Gramineae mainly underwent dispersed duplication, but maize NLP genes were mainly generated via WGD events. Other factors such as base mutations and natural selection have combined to promote the evolution of NLP genes. Intriguingly, the family in plants showed a gradual expansion during evolution with more duplications than losses, contrary to most gene families. In conclusion, this study advances the use of NLP genetic resources and the understanding of molecular evolution in cereals.

KEYWORDS

foxtail millet, NODULE-INCEPTION-like protein, transcription factor, structure, molecular evolution, codon bias

1 Introduction

Nitrogen is an important component of amino acids, nucleotides, chlorophyll, and hormones, and it is also an essential nutrient for all organisms (Tegeder and Masclaux-Daubresse, 2018; Li D. et al., 2024). The growth and development of crops depend on their ability to take up and utilize nitrogen. The main nitrogen source of terrestrial plants is nitrate. However, most terrestrial soils in the world contain less nitrogen, and the absorption and utilization of nitrogen by most terrestrial plants only account for 30%–50% of the nitrogen

application amount (Chen et al., 2020). Therefore, to improve the utilization of nitrogen fertilizer (Liu et al., 2018), research on genes related to nitrogen uptake and transport has become increasingly popular in recent years. Plants and fungi are the only multicellular organisms capable of absorbing inorganic nitrogen. In higher plants, inorganic nitrogen is primarily absorbed from the soil in the form of NO^{3-} by various complex regulatory mechanisms evolved by plants (Crawford, 1995; Castaings et al., 2009). Among them, genes encoding nitrate transporter (NRT), nitrate reductase (NIA), and nitrite reductase (NIR) play an important role in the absorption and utilization of nitrate by plants (Forde, 2000; Li Y. et al., 2024). And the NODULE-INCEPTION-like protein (NLP) transcription factors play a central role in nitrate sensing and signaling. The NLP transcription factor family is a plant-specific transcription factor (TF) (Feng et al., 2020) family that participates in nitrate signal transduction and assimilation processes (Nishida and Suzaki, 2018; Gao et al., 2022).

The earliest research on NLP family can date back to the leguminous model plant *Lotus japonicus* nodule inception (NIN) (Schauser et al., 1999). The most obvious feature of NIN protein is a highly conserved long sequence composed of 60-amino acid, known as RWP-RK sequence (also called RWP × RK motif). Some highly homologous genes to NIN have been found in legumes, which were named NIN-like protein (NLP) (Borisov et al., 2003). NLP carries mainly the RWP-RK (Chardin et al., 2014) and PB1 (Sumimoto et al., 2007) conserved domains. NLP is homologous to NIN in the RWP-RK structural domain and the *N*-terminal region (Schauser et al., 2005; Suzuki et al., 2013). The RWP-RK structural domain is highly conserved, can bind and function with DNA, and its activity is independent of the nitrate signal (Chardin et al., 2014; Liu et al., 2018). The RWP-RK structure is an activation domain for transcription-mediated nitrate signaling, and the PB1 domain is located at the carboxyl terminus and can participate in protein-protein interactions (Ge et al., 2018). Moreover, the N-terminus of NLP has a highly conserved cGMP phosphodiesterase domain in addition to the RWP-RK domain, which may be involved in signal transduction or dimerization (Yu et al., 2023). There have been many genome-wide studies of the NLP gene family, such as *Arabidopsis thaliana* (Schauser et al., 2005), rice (*Oryza sativa*) (Schauser et al., 2005), pepper (Wu et al., 2023), alfalfa (Wu et al., 2023), cucumber seedlings (Li Y. et al., 2024), tea tree (Li D. et al., 2024), *Physcomitrella patens* (Crawford, 1995), wheat (Kumar et al., 2018), maize (*Zea mays*) (Ge et al., 2018), *Brassica napus* (Feng et al., 2020), Chinese cabbage (Chen et al., 2022a) and so on. Evolutionary studies of the Chinese cabbage NLP family have revealed that the origins of duplication of the NLP gene family in the genus Brassica were almost exclusively derived from the WGD type (Chen et al., 2022a), yet the molecular evolutionary characterization of the other plant taxa is not known.

Foxtail millet (*Setaria italica*) is an ancient diploid C4 gramineous model crop (Kumar et al., 2013) with a long history of cultivation. The process of humans cultivating wild weed green foxtail (*Setaria viridis* L.) into foxtail millet can be traced back to about 11,000 years ago (Li et al., 2022). Foxtail millet is an abiotic stress-tolerant plant with a short life cycle, which can be inbreeding and self-pollination. It can be grown as a food crop in the saline-prone regions of Asia and under adverse conditions such as drought and semi-drought conditions, as well as hay and fodder in Australia, southern Europe, South America, and North Africa (Sreenivasulu et al., 2004; Sharma and Niranjan, 2018). The genome of foxtail millet has been sequenced, and it has a small diploid genome (approximately 515 Mb) with a relatively small amount of repetitive DNA (Zhang G. et al., 2012; Bennetzen et al., 2012). Foxtail millet has gradually become a model species for studying gramineous crops (Doust et al., 2009; Yang et al., 2020) and provides data resources for studying the NLP gene family in foxtail millet.

In this study, we aimed to deepen our understanding of the functions of the NLP gene family in foxtail millet by conducting a genome-wide identification and bioinformatics analysis of the NLP gene family. The analysis included codon bias analysis, gene structure analysis, protein structure analysis, phylogenetic analysis, chromosome localization analysis, homology analysis, duplication type analysis, expansion analysis, and tissue expression analysis. The results of this analysis are expected to be fully applied and contribute to the research of the NLP gene family in foxtail millet.

2 Materials and methods

2.1 Data collection

We obtained the genome data for *A. thaliana* from the TAIR database (Berardini et al., 2015) (https://www.arabidopsis.org/), for rice from the Rice Genome Database (Ouyang et al., 2007) (http://rice.plantbiology.msu.edu/), and for *Chlorella variabilis* and *Chara braunii* from the NCBI database (https://www.ncbi.nlm.nih.gov/). Additionally, we downloaded genome data for *Chlamydomonas reinhardtii* (Merchant et al., 2007), *Volvox carteri* (Prochnik et al., 2010), *Ostreococcus lucimarinus* (Palenik et al., 2007), *Amborella trichopoda* (Project et al., 2013), *P. patens* (Lang et al., 2018), *Selaginella moellendorfii* (Banks et al., 2011), maize (Hirsch et al., 2016), sorghum (*Sorghum bicolor*) (McCormick et al., 2018) and foxtail millet (Bennetzen et al., 2012) in the JGI database (https://genome.jgi.doe.gov/).

After reviewing the literature (Jagadhesan et al., 2020), we identified six members of the NLP gene family in rice and obtained the nucleic acid and protein sequences for the rice NLP gene family. Using the six NLP sequences from rice as targets, we searched for candidate NLP gene family members in other species (*C. variabilis*, *C. braunii*, *C. reinhardtii*, *V. carteri*, *O. lucimarinus*, *P. patens*, *S. moellendorfii*, *A. trichopoda*, sorghum, maize, and *A. thaliana*) by performing a local tool Blastp (Altschul et al., 1990; Altschul et al., 1997) search with an e-value less than $1e^{-5}$ against all protein sequences in each species' database. We then used the online tools Pfam (Finn et al., 2014) (http://pfam.xfam.org/) and SMART (Letunic et al., 2012) (http://smart.embl-heidelberg.de/) to confirm the presence of two conserved structural domains, RWP-RK and PB1. After filtering out the NLP sequences for each species, we manually modified the prefixes of the original IDs to the initials of the species' Latin names to facilitate analysis.

2.2 Two-dimensional and three-dimensional structure of NLP family in foxtail millet

We utilized the online tool SOPMA (Geourjon and Deleage, 1995) (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html) to predict the two-dimensional structure of proteins. SOPMA's basic principle for predicting amino acid

secondary structure is based on the conserved and physicochemical properties of the sequence, which utilizes the self-similarity of the protein sequence. By comparing and analyzing the protein sequences, SOPMA can identify the repeats and conserved structures in the sequences and predict the secondary structure of the proteins. A significant improvement of the SOPMA method is that it takes into account sequence alignment information belonging to the same family, which makes the prediction results more accurate (Geourjon and Deleage, 1995). We uploaded the protein files of each of the seven foxtail millet NLP genes (Supplementary Material S1) to the SOPMA website with the default parameters (Number of conformational states: 4 (Helix, Sheet, Turn, Coil), Similarity threshold: 8, Window width: 17), and recorded the proportions of alpha helix, extended strand, beta turn, and randon coil in the results in tabular form.

PHYRE2 is a web-based tool that uses advanced remote homology detection methods to build 3D models of proteins, predict ligand binding sites, and analyze the effects of amino acid variants (e.g., nonsynonymous snps (nssnps) on the user's protein sequence (Kelley et al., 2015). We used the online tool PHYRE2 (Kelley et al., 2015) (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) to predict the three-dimensional structure of proteins. We separately uploaded the seven NLP protein sequences of foxtail millet to the PHYRE2 website with default parameters (Modelling Mode info icon: Normal, please tick as appropriate: Other), the final result was sent to the mailbox in the form of .pdb files, and the results were visualized using the native tool VMD (Humphrey et al., 1996) (version 1.9.4a51). We adjusted the color settings to helix-ColorID0, sheet-ColorID1, turn-ColorID4, and coil-ColorID7, with a transparent background color.

2.3 Promoter analysis of foxtail millet NLP genes

We extracted a 2000 bp sequence upstream of seven NLP genes from the foxtail millet genome file as promoter sequences using a self-design Python script 'finally_promoter_genome.py' (https://github.com/ChenHuilong1223/CFVisual/blob/main/finally_promoter_genome.py) based on the gff3 data of foxtail millet. The promoter sequences (Supplementary Material S2) were analyzed using the promoter analysis website PlantCARE (Lescot et al., 2002) (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/), and the resulting promoter elements located on the negative strand were removed. The remaining promoter elements were classified by function according to the classification table. Using the Neighbor-joining method in MEGA6 (Tamura et al., 2013), seven NLP protein sequences of foxtail millet were compared, and the parameters were set as follows: repeat 1,000 times, Poisson model and pairwise deletion, with other parameters set to default. We exported the resulting tree file, which contained bootstrap values, and used CFVisual (Chen et al., 2022b) to visualize the results.

Subsequently, we used the online tool JASPAR (Castro-Mondragon et al., 2022) (https://jaspar.elixir.no/search?q=&collection=CORE&tax_group=plants) to cross-check the prediction of NLP transcription factor binding sites (TFBSs), thereby increasing the credibility of PlantCARE promoter analysis results. As far as we know, JASPAR is a comprehensive database website dedicated to predicting TFBSs, and its prediction results include experimental evidence and algorithm prediction. It is believed that the prediction results of *cis*-acting elements in the putative promoter sequence of foxtail millet NLP TFs include both TFBSs and non-TFBSs such as hormones. We chose the following scheme for cross-validation. We first investigated and classified all the *cis*-acting elements we identified, and then we selected the results of known TFBSs to verify in the JASPAR database. In the JASPAR CORE non redundant database from plants, we selected all the MYB TFs corresponding to the *cis*-acting element MBS, and predicted the binding sites with the seven NLP promoter sequences of foxtail millet, respectively. The relative contour score threshold was 100%. Finally, we removed the promoter located on the negative chain in the prediction results, and compared the prediction results with the PlantCARE promoter analysis results.

2.4 Tissue expression analysis of foxtail millet NLP genes

We downloaded the transcriptome data for the root, stem, leaf, and spica of Zhang Gu from the public database (Zhang G. et al., 2012) (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36391) and performed a two-way homology alignment of the foxtail millet NLP family sequences with the Zhang Gu sequences using the native tool Blastp. We then extracted the corresponding results with a self-made Perl script (Supplementary Program S1) and obtained the expression data (RPKM) and log2 (RPKM) for the NLP gene family in foxtail millet from the transcriptome data. Next, we used the online tool Venn (Chen et al., 2021) (http://www.ehbio.com/test/venn/#/) to perform statistical analysis on the NLP gene expression data of foxtail millet, and the online tool Hiplot (https://hiplot.com.cn/basic) to generate a tissue expression heat map for the foxtail millet NLP gene family. We also downloaded the transcriptome data for the rice expression matrix from the Rice Genome Database and calculated log2 (FPKM) values to generate a tissue expression heat map for the rice NLP gene family using the same methods.

2.5 Analysis of protein interactions in foxtail millet NLP proteins

We predicted the interaction of NLP family members with related proteins in foxtail millet by the online website STRING (Szklarczyk et al., 2021) (https://string-db.org/). We predicted interactions for each of the seven NLP families individually, with a minimum required interaction score set to high confidence (0.700) and the maximum number of interactors set to 5, with other parameters set to default values. Furthermore, we also uploaded all protein sequences of the NLP family in foxtail millet to the website in batches for prediction.

2.6 Analysis of GO and KEGG of NLP genes in foxtail millet

We performed gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses on the seven NLP gene family

members of foxtail millet using the agriGO v2.0 (Tian et al., 2017) (http://systemsbiology.cau.edu.cn/agriGOv2/) and KOBAS 3.0 (Bu et al., 2021) (http://kobas.cbi.pku.edu.cn/kobas3/?t=1) databases. We used Microsoft Office Excel to organize the GO function annotation results and KEGG analysis results.

2.7 Chromosome distribution analysis of the NLP family in foxtail millet and closely related crops

We counted the number of foxtail millet, rice, sorghum, and maize in each group based on the grouping of the phylogenetic tree. Then, we created a graph using Microsoft Excel with foxtail millet, rice, sorghum, and maize as the horizontal axis and the number as the vertical axis.

2.8 Duplication type analysis of NLP family in foxtail millet and closely related crops

We used the dupicate_gene_classifier program (Wang et al., 2012) in MCScanX to identify the five types of duplications (whole genome/segmental, tandem, proximal, dispersed, and singleton) of the whole genome and NLP gene families of foxtail millet, rice, sorghum, and maize. We also counted genome-wide collinear information tables with the NLP gene family using our previously published method (Chen et al., 2022a).

2.9 Codon bias analysis and ENC-Plot mapping of NLP families in foxtail millet and closely related crops

We downloaded the coding sequence (CDS) data of foxtail millet from the JGI database and extracted the CDSs of the NLP gene family in foxtail millet using a homemade Python program. We then screened the CDSs based on the following criteria: the sequence length was greater than or equal to 300 bp, the start codon was ATG (AUG), the stop codon was TAG (UAG), TAA (UAA), TGA (UGA), and the sequence did not contain other bases except A, T (U), G, C.

After the screening, we obtained the CDSs of the NLP genes in foxtail millet. We processed these seven CDSs with CodonW version 1.4.2 (https://sourceforge.net/projects/codonw/) and collated the results using Microsoft Office Excel. We used parameters such as codon adaptation index (CAI), effective number of codons (ENC), and relative synonymous codon usage (RSCU) for data processing. Using the *ggplot2* package in the R language to draw the ENC-plot diagram, the ENC value of each CDS is the ordinate, and the GC3 s is the abscissa to draw the two-dimensional scatter plot. The standard curve calculation formula was: ENC=2+GC3s+29/[GC3s2+(1—GC3s)2] (Wright, 1990). An RSCU value greater than 1 indicates that the codon is used more frequently, an RSCU equal to 1 indicates that the codon has no preference, and an RSCU less than 1 indicates that the codon is used less frequently (Sharp and Li, 1986). Similarly, we used the same method to analyze the codon bias of NLP genes in foxtail millet closely related species (rice, sorghum, and maize).

2.10 Determination of the optimal codon of the NLP family in foxtail millet and closely related crops

We used Microsoft Excel to determine the optimal codons. We sorted the genes based on their ENC preferences and selected 10% of genes at each end to determine the high and low expression genes. We then identified codons with a ΔRSCU (ΔRSCU = high expression - low expression) greater than 0.08 in the high and low expression genes as high expression superior codons. Finally, we selected the codon which has the highest RSCU value (RSCU > 1) in each amino acid as the high frequency superior codon. If the codon was a high-expression superior codon, it was considered the optimal codon.

2.11 Construction of a phylogenetic tree of the NLP family of representative species in plants

We used the protein sequences of 50 NLP families in *C. braunii*, *O. lucimarinus*, *A. trichopoda*, *P. patens*, *S. moellendorfii*, *A. thaliana*, rice, maize, sorghum, and foxtail millet to construct a phylogenetic tree using the local tree-building software MEGA6. Using ClustalW in MEGA6 for multiple sequence alignment, default parameters. And the Maximum Likelihood tree construction method was selected. The parameters were set as follows: the best protein model was JTT + G + I with 1,000 repetitions and pairwise deletion. We then used the online tool EvolView (Zhang H. et al., 2012) (https://www.evolgenius.info/) to build a phylogenetic tree.

2.12 Analysis of selection pressure on NLP families of representative species in plants

We used a homemade Python program to delete the termination codons of the CDS files of the ten studied species and then compared them using MEGA6. We saved the comparison files to obtain the comparison files and then constructed a phylogenetic tree using the NJ method. We saved the tree file without Branch Length. Subsequently, positive selection analysis was performed by EasyCodeML (Gao et al., 2019). In the preset mode, the site model was used to select the positive selection site; in the custom mode, the free-ratio model of the branch model was made, and the branch with $\omega > 1$ in the result was marked as the foreground branch, and the five foreground branches were a group. Then the free ratio model was changed to a double-ratio model. If the result was still $\omega > 1$, the branch was positively selected.

2.13 Exon-intron analysis of NLP families of representative species in plants

We downloaded the GFF3 data for *A. thaliana* from the TAIR database, for rice from the Rice Genome Database, for *C. braunii* from the NCBI database, and for *O. lucimarinus*, *A. trichopoda*, *P. patens*, *S. moellendorfii*, maize, sorghum, and foxtail millet from the JGI database. We used a homemade Python script to extract the NLP

gene structure information from the GFF3 data of each species and merged it into the same file. We then performed statistical analysis on the gene structure and visualized the results using CFVisual software.

2.14 Analysis of conserved motifs and structural domains of NLP families of representative species in plants

The MEME Suite web server (Bailey et al., 2009) (https://meme-suite.org/meme/tools/meme) can perform four types of motif analysis: motif discovery, motif-motif database searching, motif-sequence database searching and assignment of function. It is a good tool for discovering and searching sequence motifs. These sequence motifs represent features such as DNA binding sites and protein interaction domains. We uploaded 50 NLP protein sequences of the study species for motif discovery, and used the Multiple Em for Motif Elicitation (MEME) algorithm. We set the maximum value of the motif to 15 (Ge et al., 2018; Liu et al., 2018) and left the other parameters as default (Select the motif discovery mode: Classic mode, Select the sequence alphabet: DNA, RNA or Protein; Select the site distribution: Zero or One Occurrence Per Sequence (zoops)). We saved the MEME results (meme.xml) locally. Then we uploaded 50 NLP protein sequences of the study species to the Pfam database for analysis and saved the analysis results in text form. Finally, we used CFVisual software to draw the motif and domain together on a map based on the location information in the result files of MEME and Pfam. To determine the location of the motif and domain association.

2.15 Homology analysis of NLP families of representative species in plants

Homology analysis of the NLP family of ten species was performed using OrthoMCL (Li et al., 2003) software. The parameter settings were percent Match Cutoff of 75 and an e-value Exponent Cutoff of −10. We created orthologous network maps using the native software Cytoscape (Shannon et al., 2003), and homologous radar maps were created using Microsoft Excel.

2.16 Expansion analysis of the NLP family of representative species in plants

We identified the evolutionary relationships of the ten studied species containing NLP genes from the available literature (Geourjon and Deleage, 1995; Humphrey et al., 1996; Letunic et al., 2012; Kelley et al., 2015; Chen et al., 2021; Szklarczyk et al., 2021). We then constructed a species evolutionary tree using MEGA6 to understand the evolutionary process and genetic relationship among the studied species. Based on species and phylogenetic trees, we determined the gains and losses of NLP genes in the evolution of ten species through software Notung (Chen et al., 2000).

3 Results

3.1 Data collection

After a homology search, we identified seven sequences in the foxtail millet genome: *Si5G004100.1*, *Si3G084600.1*, *Si9G553000.1*, *Si8G074000.1*, *Si2G298700.1*, *Si1G094300.1*, and *Si6G248300.1*. We also identified nine sequences in the dicotyledonous plant *Arabidopsis* and nine, five, and seven sequences in the closely related species maize, sorghum, and rice, respectively. Three sequences were identified in the basal angiosperm *A. trichopoda*, while two and eight sequences were identified in the lower angiosperm *S. moellendorfii* and *P. patens*, respectively. Among the algae, we identified one sequence in each genome of *C. braunii* and *O. lucimarinus*. None were detected in *C. reinhardtii*, *C. variabilis* and *V. cariabilis*. By quantitatively comparing the results, we found a trend of NLP family size amplification from lower to higher plants. For instance, the number of NLPs in algae (*C. braunii*, *C. reinhardtii*, *C. variabilis*, *V. cariabilis*, and *O. lucimarinus*) ranged from zero to one, while lower plants (*S. moellendorfii*) had two NLPs, basal angiosperms (*A. trichopoda*) had three NLPs, and higher plants (*A. thaliana*, maize, sorghum, foxtail millet, and rice) had five to nine NLPs. Notably, the lower plant *P. patens* had more NLPs (eight), and maize had the most NLPs (nine) in the Graminaceae family.

3.2 Two-dimensional and three-dimensional structure of NLP family in foxtail millet

We predicted the two-dimensional and three-dimensional structure of the NLP family in foxtail millet. According to the prediction analysis by the online tool SOPMA, all seven foxtail millet proteins had extremely similar percentages of two-dimensional structures (Table 1). Among them, α-helix and random coil were the dominant structures, with random coil accounting for the largest proportion (46.73%–55.61%), followed by α-helix (25.99%–34.18%), β-sheet (11.60%–14.67%), and β-turn being the smallest (4.15%–5.19%). To better understand the three-dimensional structure of the NLP proteins in foxtail millet, we used the online tool PHYRE2 to establish a three-dimensional protein model through homology modeling (Figure 1). We observed that the three-dimensional structures of the proteins encoded by genes *Si5G004100.1*, *Si2G298700.1*, and *Si1G094300.1* were similar. The three-dimensional structures of the proteins encoded by genes *Si3G084600.1*, *Si9G553000.1*, and *Si8G074000.1* were also similar. However, the three-dimensional structure of the protein encoded by gene *Si6G248300.1* differed from that of the remaining six protein sequences, with fewer β-folded parts.

3.3 Promoter analysis of foxtail millet NLP genes

We selected the promoter sequences of seven NLP family genes and predicted the *cis*-acting elements of the 2 kb base sequence upstream of the transcription start site. The 31 *cis*-acting elements were divided into four major categories: light-responsive elements,

TABLE 1 NLP family protein two-dimensional structure prediction results of foxtail millet.

Gene	Alpha helix (%)	Extended strand (%)	Beta turn (%)	Randon coil (%)
Si5G004100.1	25.99	13.69	4.71	55.61
Si3G084600.1	27.52	13.76	4.81	53.91
Si9G553000.1	31.11	12.55	4.15	52.18
Si8G074000.1	27.58	14.19	4.77	53.46
Si2G298700.1	29.35	14.67	5.19	50.79
Si1G094300.1	34.18	14.29	4.81	46.73
Si6G248300.1	31.43	11.60	5.14	51.82

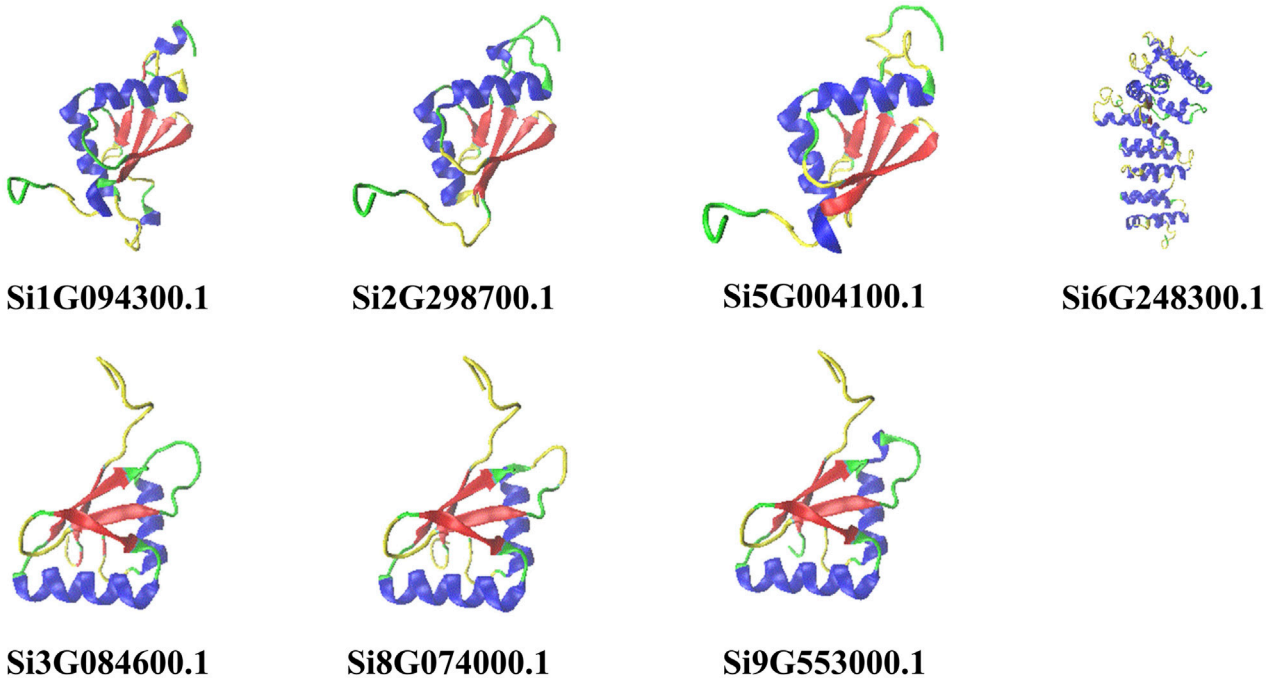

FIGURE 1
Prediction of the tertiary structure of NLP family proteins in foxtail millet. Note: Helix-blue; sheet-red; turn-yellow; coil-green.

stress-responsive elements, hormone-responsive elements, and other *cis*-acting elements (Figure 2A; Supplementary Table S1). We found a total of 196 *cis*-acting elements in the promoter sequences of the seven NLP families, including 15 light-responsive elements, six stress-responsive elements, seven hormone-responsive elements, and three other types of *cis*-acting elements. These include *cis*-acting regulatory elements associated with phloem expression, *cis*-acting regulatory elements involved in endosperm expression, and *cis*-acting regulatory elements involved in regulating maize alcohol-soluble protein metabolism. The promoter of *Si9G553000.1* was the only one containing all three *cis*-acting elements at the same time. Among the 15 light response elements, the G-box was the most abundant in the promoter of *Si9G553000.1*. Among the six stress response elements, the most abundant *cis*-acting element was ARE, which was present in all the promoters except for *Si3G084600.1*. Among the seven hormone-responsive action elements, ABRE, CGTCA-motif, and TGACG-motif were more abundant. As shown in Figure 2B, the promoter of *Si9G553000.1* contains many light-responsive action elements and hormone-responsive action elements. *Si8G074000.1* has more light-responsive action elements and stress-responsive action elements.

In the JASPAR prediction results, we found that the gene promoters (*Si3G084600.1*, *Si9G553000.1*, *Si8G074000.1*, *Si2G298700.1*, *Si1G094300.1*, and *Si6G248300.1*) all contained MBS elements with the numbers of 3, 16, 4, 1, 2, and 3, respectively. While gene *Si5G004100.1* did not contain MBS elements in its promoter. This prediction was highly consistent with the results of PlantCARE's promoter analysis (Figure 2B). The reliability of PlantCARE's promoter analysis results was also verified by the fact that all other six members of the foxtail millet NLP gene family have MBS elements, except for the absence of MBS elements in the promoter sequence of gene *Si5G004100.1* in PlantCARE's analysis.

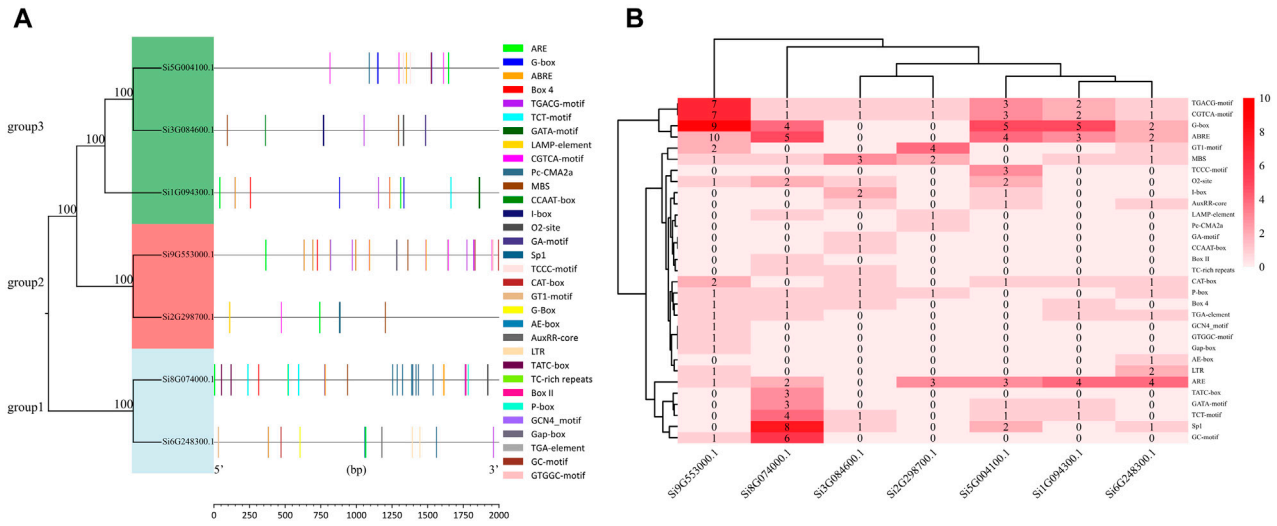

FIGURE 2
Analysis of promoter in foxtail millet NLP genes. **(A)** Distribution of promoter elements of the NLP gene family in foxtail millet. **(B)** Number of promoter elements of the NLP gene family in foxtail millet.

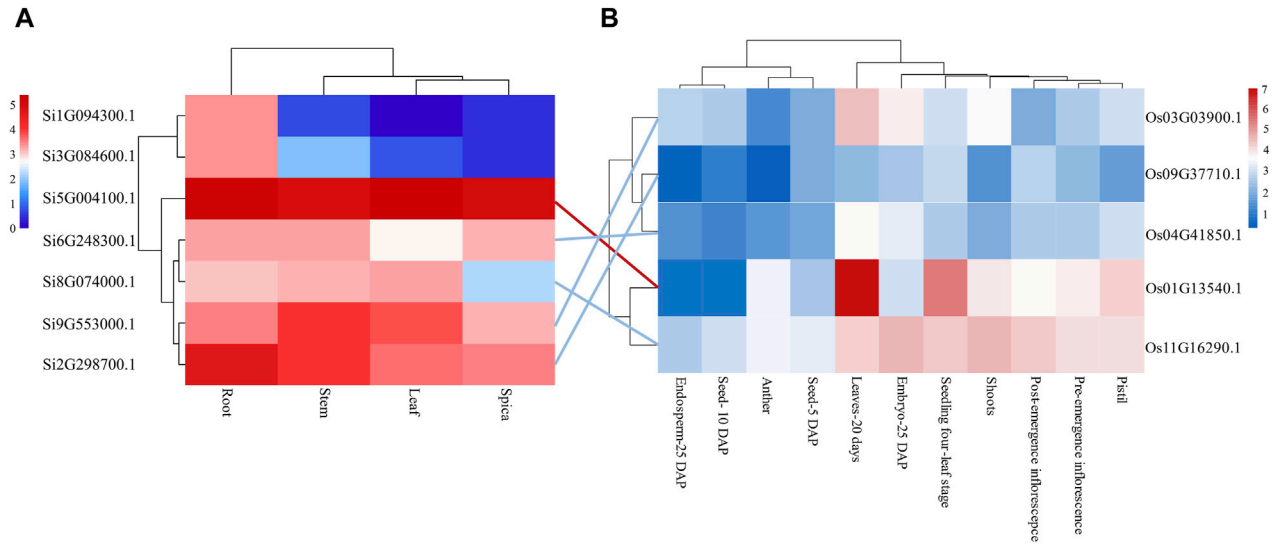

FIGURE 3
Tissue expression analysis of NLP gene family in foxtail millet and rice. **(A)** Tissue expression profile of foxtail millet NLP genes. **(B)** Tssue expression profile of rice NLP genes. Note: Lines represent orthologous gene pairs identified by OrthoMCL software. Red indicates that the expression level of the orthologous pair is higher, and blue indicates that the expression level of the orthologous pair is relatively low.

3.4 Tissue expression analysis of foxtail millet NLP genes

Analyzing gene expression helps to speculate about the function of genes. Therefore, we downloaded transcriptome data of the root, stem, leaf, and spica of foxtail millet from a public database and analyzed the tissue expression of the seven NLP genes in foxtail millet (Figure 3). The results showed that gene *Si1G094300.1* was expressed in three tissues (root, stem, and spica) but not leaf tissue. Except for gene *Si1G094300.1*, the remaining six NLP family members (*Si2G298700.1*, *Si3G084600.1*, *Si5G004100.1*, *Si6G248300.1*, *Si8G074000.1*, and *Si9G553000.1*) of foxtail millet were expressed to varying degrees in all four tissues (root, stem, leaf, and spica). Importantly, the gene *Si5G004100*.1 was highly expressed in all four tissues (root 36.30 RPKM, stem 28.65 RPKM, leaf 41.44 RPKM, spica 32.65 RPKM), indicating that the gene plays a great role in the growth and development of foxtail millet. The NLP gene family of foxtail millet exhibited tissue bias, mainly expressed in roots.

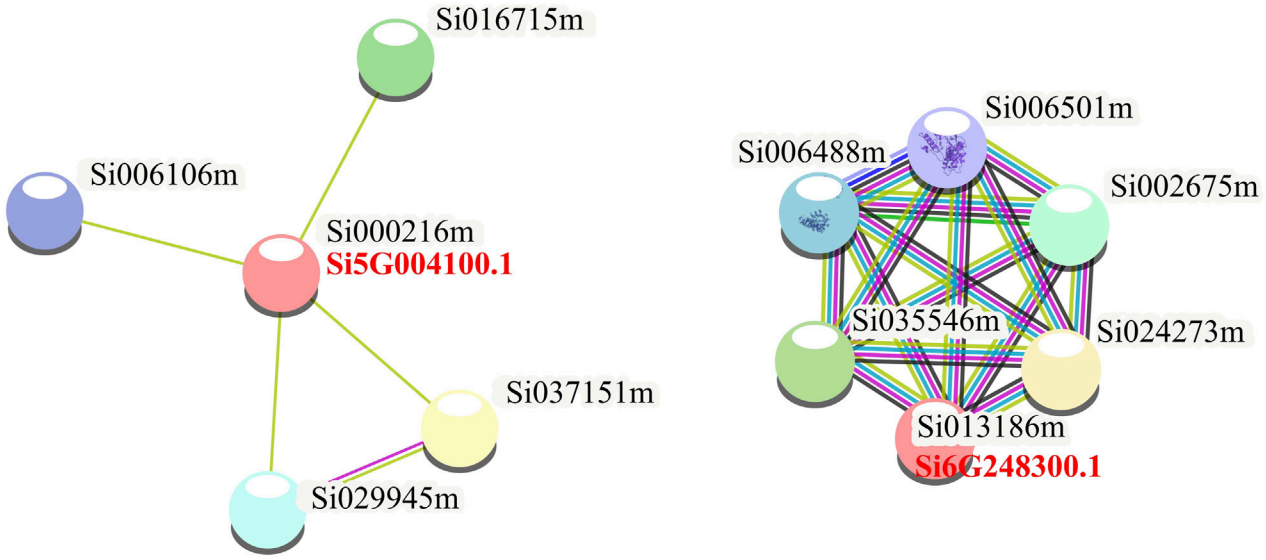

FIGURE 4
Protein interactions analysis of foxtail millet NLP gene family.

In addition, we analyzed the expression of NLP gene family in model organism rice. The NLP genes in rice were poorly expressed in Seed-5 DAP, Anther, Seed-10 DAP, and Endosperm-25 DAP but were expressed in all other tissues. However, the expression of genes *Os09G37710.1* and *Os04G41850.1* was low in each tissue. The gene *Os03G03900.1* was only weakly expressed in Leaves-20 days and Embryo-25 DAP. *Os01G13540.1* and *Os11G16290.1* were almost expressed in Pistil, Pre-emergence inflorescence, Post-emergence inflorescence, Shoots, Seedling four-leaf stage, Embryo-25 DAP, and Leaves-20 days. The gene *Os01G13540.1* had the highest expression in the leaves, no expression in the Post-emergence inflorescence, and lesser expression in the Embryo-25 DAP. There were five orthologous gene pairs in foxtail millet and rice, and the expression levels of the five orthologous gene pairs in leaves were similar. Among them, both the orthologous gene pairs *Si5G004100.1* and *Os01G13540.1* had higher expression in leaves, and both the orthologous gene pair *Si6G248300.1* and *Os04G41850.1* were not significantly expressed in leaves. However, the orthologous gene pairs *Si2G298700.1* and *Os09G37710.1* showed differential expression in leaf tissue, implying that the orthologous gene pairs underwent functional divergence after divergence between foxtail millet and rice.

3.5 Analysis of protein interactions in foxtail millet NLPs

To further explore the function of NLP in foxtail millet, we performed protein-protein interaction analysis of the proteins expressed by the seven NLP genes of foxtail millet (Figure 4). The results showed that two NLP proteins (Si5G004100.1 and Si6G248300.1) could form protein-interacting networks with other proteins. Si5G004100.1 was able to form an interaction network with four proteins (Si016715m, Si037151m, Si029945m, and Si006106m). According to the annotation results of the String database, Si5G004100.1 interacted with iron redox protein nitrite reductase (Si016715m), SPX domain-containing protein/4 (Si037151m), HTH myb domain-containing protein (Si029945m), and foxtail millet nitrite transporter (Si006106m). Moreover, we found that Si6G248300.1 could form a protein complex interaction network with five proteins (Si035546m, Si006488m, Si006501m, Si002675m, and Si024273m). According to the annotation results of the String database, Si6G248300.1 interacted with proteasome subunit α (Si002675m), PCI domain-containing protein (Si035546m), two AAA domain-containing proteins (Si006488m and Si006501m), and MPN domain-containing protein (Si024273m). In addition, we also used the String database to construct a protein interaction network diagram of the seven NLP protein families as a whole and found that all members were independent of each other, indicating that the seven NLPs of foxtail millet may play independent biological roles in foxtail millet.

3.6 Analysis of GO and KEGG of NLP genes in foxtail millet

We also performed further functional predictions on the NLP family proteins of foxtail millet (Table 2). In the GO function annotation analysis results, the enriched GO entries in the NLP family of foxtail millet belong to molecular functions, and each GO entry contained seven NLP proteins, indicating that the foxtail millet NLP family played a role in binding activity or catalytic activity. In the KEGG function annotation analysis, only the Si6G248300.1 protein participated in the biological metabolism process.

10.3389/fgene.2024.1395224

TABLE 2 Analysis of GO and KEGG of NLP genes in foxtail millet.

GO ID and pathway ID	Term type	GO Term and pathway term	The number of genes	Corrected *p*-value	Gene ID
GO:0005515	MF	protein binding	7	0.0014	//Si6G248300//Si5G004100//Si2G298700//Si3G084600//Si8G074000//Si9G553000//Si1G094300
GO:0005488	MF	binding	7	0.29	//Si6G248300//Si5G004100//Si2G298700//Si3G084600//Si8G074000//Si9G553000//Si1G094300
sita03050		proteasome	1	0.014809	Si6G248300.1

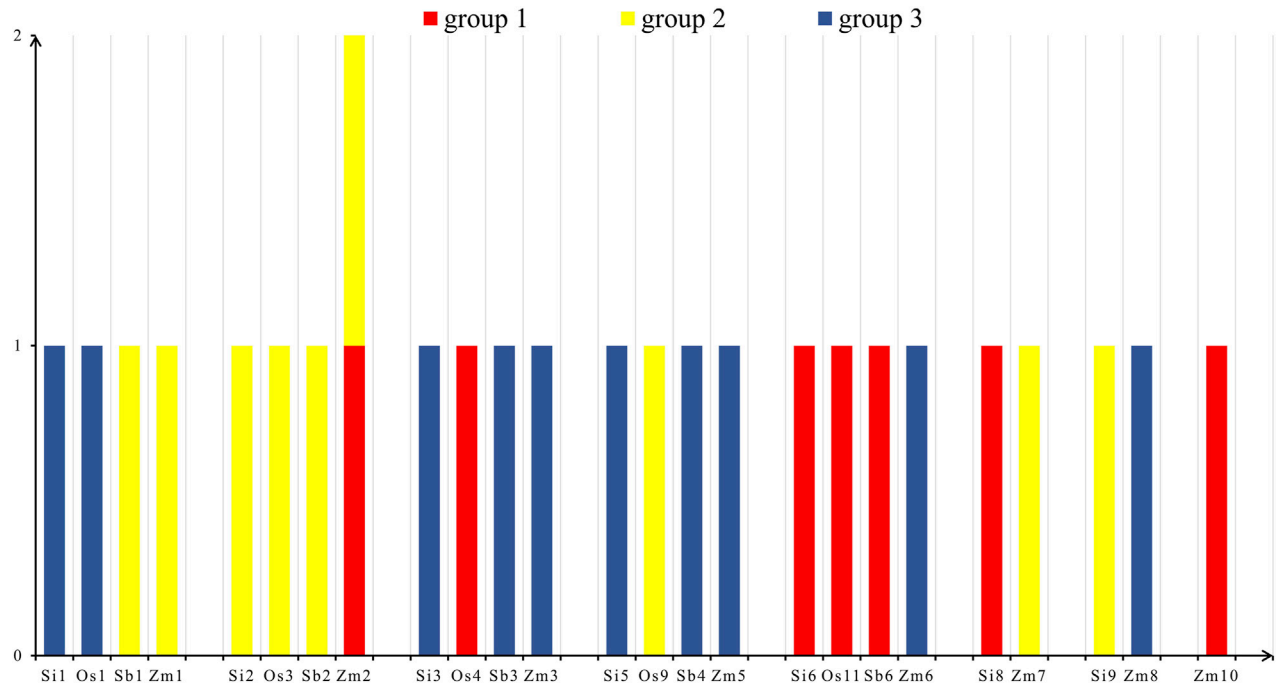

FIGURE 5
Chromosome distribution. Note: The horizontal axis is the four species of foxtail millet, rice, sorghum and maize, and the vertical axis is the number of NLP genes. Os, Si, Sb and Zm are abbreviations of rice, foxtail millet, sorghum and maize, respectively.

3.7 Chromosome distribution analysis of the NLP family in foxtail millet and closely related crops

According to the statistical results (Figure 5), we found that there were seven NLP genes in group 1, nine NLP genes in group 2, and ten NLP genes in group 3. The seven NLP genes of foxtail millet were located on chromosomes 1 (*Si1G094300.1*), 2 (*Si2G298700.1*), 3 (*Si3G084600.1*), 5 (*Si5G004100.1*), 6 (*Si6G248300.1*), 8 (*Si8G074000.1*), and 9 (*Si9G553000.1*). Meanwhile, the five NLP genes of rice are distributed on five chromosomes: chromosomes 1, 3, 4, 9, and 11, respectively. The five NLP genes of sorghum were distributed on chromosomes 1, 2, 3, 4, and 6. The nine genes of maize were distributed on chromosomes 1, 2, 3, 5, 6, 7, 8, and 10, respectively, with two NLP genes on chromosome 2. On the chromosomes of the four species, the distribution of NLP genes is relatively dispersed and there is no clustering phenomenon.

3.8 Duplication type analysis of NLP family in foxtail millet and closely related crops

We investigated the duplication types of NLP genes in four species, including rice, foxtail millet, sorghum, and maize (Table 3). We found that none of the species had singleton, proximal, or tandem duplication types for NLP genes. Dispersed duplication was the main type of NLP gene duplication in rice and sorghum. In foxtail millet, five genes underwent dispersed duplication, while two genes experienced WGD or segmental duplication. Nine NLP genes were identified in maize, of which three genes underwent dispersed duplication, and six genes experienced WGD or segmental duplication. In the collinearity analysis of the four species (Table 4), we found that the proportion of NLP genes in collinear blocks and in genome-wide collinear blocks was different. The NLP genes of rice and sorghum were not in collinear blocks, and the proportion of collinear blocks in NLP

TABLE 3 Analysis of duplication types of NLP families.

Species	Singleton		Dispersed		Proximal		Tandem		WGD or segmental		Total	
	Genome	NLP	Genome	NLP	Genome	NLP	Genome	NLP	Genome	NLP	Genome	NLP
Os	8,758	0	33,184	5	3,965	0	3,822	0	6,072	0	55,801	5
Si	5,541	0	15,806	5	2,223	0	4,585	0	6,109	2	34,264	7
Sb	7,384	0	16,178	5	2,233	0	4,165	0	5,504	0	35,464	5
Zm	13,597	0	29,377	3	2,656	0	2,458	0	11,871	6	59,959	9

TABLE 4 Whole gene and NLP collinear block analysis.

Species	All genes				NLP genes			
	Total collinear blocks	Gene number in collinear blocks	Total genes	Percentage (%)	Collinear blocks contained NLP gene	NLP genes in collinear block	Total NLP genes	Percentage (%)
Os	180	5,731	55,801	10.27	0	0	5	0.00
Si	177	5,905	34,264	17.23	1	2	7	28.57
Sb	136	5,161	35,464	14.55	0	0	5	0.00
Zm	394	11,540	59,959	19.25	3	6	9	66.67

genes of foxtail millet and maize (28.57% and 66.67%) was higher than that in genome (17.23% and 19.25%). These results provide insight into the evolution of NLP genes.

3.9 Codon bias analysis and ENC-Plot mapping of NLP families in foxtail millet and closely related crops

Studying the codon bias of gene families can provide valuable references for transgenic technology and also offer a new perspective for understanding the evolutionary history of a family. Maize, sorghum, and rice are closely related to foxtail millet and are common economic crops. To deepen our understanding of the evolution of these four species, we analyzed the codon bias of the NLP gene family in foxtail millet and its related species. We followed the literature standard (Wu et al., 2007) to reduce calculation errors and screened seven suitable CDSs of the NLP family of foxtail millet (Supplementary Table S2). Using CodonW, we obtained the results shown in Table 5, which includes the ENC used in the gene. The reference range of ENC value is 20–61, which can reflect the degree of preference for the unbalanced use of synonymous codons. The lower the ENC value, the stronger the preference. The ENC values of the NLP family genes of foxtail millet, rice, sorghum, and maize ranged from 45.55 to 58.74, 54.32 to 58.58, 40.74 to 58.6, and 41.64 to 59.39, respectively, with mean values of 55.3071, 56.512, 52.804, and 55.2756, respectively. The ENC value range and the average value of the four species were close to 60, indicating weak codon bias in the NLP family genes of these species.

The CAI measures the codon preference of a gene concerning a group of highly expressed genes. CAI values range from 0 to 1, with values closer to 1 indicating that the gene uses codons that are exclusively preferred by highly expressed genes. The CAI values of the NLP family genes in foxtail millet ranged from 0.205 to 0.258, with an average of 0.225. For rice, the CAI values of NLP family genes ranged from 0.199 to 0.233, with an average of 0.211. In sorghum, the CAI values of NLP family genes ranged from 0.211 to 0.268, with an average of 0.2372. In maize, the CAI values of NLP family genes ranged from 0.2 to 0.25, with an average of 0.2204. The codon bias index (CBI) is used to elucidate the components of all optimal codons in a particular gene. The CBI values of NLP family genes in foxtail millet ranged from −0.044 to 0.157, with an average of 0.0269. For rice, the CBI values of NLP family genes ranged from −0.005 to −0.01, with an average of −0.0182. In sorghum, the CBI values of NLP family genes ranged from −0.043 to 0.197, with an average of 0.0612. In maize, the CBI values of NLP family genes ranged from −0.042 to 0.193, with an average of 0.033. GC content of the third base of codon (GC3s) is another measure of codon preference. In monocots and dicots, the smaller the GC3s, the greater the influence of natural selection on codon preference (Gu et al., 2004).

Table 6 lists four species, each with a range of frequency (T3s) for the third base T of their synonymous codon: 0.1695–0.4161, 0.2941–0.4155, 0.1278–0.3947, and 0.1297–0.4194, respectively. The third base A also has a frequency range (A3s) in each species: 0.131–0.3407, 0.2452–0.339, 0.1045–0.348, and 0.0978–0.3599, respectively. Additionally, the third base G (G3s) has a frequency range of 0.2564–0.4344, 0.2485–0.3652, 0.2609–0.4566, and 0.2437–0.4711, respectively, while the frequency range of the third base C (C3s) is 0.2323–0.4757, 0.247–0.3347, 0.2837–0.5192, and 0.2411–0.4933, respectively. The frequency range of the third base GC (GC3s) for the four species is as

10.3389/fgene.2024.1395224

TABLE 5 Codon bias parameters.

Species	Statistic	T3s	C3s	A3s	G3s	CAI	CBI	Fop	ENC	GC3s	GC
Si	range	0.1695–0.4161	0.2323–0.4757	0.131–0.3268	0.2564–0.4344	0.205–0.258	−0.044–0.157	0.397–0.512	45.55–58.74	0.4–0.748	0.451–0.64
	average	0.3282	0.3229	0.2772	0.3163	0.225	0.0269	0.4377	55.3071	0.508	0.5057
Os	range	0.2941–0.4155	0.247–0.3347	0.2452–0.339	0.2485–0.3652	0.199–0.233	−0.005–−0.01	0.382–0.451	54.32–58.58	0.394–0.55	0.489–0.462
	average	0.3539	0.2818	0.3122	0.3017	0.211	−0.0182	0.4124	56.512	0.4598	0.482
Sb	range	0.1278–0.3947	0.2837–0.5192	0.1045–0.348	0.2609–0.4566	0.211–0.268	−0.043–0.197	0.397–0.535	40.74–58.6	0.432–0.805	0.469–0.662
	average	0.2975	0.3564	0.2644	0.325	0.2372	0.0612	0.4584	52.804	0.5442	0.5228
Zm	range	0.1297–0.4194	0.2411–0.4933	0.0978–0.3599	0.2437–0.4711	0.2–0.25	−0.042–0.193	0.397–0.53	41.64–59.39	0.373–0.806	0.433–0.676
	average	0.3129	0.3358	0.2737	0.3171	0.2204	0.033	0.4403	55.2756	0.5219	0.5122

Note: Fop, the optimal codon usage frequency.

follows: 0.4–0.748, 0.394–0.55, 0.432–0.805, and 0.373–0.806, respectively. The frequency range of the total codon GC (GC) is 0.451–0.64, 0.489–0.462, 0.469–0.662, and 0.433–0.676, respectively. Overall, there appears to be no clear base preference in the coding region of the NLP family genes of these four species, nor in the selection of bases in the third position of their codons.

The ENC-plot is a useful tool for detecting the impact of base composition on codon bias. A gene distributed along or near the standard curve suggests that the codon bias of the gene is solely influenced by mutations. Conversely, if a gene falls far from the standard curve, it indicates that the codon bias of the gene is affected by selection pressure and other factors. Based on the ENC-plot diagram for the four species (Figure 6), it can be observed that most of their genes are located close to or below the standard curve. We can draw the following inference: compared with base mutations, natural selection and other factors had a more significant effect on the codon bias of the NLP gene family.

3.10 Determination of the optimal codon of the NLP family in foxtail millet and closely related crops

RSCU is the ratio of the observed value of synonymous codons to the average expected value of synonymous codon usage in a sample, which intuitively reflects the degree of codon usage bias independent of amino acid usage and codon abundance (Sharp and Li, 1986). We selected one gene from each end (10%, rounded) and sorted it according to the ENC value from small to large, so as to obtain the genes with high expression and low expression in the NLP family of each species. The highly expressed gene of foxtail millet was *Si1G094300.1*, and the weakly expressed gene was *Si9G553000.1*. Similarly, the highly expressed gene of rice was *Os04G41850.1*, and the weakly expressed gene was *Os01G13540.1*. For sorghum, the highly expressed gene was *Sb04G038000.1*, and the weakly expressed gene was *Sb03G003700.1*. Lastly, the highly expressed gene of maize was *Zm2G042278_P01*, and the weakly expressed gene was *Zm2G475305_P01*. Preference libraries were established separately by species, and highly expressed superior codons were obtained for each species based on ΔRSCU >0.08 (27 for foxtail millet, 22 for rice, 25 for sorghum, and 28 for maize). Next, optimal codons were identified for each species based on the highest RSCU values of codons in each amino acid (Supplementary Table S3).

Foxtail millet had 11 optimal codons, with six ending in A/U (T) and five ending in G/C; rice had seven optimal codons, with four ending in A/U (T) and three ending in G/C; sorghum had 13 optimal codons, all ending in G/C; maize had 11 optimal codons, all ending in G/C. Therefore, compared to sorghum and maize, the optimal codon bias of foxtail millet and rice was weaker. Figure 7 shows that the four species shared two optimal codons (AAG and UUC). In addition, foxtail millet, sorghum, and maize shared three optimal codons (CUG, UAC, and CAG). Sorghum and maize had five identical optimal codons (AUC, AAC, GAG, UGC, and GGC). Foxtail millet and rice shared one identical optimal codons (CCA).

TABLE 6 Exon—intron structure information of NLP family.

Group	Gene	Length	Intron	CDS	UTR	Group	Gene	Length	Intron	CDS	UTR
group 3	*Sb03G003700.1*	12,323	4	5	1	group 2	*Zm2G048582_P01*	5,233	3	4	2
	Zm2G375675_P01	6,192	4	5	2		*Sb02G302500.1*	5,350	5	4	4
	Zm2G475305_P01	4,391	4	5	2		*Zm2G053298_P01*	5,333	4	4	2
	Si5G004100.1	5,891	4	5		group 1	*At3G59580.1*	3,716	5	5	3
	Os01G13540.1	5,582	4	5	2		*At2G43500.1*	4,715	7	6	4
	Si3G084600.1	6,431	4	5	2		*Amscaffold00066.150*	8,425	4	5	0
	Zm2G176655_P01	9,204	4	5	2		*Os11G16290.1*	6,997	4	5	1
	Amscaffold00058.115	6,863	3	4	0		*Si8G074000.1*	4,668	5	5	3
	At1G64530.1	3,421	5	6	2		*Os04G41850.1*	6,468	5	5	3
	At4G24020.1	4,242	4	5	2		*Si6G248300.1*	5,424	9	10	2
	Amscaffold00080.66	8,312	3	4	0		*Zm2G031398_P02*	8,724	9	10	0
	Si1G094300.1	3,192	4	5	2		*Sb06G148100.1*	5,881	4	5	2
	Sb04G038000.1	3,520	4	5	2		*Zm2G105004_P01*	5,045	4	5	2
	Zm2G042278_P01	3,286	3	4	2		*Sm172537*	3,057	3	4	1
group 2	*At2G17150.1*	4,418	4	4	3		*Cb84175.1*	17,691	1	2	2
	At4G35270.1	3,487	3	4	1		*Pp3c17_4370*	6,312	2	3	2
	At4G38340.1	3,119	3	4	0		*Pp3c17_4375*	6,068	3	3	3
	At1G76350.1	3,510	4	4	3		*Pp3c12_2070*	6,281	2	3	2
	At1G20640.1	3,659	4	4	3		*Pp3c9_14600*	6,907	4	4	3
	Sb01G552900.1	4,488	4	4	2		*Pp3c15_9180*	6,254	4	4	3
	Zm2G109509_P01	4,700	5	5	1		*Pp3c19_2670*	7,268	3	4	2
	Si9G553000.1	5,643	5	4	4		*Pp3c22_6370*	6,052	3	4	2
	Os03G03900.1	4,629	4	4	3		*Pp3c22_6360*	5,840	3	4	2
	Os09G37710.1	5,194	4	4	3		*Sm61084*	2019	3	4	0
	Si2G298700.1	5,211	4	5	2		*Ol24740*	2,172	0	1	0

3.11 Phylogenetic tree construction and selection pressure analysis of NLP family of representative species in plants

To explore the phylogenetic relationship of NLP, we aligned all NLP protein sequences of ten species (*A. thaliana*, maize, sorghum, foxtail millet, rice, *A. trichopoda*, *S. moellendorfii*, *P. patens*, *C. braunii*, and *O. lucimarinus*) containing NLP and constructed a phylogenetic tree (Figure 8). Based on previous classification criteria (Schauser et al., 2005; Hachiya et al., 2011) and the topology of the phylogenetic tree, we divided it into three groups. Group 1 contained two foxtail millet NLP genes (*Si6G248300.1* and *Si8G074000.1*), two *A. thaliana* NLP genes, one sorghum NLP gene, two maize NLP genes, two rice NLP genes, two *S. moellendorfii* NLP genes, eight *P. patens* NLP genes, one *A. trichopoda* NLP gene, one *C. braunii* NLP gene, and one *O. lucimarinus* NLP gene. Group 2 contained two foxtail millet NLP genes (*Si2G298700.1* and *Si9G553000.1*), three maize NLP genes, two sorghum NLP genes, two rice NLP genes, and five *A. thaliana* NLP genes. Group 3 contained three foxtail millet NLP genes (*Si1G094300.1*, *Si3G084600.1*, and *Si5G004100.1*), four maize NLP genes, two sorghum NLP genes, one rice NLP gene, two *A. trichopoda* NLP genes, and two *A. thaliana* NLP genes.

Based on the phylogenetic tree and sequence alignment files, we analyzed the selection pressure of representative plant species in the NLP family. The results showed that under the site model condition, no amino acid sites under positive selection pressure were detected (Supplementary Table S4). In the branch model, we labeled four foreground branches for the double-ratio model based on the results of the free-ratio model. However, the results did not identify any branches that were subjected to positive selection pressure.

3.12 Exon-intron analysis of NLP families of representative species in plants

We analyzed the differences in gene structure among ten species and used CFVisual software to determine the number of NLP gene

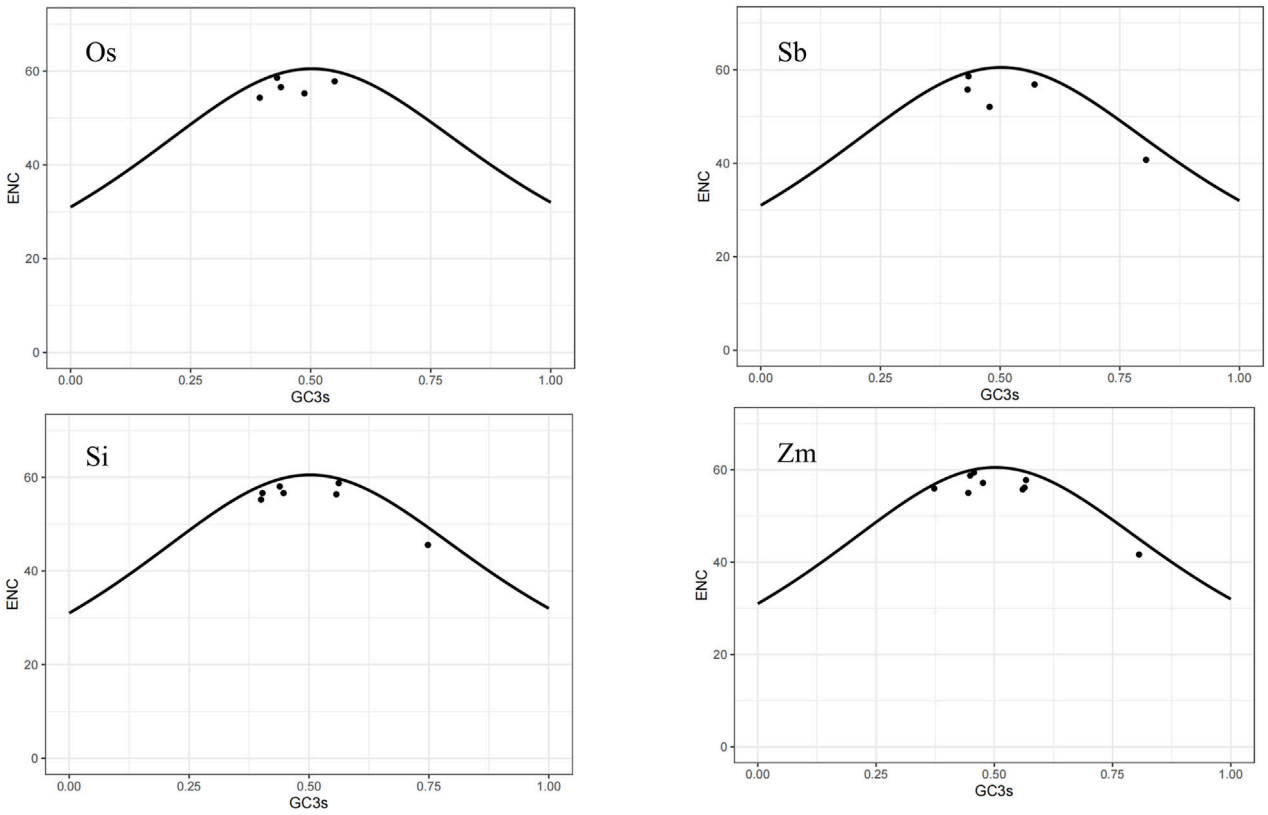

FIGURE 6
ENC-plot diagram.

structural features for each species (Table 6; Supplementary Figure S1). Among the five NLP sequences of foxtail millet, all except for *Si8G074000.1* and *Si9G553000.1* had one untranslated region (UTR) sequence at both ends. In group 1, *Si6G248300.1* had ten CDS and nine intron sequences, while *Si8G074000.1* had five CDS and five intron sequences. In group 2, *Si2G298700.1* had five CDS and four intron sequences, while *Si9G553000.1* had four CDS and five intron sequences. The three NLP sequences in group 3 contained five CDS and four intron sequences.

Among the nine NLP sequences in *A. thaliana*, *At4G38340.1* lacked a UTR sequence, *At4G35270.1* had one UTR sequence, *At1G64530.1* and *At4G24020.1* both had two UTR sequences, *At2G43500.1* had four UTR sequences, and the remaining four NLP sequences had three UTR sequences. In group 1, *At3G59580.1* had five CDS and five intron sequences, while *At2G43500.1* had six CDS and seven intron sequences. In group 2, *At4G35270.1* and *At4G38340.1* had four CDS and three intron sequences, while *At2G17150.1*, *At1G76350.1*, and *At1G20640.1* had four CDS and four intron sequences. In group 3, *At1G64530.1* had six CDS and five intron sequences, and *At4G24020.1* had five CDS and four intron sequences.

Among the nine NLP sequences of maize, *Zm2G031398_P02* lacked a UTR sequence, and *Zm2G109509_P01* had one UTR sequence, while the remaining seven NLP sequences had two UTR sequences. In group 1, *Zm2G105004_P01* had five CDS and four intron sequences, and *Zm2G031398_P02* had one CDS and nine intron sequences. In group 2, *Zm2G109509_P01* had five CDS and five intron sequences, while *Zm2G048582_P01* and *Zm2G053298_P01* had four CDS and three or four intron sequences. In group 3, except for *Zm2G042278_P01*, which had four CDS and three intron sequences, the remaining three NLP sequences had five CDS and four intron sequences.

Among the five NLP sequences of sorghum, *Sb03G003700.1* had one UTR sequence, and *Sb02G302500.1* had four UTR sequences, while the other three NLP sequences had two UTR sequences. In group 1, *Sb06G148100.1* had five CDS and four intron sequences; *Sb02G302500.1* had four CDS and five intron sequences, and *Sb01G552900.1* had four CDS and four intron sequences. In group 3, both *Sb03G003700.1* and *Sb04G038000.1* had five CDS and four intron sequences.

Among the five NLP sequences in rice, *Os11G16290.1* had one UTR sequence, and *Os01G13540.1* had two UTR sequences, while the remaining three NLP sequences had three UTR sequences. In group 1, *Os04G41850.1* had five CDS and five intron sequences, and *Os11G16290.1* had five CDS and four intron sequences. In group 2, *Os03G03900.1* and *Os09G37710.1* had four CDS and four intron sequences. In group 3, *Os01G13540.1* had five CDS and four intron sequences.

The three NLP sequences of *A. trichopoda* lacked a UTR sequence. In group 1, *Amscaffold00066.150* had five CDS and four intron sequences, while in group 3, *Amscaffold00080.66* and *Amscaffold00058.115* had four CDS and three intron sequences.

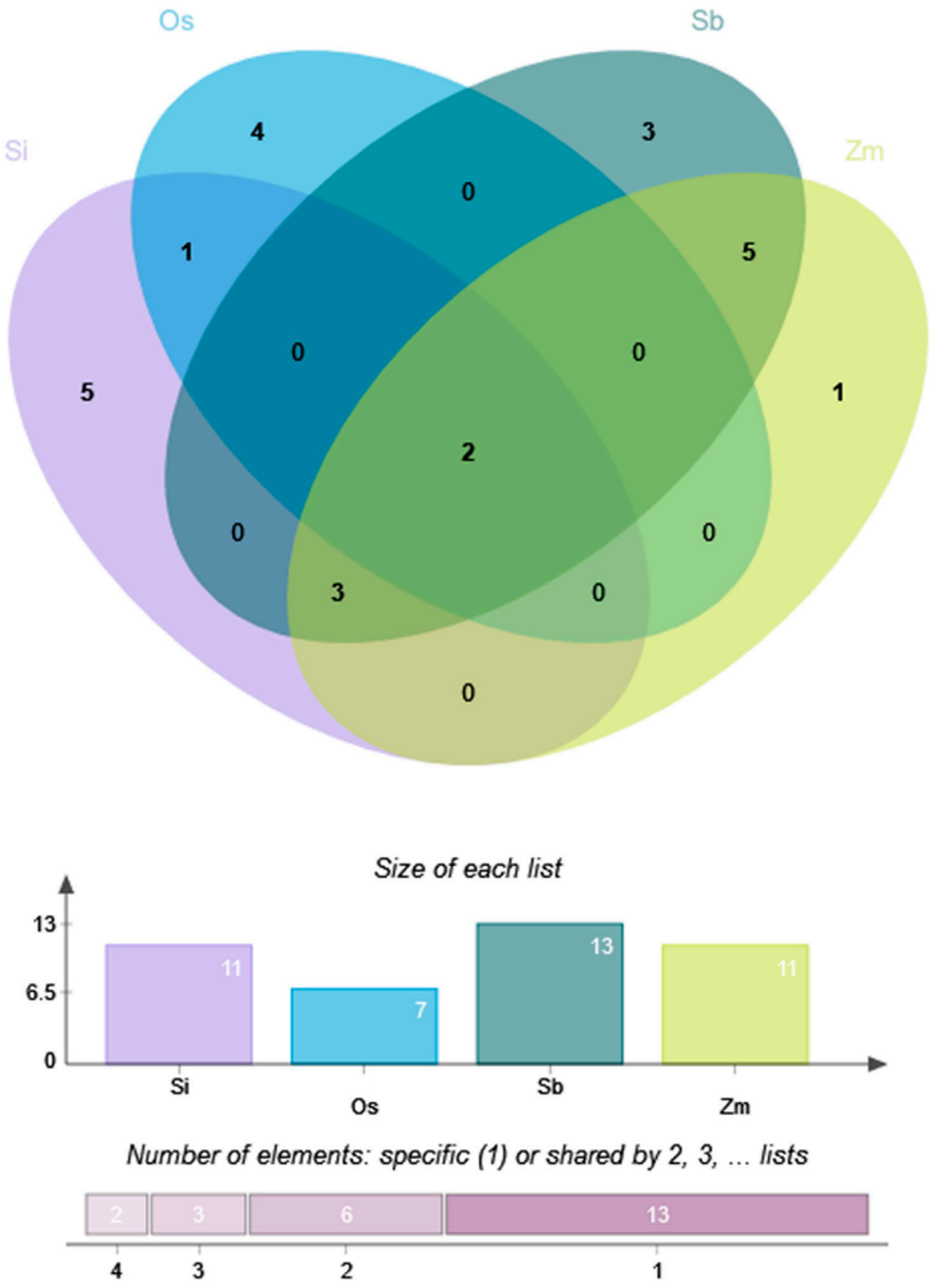

FIGURE 7
Optimal codon Venn diagram of NLP family in foxtail millet and closely related crops. Note: Os, Si, Sb and Zm are abbreviations of rice, foxtail millet, sorghum and maize, respectively.

Both of the NLP sequences in *S. moellendorfii*, *Sm61084* and *Sm172537*, belong to group 1. *Sm61084* lacked a UTR sequence, while *Sm172537* had only one UTR sequence. Both sequences contained four CDS and three intron sequences.

Among the eight NLP sequences of *P. patens*, *Pp3c17_4375*, *Pp3c9_14600*, and *Pp3c15_9180* had three UTR sequences, while the remaining five NLP sequences had two UTR sequences. All eight NLP sequences were in group 1. *Pp3c17_4375* had three CDS and three intron sequences, *Pp3c17_4370* and *Pp3c12_2070* had three CDS and two intron sequences. *Pp3c9_14600* and *Pp3c15_9180* had four CDS and four intron sequences, and the other three NLP sequences had four CDS and three intron sequences.

C. braunii had only one NLP sequence, namely, *Cb84175.1*, which had two UTR sequences. It was located in group 1 and had two CDS and one intron sequence. *O. lucimarinus* had one NLP sequence, *Ol24740*, which was located in group 1 and had only one CDS with no UTR or intron sequences.

Among the researched species, most NLP gene families had a UTR sequence at each end of the sequence, and most gene families

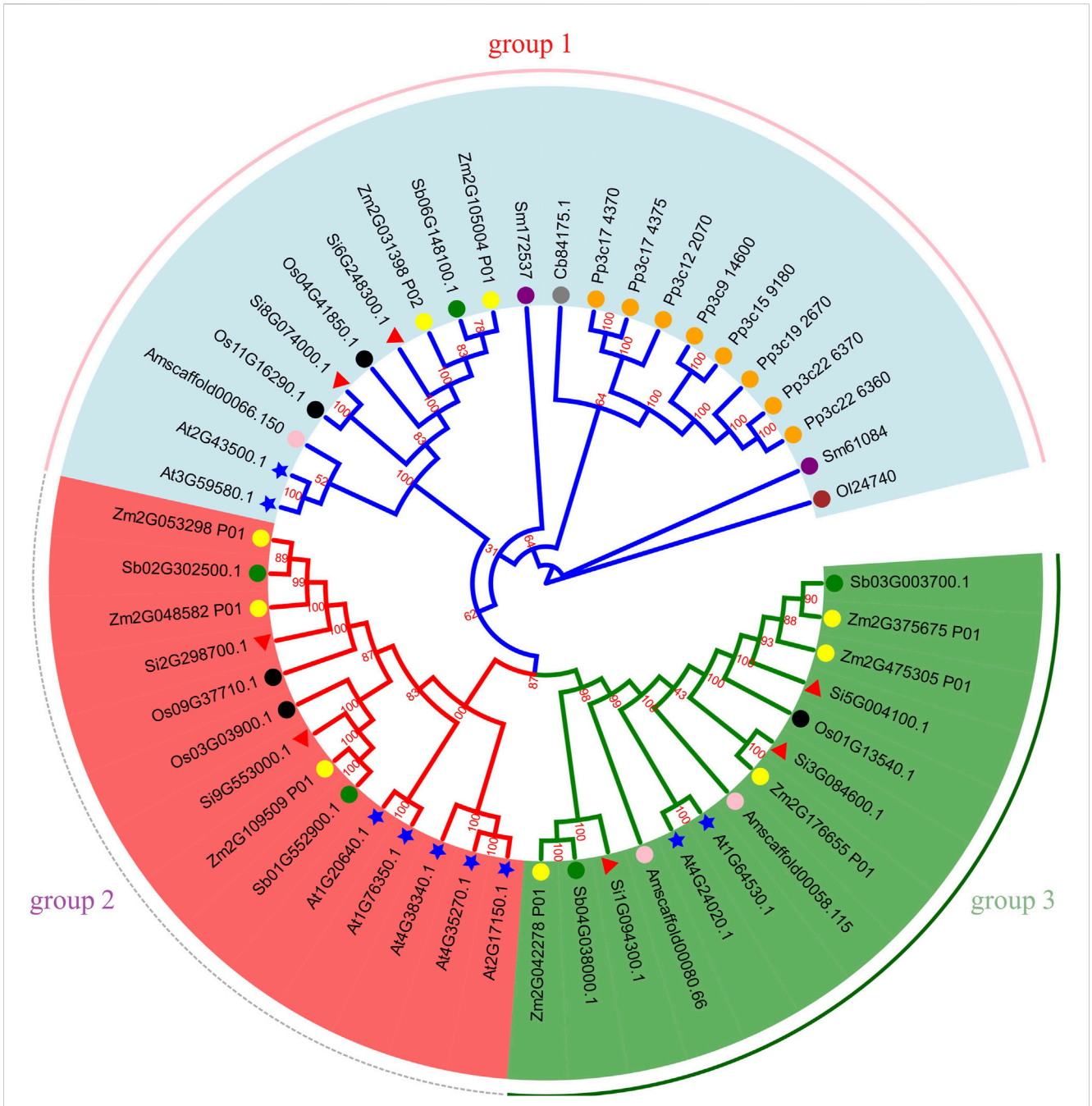

FIGURE 8
Phylogenetic tree of NLP gene family. Note: Gene IDs show their respective sources: Os represents rice, Si represents foxtail millet, Sb represents sorghum, Zm represents maize, Am represents *A. trichopoda*, At represents *A. thaliana*, Sm represents *S. moellendorfii*, Cb represents *C. braunii*, Pp represents *P. patens*, Ol represents *O. lucimarinus*. We use shape and color to distinguish different species. Green circles, yellow circles, black circles, pink circles, blue stars, orange circles, purple circles, brown circles and red triangles represent sorghum, maize, rice, *A. trichopoda*, *A. thaliana*, *P. patens*, *S. moellendorfii*, *O. lucimarinus* and foxtail millet, respectively. The number on the branch is the bootstrap value.

had one more intron sequence than CDS. The UTR at the 5′end of *Sb06G148100.1* was the longest (1,149 bp), while *At2G43500.1* had the shortest UTR (21 bp). The UTR at the 3′end of *Pp3c19_2670* was the longest (1972 bp), and *At3G59580.1* had the shortest UTR (45 bp). The CDS of the *Cb84175.1* gene of *C. braunii* was the longest (3,285 bp), and the intron inserted into the *Cb84175.1* sequence was the longest (11,355 bp) (Table 6).

3.13 Analysis of conserved motifs and structural domains of NLP families of representative species in plants

We explored the differences in the protein structure of the NLP gene family, specifically its motifs and structural domains, among different plant species (Supplementary Figure S2). Regarding

conserved motifs, the types and positions of conserved motifs in the NLP gene family of higher plants (foxtail millet, sorghum, rice, maize, and *A. thaliana*) were mostly similar, but there were some differences. For example, in group 1, the At2G43500.1 sequence lacked motif 1, and in group 2, At2G17150.1 lacked motif 6, At4G38340.1 lacked motif 12 and motif 6. Compared to most NLP gene families, the sequence Sb02G302500.1 lacked multiple conserved motifs such as motif 14, motif 12, motif 13, motif 3, and motif 6, and the positions of the conserved motifs were also different; motif 2 was located between motif 8 and motif 10, not before motif 7. In group 3, the sequence Si3G084600.1 and Zm2G176655_P01 lacked motif 10, the sequence Amscaffold00080.66 lacked motif 11 and motif 4, and the sequence Si1G094300.1, Sb04G038000.1, and Zm2G042278_P01 lacked both motif 6 and motif 10. Additionally, Zm2G042278_P01 lacked motif 3. In group 3 of lower plants (*A. trichopoda*, *S. moellendorfii*, *P. patens*, *C. reinhardtii*, *C. variabilis*, *V. carteri*, and *O. lucimarinus*), the conservative motif of *O. lucimarinus* lacked several conserved motifs such as motif 12, motif 7, motif 5, motif 8, motif 3, motif 6, motif 10, motif 9, and motif 15, and the position of these motifs had changed significantly. The types and positions of the conserved motifs of the NLP gene family sequences in lower plants were generally the same and consistent with most higher plants. From the perspective of structural domains, the members of the NLP gene family all contained two structural domains, RWP-RK and PB1. The RWP-RK domain consisted of motif 11, motif 15, and motif 4, while the PB1 domain contained only one conserved group of motif 1. This is an important sequence feature of the NLP gene family. The positions of RWP-RK and PB1 domains were relatively fixed. The PB1 domain was located at the *C*-terminus, and its right side was the RWP-RK domain. In particular, the Ol24740 sequence in group 1 had two PB1 domains, with the left PB1 domain containing only motif 4 and the right PB1 domain containing only motif 11. The Si6G248300.1 sequence in group 1 had an RPN1_RPN2_N domain located at the *N*-terminus and did not contain a conserved motif. The RWP-RK domain of At2G43500.1 had no motif 1. In group 3, the PB1 domain of Amscaffold00080.66 was incomplete, with only one conserved motif 15 detected, while motif 11 and motif four were missing.

Overall, the conserved motif 2 was found to be present in the protein sequences of various NLP gene family research species, and it may serve as a characteristic sequence of the functional unit of the NLP gene family.

3.14 Homology analysis of NLP families of representative species in plants

We constructed orthologous network relationships for the NLP families of the ten studied species and analyzed the number of paralogous gene pairs for each species using Excel statistics (Figure 9). In the orthologous network diagram, we observed nine pairs of orthologous genes between foxtail millet and maize, indicating that they had the most orthologous gene pairs compared to other species. On the other hand, maize, sorghum, and foxtail millet had relatively few orthologous gene pairs, each with only five pairs. Additionally, there were no orthologous gene pairs between foxtail millet and the remaining nine researched species (*A. thaliana*, *A. trichopoda*, *S. moellendorfii*, *P. patens*, *C. braunii*, *C. reinhardtii*, *C. variabilis*, *V. carteri*, and *O. lucimarinus*). We also analyzed the paralogous radar chart and found that *P. patens* had the most paralogous gene pairs with 11 pairs, followed by *A. thaliana* with five pairs and maize with three pairs. The remaining ten researched species did not have a pair of paralogous genes. This suggests that each of these species experienced a different scale of gene duplication events after differentiation. In ancient species, *P. patens* had the most paralogous gene pairs.

3.15 Expansion analysis of the NLP family of representative species in plants

We analyzed the duplication and loss of NLP genes during evolution in the ten studied species (Figure 10). Our analysis revealed that the number of genes in the NLP family tended to expand during evolution. Initially, we calculated that the number of ancestral NLP genes in the representative species was one. The common ancestor of the representative species did not experience gene duplication and loss (0:0), and the same was true for the algae (0:0), where no duplication and loss of NLP genes occurred in both *C. braunii* and *O. lucimarinus*, with only one NLP gene present in both. Subsequently, three NLP genes were duplicated, and the common ancestor of *A. thaliana*, maize, sorghum, foxtail millet, and rice, *A. trichopoda*, *S. moellendorfii*, and *P. patens* possessed four NLP genes (3:0). After a WGD (Whole Genome Duplication) event (θ), six NLP genes were duplicated, and two NLP genes were lost (6:2), resulting in the detection of eight NLP genes in *P. patens*. Subsequently, the common ancestor of *A. thaliana*, maize, sorghum, foxtail millet, and rice, *A. trichopoda*, and *S. moellendorfii* experienced the loss of one NLP gene, creating a quantitative size of three NLP gene families. After two WGD events (ε and ζ), six NLP genes were replicated, and two NLP genes were lost, resulting in the expansion of the ancestral NLP genes in the seven angiosperms. *S. moellendorfii* experienced a loss of one NLP gene, resulting in the currently observed number of two NLP genes. For the basal angiosperm *A. trichopoda*, four NLP genes were lost, resulting in the currently observed size of three NLP gene numbers. However, ancestral species experienced one NLP gene duplication and one NLP gene loss (1:1), resulting in a sizeable number of seven NLP genes in the monocotyledonous and dicotyledonous common ancestor, with a WGT (Whole Genome Triplication, γ) and two WGD (β and α) events, five NLP genes were duplicated, and three NLP genes were lost (5:3), resulting in the currently observed NLP gene population size of nine NLP genes in *A. thaliana*. After three WGD (τ, σ, and ρ) events, three NLP genes were duplicated, and three NLP genes were lost (3:3). The number of NLP gene families in the monocotyledonous ancestor remained at seven. Subsequently, it experienced the loss of two NLP genes, resulting in the currently observed size of five NLP gene numbers in rice. The number of NLP genes in the common ancestor of maize, sorghum, and foxtail millet was maintained at seven quantitative scales of foxtail millet, with no duplication and loss occurring. However, two NLP genes were duplicated, and one NLP gene was lost (2:1), forming the number of eight NLP genes in the ancestors of maize, sorghum. With the occurrence of a WGD event (θ) in maize, after a single NLP gene duplication, the expansion was made to the currently observed size of nine NLP gene numbers in maize. In contrast, sorghum experienced the loss of three NLP genes, resulting in the currently observed number of five NLP genes. In summary, we were surprised to find an

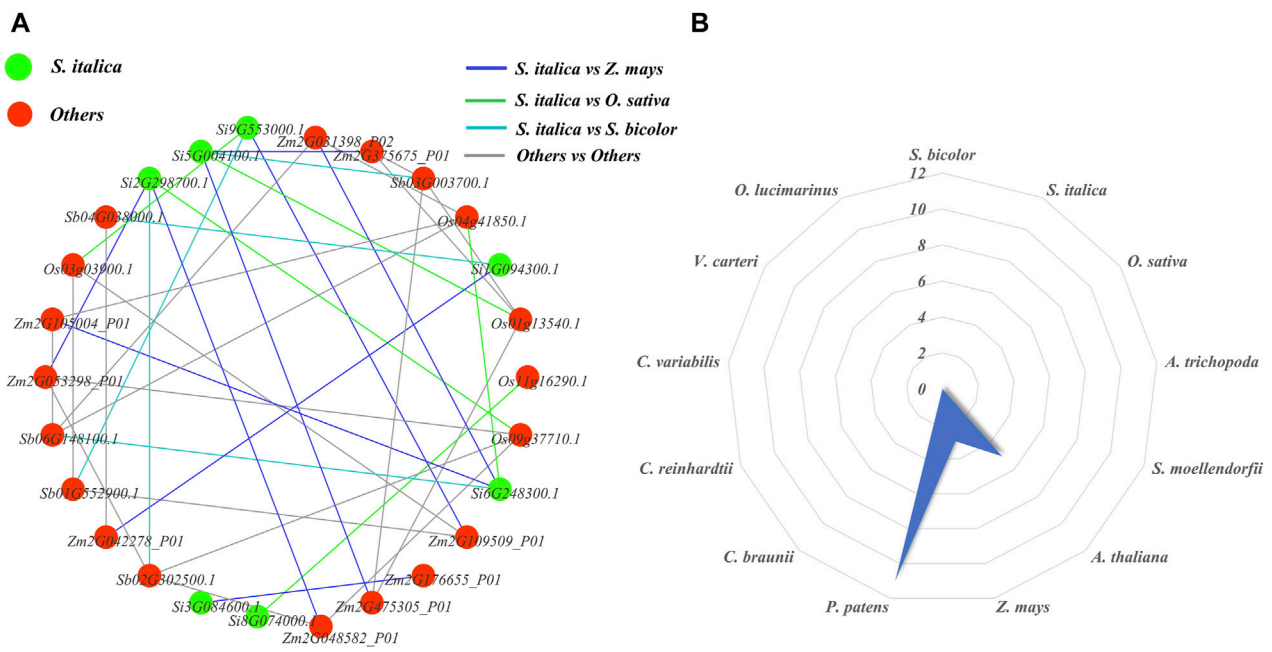

FIGURE 9
NLP family homology map and radar map. **(A)** Orthologous networks of NLP gene families in representative plants. **(B)** Paralogous homology radargram of NLP gene families in representative plants.

evolutionary trend of gradual expansion of the NLP gene family as a whole, with more duplications than losses throughout the expansion journey.

4 Discussion

Nitrogen is a major nutrient, essential to the survival of all plants (Liu et al., 2018). Therefore, nitrogen uptake and utilization directly affects plant growth and development as well as crop yields. Among the complex pathways regulating nitrate metabolism and assimilation in plants, the NLP protein family is a key component that can improve the efficiency of nitrogen utilization in plants (Castaings et al., 2009). Therefore, studying NLP family characterization and its molecular evolution in crops can provide a valuable clue to crop breeding and the evolutionary history of the family. Here, we comprehensively characterized the NLP family in foxtail millet, a C4 model crop that has not yet been resolved, then identified and investigated its molecular evolutionary trajectory by selecting representative species from each plant taxa. We obtained plenty of interesting and novel results that can be used as a resource and reference for studying NLP genes.

4.1 Molecular characterization and two significantly important genes of the NLP gene family in foxtail millet

We identified seven NLP genes in the foxtail millet genome, all of which were individually and separately distributed on different chromosomes. Such a small number of families implies that there may be no functional redundancy within the family members. Moreover, they were not structurally identical to each other, suggesting that these seven genes assume incompletely aligned functional roles. To the best of our knowledge, it has been shown in many studies with a large number of gene families that too many members may have functional redundancy (Wang et al., 2011; Wang et al., 2013). Thus, the small number of family sizes reflects the importance of each NLP gene. Through our analysis, we found that the foxtail millet NLP gene promoter contains a large number of light-responsive elements and hormone-responsive elements, which can enable foxtail millet to use light and related hormones for nitrogen assimilation when absorbing nitrogen. Furthermore, RNA-seq data showed that these seven NLP genes were indeed expressed in different tissues of foxtail millet and mainly in the root, which supports the known biological function of NLP genes, i.e., plants enhance nitrogen absorption and assimilation from the environment through the root via NLP proteins (Liu et al., 2022).

As the analysis deepened, we unearthed two genes that were extremely important for foxtail millet: *Si5G004100.1* and *Si6G248300.1*. The reason is as follows: we found that among the seven NLP proteins in foxtail millet, only the proteins expressed by these two genes were able to form protein-interaction networks with other proteins. For *Si5G004100.1*, only this gene maintained a stable and high expression in all four tissues (root, stem, leaves, and spica) of foxtail millet, suggesting that it plays a crucial role in the growth and development of foxtail millet. With regard to *Si6G248300.1*, its three-dimensional structure was the most unique among the seven NLP proteins, and unlike the other six members, it have almost no β-folding. The unique structure hints at its distinctive function (Ge et al., 2022). Also, only *Si6G248300.1* was involved in biometabolic processes

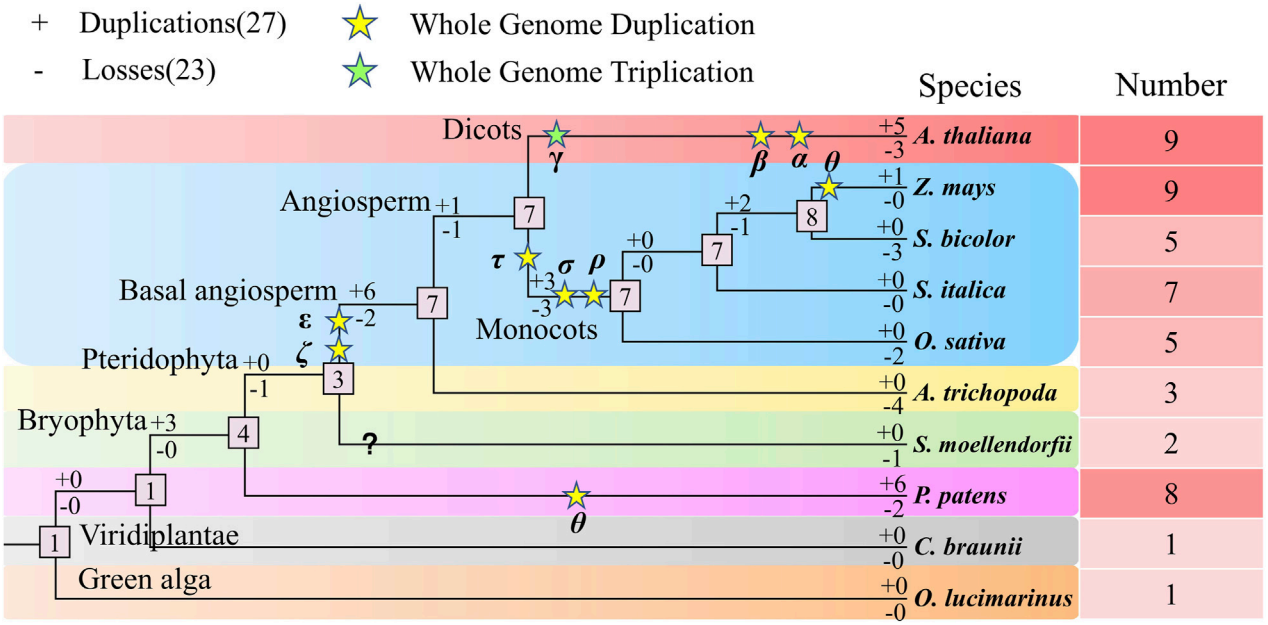

FIGURE 10
Polyploidy and amplification analysis of the NLP family in plants.

and had the most complex and robust protein interaction network. Thus, the above general phenomena reveal the specificity and importance of these two genes. Moreover, information from the String database supported by experimental evidence showed that the protein encoded by *Si5G004100.1* interacts with the iron redox protein nitrite reductase, indicating its partial function in the nitrate signaling pathway in foxtail millet. On the other hand, Si6G248300.1 was found to interact with the proteasome subunit alpha type of the T1A family of peptidases, a multicatalytic protease complex that cleaves polypeptides with arginine, phenylalanine, tyrosine, leucine, and glutamate residues under neutral or slightly alkaline pH conditions (Bochtler et al., 1999). These results could explain to some extent how these two important genes perform their biological functions.

4.2 Molecular evolutionary studies reveal the origin of the foxtail millet NLP genes and the expansion of NLP genes in plants

From a duplication perspective, gene production can be traced to a variety of duplication mechanisms, such as WGD and tandem duplication (Chen et al., 2023; Feng et al., 2024). By duplication origin analysis, we found that the seven NLP genes in foxtail millet were derived from dispersed duplication (*Si1G094300.1*, *Si2G298700.1*, *Si6G248300.1*, *Si8G074000.1*, and *Si9G553000.1*) and WGD or segmental duplication (*Si3G084600.1* and *Si5G004100.1*). Deeper comparative analyses showed that the NLP family genes of sorghum and rice, close relatives of foxtail millet, were all derived from dispersed duplication, whereas the maize NLP family genes were more derived from WGD or segmental duplication (six out of nine NLP genes), which can be explained by the fact that maize underwent another separate WGD event (θ) after species formation (Wang et al., 2015). Our previous study showed that almost all NLP gene families in Brassica spp originated from WGD or segmental duplication (Chen et al., 2022a). Thus, our results suggest that NLP gene duplication origins differ significantly across plant taxa. The identification of orthologous gene pairs can help determine the origin of genes across species. The results of our analysis showed that the foxtail millet NLP family genes could form clear orthologous gene pairs with NLP family members in closely related species (maize, rice, and sorghum), although expression differences in the homologous genes indicated that their functions had diverged (Figure 3). However, it was not possible to form orthologs with NLPs from species represented in other species taxa, indicating that the foxtail millet NLP gene family may have originated from the common ancestor of monocots.

In addition to gene duplication, the evolution of gene families can be driven by a combination of factors such as base mutations and natural selection (Chen et al., 2023). Codon usage biases have been hypothesized to potentially contribute to gene evolution (Shenton et al., 2006), so we performed a comprehensive and detailed codon bias analysis. Not only were the optimal codons for NLP family of each species identified for researchers to choose from, but factors such as base mutations and natural selection were found to contribute to the evolution of NLP gene families to varying degrees. Compared with base mutations, natural selection and other factors had a more significant effect on the codon bias of the NLP gene family. Moreover, we did not detect a positive selection branch in the selection pressure analysis, suggesting that the NLP gene may be subject to purifying selection.

More deeply, we explored the expansion of the NLP gene family across the plant kingdom. We were surprised to find an evolutionary trend of gradual expansion of the NLP gene family as a whole, with more duplications than losses throughout the expansion journey. This is because all other gene families that we know about have more losses than duplications (Song et al., 2018; Chen and Ge, 2022; Ge et al., 2022; Yu et al., 2022; Zuo et al., 2022). This reflects the

evolutionary particularity of NLP genes, reflects the plant's demand for NLP genes, and also reflects the functional importance of NLP genes. Combined with the results of paralogous homology analyses, we also found the phenomenon and possible reasons for the large size of NLP gene families in the lower plant *P. patens*, i.e., each of these species experienced a different scale of gene duplication events after differentiation, whereas *P. patens* experienced the greatest number of independent gene duplications. In addition, we have also deeply compared and elucidated the differences and variations in NLP gene structure and motif sequence features of different plant taxa. Hence, we comprehensively analyzed the molecular origin of NLP genes in foxtail millet and discovered the expansion of NLP gene families in plants.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

HC: Conceptualization, Data curation, Methodology, Software, Supervision, Writing–original draft, Writing–review and editing. FL: Data curation, Formal Analysis, Methodology, Project administration, Software, Writing–original draft, Writing–review and editing. JC: Investigation, Methodology, Project administration, Resources, Software, Writing–review and editing. KJ: Investigation, Methodology, Software, Validation, Writing–review and editing. YC: Investigation, Project administration, Resources, Software, Supervision, Writing–review and editing. WG: Data curation, Formal Analysis, Investigation, Resources, Writing–review and editing. ZW: Data curation, Funding acquisition, Investigation, Methodology, Software, Writing–review and editing, Formal Analysis.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported by the National Natural Science Foundation of China (32000405 to ZW), Tangshan Science and Technology Program Project (21130228C to ZW).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1395224/full#supplementary-material

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215 (3), 403–410. doi:10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids Res.* 25 (17), 3389–3402. doi:10.1093/nar/25.17.3389

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids Res.* 37 (2), W202–W208. doi:10.1093/nar/gkp335

Banks, J. A., Nishiyama, T., Hasebe, M., Bowman, J. L., Gribskov, M., DePamphilis, C., et al. (2011). The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *science* 332 (6032), 960–963. doi:10.1126/science.1203810

Bennetzen, J. L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A. C., et al. (2012). Reference genome sequence of the model plant Setaria. *Nat. Biotechnol.* 30 (6), 555–561. doi:10.1038/nbt.2196

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., et al. (2015). The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *genesis* 53 (8), 474–485. doi:10.1002/dvg.22877

Bochtler, M., Ditzel, L., Groll, M., Hartmann, C., and Huber, R. (1999). The proteasome. *Annu. Rev. biophysics Biomol. Struct.* 28 (1), 295–317. doi:10.1146/annurev.biophys.28.1.295

Borisov, A. Y., Madsen, L. H., Tsyganov, V. E., Umehara, Y., Voroshilova, V. A., Batagov, A. O., et al. (2003). The Sym35 gene required for root nodule development in pea is an ortholog of Nin from Lotus japonicus. *Plant physiol.* 131 (3), 1009–1017. doi:10.1104/pp.102.016071

Bu, D., Luo, H., Huo, P., Wang, Z., Zhang, S., He, Z., et al. (2021). KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic acids Res.* 49 (W1), W317–W325. doi:10.1093/nar/gkab447

Castaings, L., Camargo, A., Pocholle, D., Gaudon, V., Texier, Y., Boutet-Mercey, S., et al. (2009). The nodule inception-like protein 7 modulates nitrate sensing and metabolism in Arabidopsis. *Plant J.* 57 (3), 426–435. doi:10.1111/j.1365-313X.2008.03695.x

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic acids Res.* 50 (D1), D165–D173. doi:10.1093/nar/gkab1113

Chardin, C., Girin, T., Roudier, F., Meyer, C., and Krapp, A. (2014). The plant RWP-RK transcription factors: key regulators of nitrogen responses and of gametophyte development. *J. Exp. Bot.* 65 (19), 5577–5587. doi:10.1093/jxb/eru261

Chen, H., and Ge, W. (2022). Identification, molecular characteristics, and evolution of GRF gene family in foxtail millet (Setaria italica L). *Front. Genet.* 12, 727674. doi:10.3389/fgene.2021.727674

Chen, H., Ji, K., Li, Y., Gao, Y., Liu, F., Cui, Y., et al. (2022a). Triplication is the main evolutionary driving force of NLP transcription factor family in Chinese cabbage and related species. *Int. J. Biol. Macromol.* 201, 492–506. doi:10.1016/j.ijbiomac.2022.01.082

Chen, H., Song, X., Shang, Q., Feng, S., and Ge, W. (2022b). CFVisual: an interactive desktop platform for drawing gene structure and protein architecture. *BMC Bioinforma.* 23 (1), 178. doi:10.1186/s12859-022-04707-w

Chen, H., Zhang, Y., and Feng, S. (2023). Whole-genome and dispersed duplication, including transposed duplication, jointly advance the evolution of TLP genes in seven

representative Poaceae lineages. *BMC genomics* 24 (1), 290. doi:10.1186/s12864-023-09389-z

Chen, K., Durand, D., and Farach-Colton, M. (2000). NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* 7 (3-4), 429–447. doi:10.1089/106652700750050871

Chen, K.-E., Chen, H.-Y., Tseng, C.-S., and Tsay, Y.-F. (2020). Improving nitrogen use efficiency by manipulating nitrate remobilization in plants. *Nat. plants* 6 (9), 1126–1135. doi:10.1038/s41477-020-00758-0

Chen, T., Zhang, H., Liu, Y., Liu, Y.-X., and Huang, L. (2021). EVenn: easy to create repeatable and editable Venn diagrams and Venn networks online. *J. Genet. genomics= Yi chuan xue bao* 48 (9), 863–866. doi:10.1016/j.jgg.2021.07.007

Crawford, N. M. (1995). Nitrate: nutrient and signal for plant growth. *plant Cell* 7 (7), 859–868. doi:10.1105/tpc.7.7.859

Doust, A. N., Kellogg, E. A., Devos, K. M., and Bennetzen, J. L. (2009). Foxtail millet: a sequence-driven grass model system. *Plant physiol.* 149 (1), 137–141. doi:10.1104/pp.108.129627

Feng, H., Fan, X., Miller, A. J., and Xu, G. (2020). Plant nitrogen uptake and assimilation: regulation of cellular pH homeostasis. *J. Exp. Bot.* 71 (15), 4380–4392. doi:10.1093/jxb/eraa150

Feng, S., Li, N., Chen, H., Liu, Z., Li, C., Zhou, R., et al. (2024). Large-scale analysis of the ARF and Aux/IAA gene families in 406 horticultural and other plants. *Mol. Hortic.* 4 (1), 13. doi:10.1186/s43897-024-00090-7

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic acids Res.* 42 (D1), D222–D230. doi:10.1093/nar/gkt1223

Forde, B. G. (2000). Nitrate transporters in plants: structure, function and regulation. *Biochimica Biophysica Acta (BBA)-Biomembranes* 1465 (1-2), 219–235. doi:10.1016/s0005-2736(00)00140-1

Gao, F., Chen, C., Arab, D. A., Du, Z., He, Y., and Ho, S. Y. (2019). EasyCodeML: a visual tool for analysis of selection using CodeML. *Ecol. Evol.* 9 (7), 3891–3898. doi:10.1002/ece3.5015

Gao, Y., Quan, S., Lyu, B., Tian, T., Liu, Z., Nie, Z., et al. (2022). Barley transcription factor HvNLP2 mediates nitrate signaling and affects nitrogen use efficiency. *J. Exp. Bot.* 73 (3), 770–783. doi:10.1093/jxb/erab245

Ge, M., Liu, Y., Jiang, L., Wang, Y., Lv, Y., Zhou, L., et al. (2018). Genome-wide analysis of maize NLP transcription factor family revealed the roles in nitrogen response. *Plant growth Regul.* 84, 95–105. doi:10.1007/s10725-017-0324-x

Ge, W., Chen, H., Zhang, Y., Feng, S., Wang, S., Shang, Q., et al. (2022). Integrative genomics analysis of the ever-shrinking pectin methylesterase (PME) gene family in foxtail millet (Setaria italica). *Funct. Plant Biol.* 49 (10), 874–886. doi:10.1071/FP21319

Geourjon, C., and Deleage, G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics* 11 (6), 681–684. doi:10.1093/bioinformatics/11.6.681

Gu, W., Zhou, T., Ma, J., Sun, X., and Lu, Z. (2004). The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. *Biosystems* 73 (2), 89–97. doi:10.1016/j.biosystems.2003.10.001

Hachiya, T., Mizokami, Y., Miyata, K., Tholen, D., Watanabe, C. K., and Noguchi, K. (2011). Evidence for a nitrate-independent function of the nitrate sensor NRT1. 1 in *Arabidopsis thaliana*. *J. Plant Res.* 124, 425–430. doi:10.1007/s10265-010-0385-7

Hirsch, C. N., Hirsch, C. D., Brohammer, A. B., Bowman, M. J., Soifer, I., Barad, O., et al. (2016). Draft assembly of elite inbred line PH207 provides insights into genomic and transcriptome diversity in maize. *Plant Cell* 28 (11), 2700–2714. doi:10.1105/tpc.16.00353

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* 14 (1), 33–38. doi:10.1016/0263-7855(96)00018-5

Jagadhesan, B., Sathee, L., Meena, H. S., Jha, S. K., Chinnusamy, V., Kumar, A., et al. (2020). Genome wide analysis of NLP transcription factors reveals their role in nitrogen stress tolerance of rice. *Sci. Rep.* 10 (1), 9368. doi:10.1038/s41598-020-66338-6

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10 (6), 845–858. doi:10.1038/nprot.2015.053

Kumar, A., Batra, R., Gahlaut, V., Gautam, T., Kumar, S., Sharma, M., et al. (2018). Genome-wide identification and characterization of gene family for RWP-RK transcription factors in wheat (*Triticum aestivum* L). *PloS one* 13 (12), e0208409. doi:10.1371/journal.pone.0208409

Kumar, K., Muthamilarasan, M., and Prasad, M. (2013). Reference genes for quantitative real-time PCR analysis in the model plant foxtail millet (Setaria italica L.) subjected to abiotic stress conditions. *Plant Cell, Tissue Organ Cult. (PCTOC)* 115, 13–22. doi:10.1007/s11240-013-0335-x

Lang, D., Ullrich, K. K., Murat, F., Fuchs, J., Jenkins, J., Haas, F. B., et al. (2018). The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J.* 93 (3), 515–533. doi:10.1111/tpj.13801

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic acids Res.* 30 (1), 325–327. doi:10.1093/nar/30.1.325

Letunic, I., Doerks, T., and Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic acids Res.* 40 (D1), D302–D305. doi:10.1093/nar/gkr931

Li, D., Jin, Y., Lu, Q.-H., Ren, N., Wang, Y.-Q., and Li, Q.-S. (2024a). Genome-wide identification and expression analysis of NIN-like protein (NLP) genes: exploring their potential roles in nitrate response in tea plant (Camellia sinensis). *Plant Physiology Biochem.* 207, 108340. doi:10.1016/j.plaphy.2024.108340

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13 (9), 2178–2189. doi:10.1101/gr.1224503

Li, X., Gao, J., Song, J., Guo, K., Hou, S., Wang, X., et al. (2022). Multi-omics analyses of 398 foxtail millet accessions reveal genomic regions associated with domestication, metabolite traits, and anti-inflammatory effects. *Mol. Plant* 15 (8), 1367–1383. doi:10.1016/j.molp.2022.07.003

Li, Y., Feng, C., Xing, Y., Li, M., Wang, X., Du, Q., et al. (2024b). The NIN-LIKE PROTINE1 (CsNLP1) transcription factor is involved in modulating the nitrate response in cucumber seedlings. *Environ. Exp. Bot.* 217, 105581. doi:10.1016/j.envexpbot.2023.105581

Liu, K.-H., Liu, M., Lin, Z., Wang, Z.-F., Chen, B., Liu, C., et al. (2022). NIN-like protein 7 transcription factor is a plant nitrate sensor. *Science* 377 (6613), 1419–1425. doi:10.1126/science.add1104

Liu, M., Chang, W., Fan, Y., Sun, W., Qu, C., Zhang, K., et al. (2018). Genome-wide identification and characterization of NODULE-INCEPTION-like protein (NLP) family genes in Brassica napus. *Int. J. Mol. Sci.* 19 (8), 2270. doi:10.3390/ijms19082270

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., et al. (2018). The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J.* 93 (2), 338–354. doi:10.1111/tpj.13781

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., et al. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science* 318 (5848), 245–250. doi:10.1126/science.1143609

Nishida, H., and Suzaki, T. (2018). Nitrate-mediated control of root nodule symbiosis. *Curr. Opin. plant Biol.* 44, 129–136. doi:10.1016/j.pbi.2018.04.006

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2007). The TIGR rice genome annotation resource: improvements and new features. *Nucleic acids Res.* 35 (1), D883–D887. doi:10.1093/nar/gkl976

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., et al. (2007). The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci.* 104 (18), 7705–7710. doi:10.1073/pnas.0611046104

Prochnik, S. E., Umen, J., Nedelcu, A. M., Hallmann, A., Miller, S. M., Nishii, I., et al. (2010). Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science* 329 (5988), 223–226. doi:10.1126/science.1188800

Project, A. G., Albert, V. A., Barbazuk, W. B., dePamphilis, C. W., Der, J. P., Leebens-Mack, J., et al. (2013). The Amborella genome and the evolution of flowering plants. *Science* 342 (6165), 1241089. doi:10.1126/science.1241089

Schauser, L., Roussis, A., Stiller, J., and Stougaard, J. (1999). A plant regulator controlling development of symbiotic root nodules. *Nature* 402 (6758), 191–195. doi:10.1038/46058

Schauser, L., Wieloch, W., and Stougaard, J. (2005). Evolution of NIN-like proteins in Arabidopsis, rice, and Lotus japonicus. *J. Mol. Evol.* 60, 229–237. doi:10.1007/s00239-004-0144-2

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13 (11), 2498–2504. doi:10.1101/gr.1239303

Sharma, N., and Niranjan, K. (2018). Foxtail millet: properties, processing, health benefits, and uses. *Food Rev. Int.* 34 (4), 329–363. doi:10.1080/87559129.2017.1290103

Sharp, P. M., and Li, W.-H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *J. Mol. Evol.* 24, 28–38. doi:10.1007/BF02099948

Shenton, M., Fontaine, V., Hartwell, J., Marsh, J. T., Jenkins, G. I., and Nimmo, H. G. (2006). Distinct patterns of control and expression amongst members of the PEP carboxylase kinase gene family in C4 plants. *Plant J.* 48 (1), 45–53. doi:10.1111/j.1365-313X.2006.02850.x

Song, X., Ma, X., Li, C., Hu, J., Yang, Q., Wang, T., et al. (2018). Comprehensive analyses of the BES1 gene family in Brassica napus and examination of their evolutionary pattern in representative species. *BMC genomics* 19, 346–415. doi:10.1186/s12864-018-4744-4

Sreenivasulu, N., Miranda, M., Prakash, H. S., Wobus, U., and Weschke, W. (2004). Transcriptome changes in foxtail millet genotypes at high salinity: identification and characterization of a PHGPX gene specifically up-regulated by NaCl in a salt-tolerant line. *J. plant physiology* 161 (4), 467–477. doi:10.1078/0176-1617-01112

Sumimoto, H., Kamakura, S., and Ito, T. (2007). Structure and function of the PB1 domain, a protein interaction module conserved in animals, fungi, amoebas, and plants. *Science's STKE* 2007 (401), re6. doi:10.1126/stke.4012007re6

Suzuki, W., Konishi, M., and Yanagisawa, S. (2013). The evolutionary events necessary for the emergence of symbiotic nitrogen fixation in legumes may involve a loss of nitrate responsiveness of the NIN transcription factor. *Plant Signal. Behav.* 8 (10), e25975. doi:10.4161/psb.25975

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids Res.* 49 (D1), D605–D612. doi:10.1093/nar/gkaa1074

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30 (12), 2725–2729. doi:10.1093/molbev/mst197

Tegeder, M., and Masclaux-Daubresse, C. (2018). Source and sink mechanisms of nitrogen transport and use. *New phytol.* 217 (1), 35–53. doi:10.1111/nph.14876

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., et al. (2017). agriGO v2. 0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic acids Res.* 45 (W1), W122–W129. doi:10.1093/nar/gkx382

Wang, M., Yuan, D., Gao, W., Li, Y., Tan, J., and Zhang, X. (2013). A comparative genome analysis of PME and PMEI families reveals the evolution of pectin metabolism in plant cell walls. *PloS one* 8 (8), e72082. doi:10.1371/journal.pone.0072082

Wang, X., Wang, J., Jin, D., Guo, H., Lee, T.-H., Liu, T., et al. (2015). Genome alignment spanning major Poaceae lineages reveals heterogeneous evolutionary rates and alters inferred dates for key evolutionary events. *Mol. plant* 8 (6), 885–898. doi:10.1016/j.molp.2015.04.004

Wang, Y., Tang, H., DeBarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic acids Res.* 40 (7), e49. doi:10.1093/nar/gkr1293

Wang, Y., Wang, X., Tang, H., Tan, X., Ficklin, S. P., Feltus, F. A., et al. (2011). Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. *PloS one* 6 (12), e28150. doi:10.1371/journal.pone.0028150

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene* 87 (1), 23–29. doi:10.1016/0378-1119(90)90491-9

Wu, X.-M., Wu, S.-F., Ren, D.-M., Zhu, Y.-P., and He, F.-C. (2007). The analysis method and progress in the study of codon bias. *Yi Chuan= Hered.* 29 (4), 420–426. doi:10.1360/yc-007-0420

Wu, Y., Su, S.-x., Wang, T., Peng, G.-H., He, L., Long, C., et al. (2023). Identification and expression characteristics of NLP (NIN-like protein) gene family in pepper (Capsicum annuum L). *Mol. Biol. Rep.* 50 (8), 6655–6668. doi:10.1007/s11033-023-08587-y

Yang, Z., Zhang, H., Li, X., Shen, H., Gao, J., Hou, S., et al. (2020). A mini foxtail millet with an Arabidopsis-like life cycle as a C4 model system. *Nat. plants* 6 (9), 1167–1178. doi:10.1038/s41477-020-0747-7

Yu, J., Yuan, Y., Dong, L., and Cui, G. (2023). Genome-wide investigation of NLP gene family members in alfalfa (Medicago sativa L.): evolution and expression profiles during development and stress. *BMC genomics* 24 (1), 320. doi:10.1186/s12864-023-09418-x

Yu, T., Bai, Y., Liu, Z., Wang, Z., Yang, Q., Wu, T., et al. (2022). Large-scale analyses of heat shock transcription factors and database construction based on whole-genome genes in horticultural and representative plants. *Hortic. Res.* 9, uhac035. doi:10.1093/hr/uhac035

Zhang, G., Liu, X., Quan, Z., Cheng, S., Xu, X., Pan, S., et al. (2012a). Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nat. Biotechnol.* 30 (6), 549–554. doi:10.1038/nbt.2195

Zhang, H., Gao, S., Lercher, M. J., Hu, S., and Chen, W.-H. (2012b). EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic acids Res.* 40 (W1), W569–W572. doi:10.1093/nar/gks576

Zuo, C., Zhang, L., Yan, X., Guo, X., Zhang, Q., Li, S., et al. (2022). Evolutionary analysis and functional characterization of BZR1 gene family in celery revealed their conserved roles in brassinosteroid signaling. *BMC genomics* 23 (1), 568. doi:10.1186/s12864-022-08810-3