Genome-Wide Identification of *bHLH* Gene Family and its Response to Cadmium Stress in *Populus* × *canescens*

Yuneng Yao[1, 2, 3#], Zhengquan He[1#], Xinmeng Li[1, 2, 3], Jing Xu[2, 3], Xiaojiao Han[2, 3], Hongwei Liang[1], Renying Zhuo[2, 3], Wenmin Qiu[2, 3*]

1 Key Laboratory of Three Gorges Regional Plant Genetic & Germplasm Enhancement (CTGU)/ Biotechnology Research Center, China Three Gorges University, Yichang, 443002, Hubei, China.

2 State Key Laboratory of Tree Genetic and Breeding, Chinese Academy of Forestry, Beijing, China.

3 Key Laboratory of Tree Breeding of Zhejiang Province, The Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou, 311400, P.R. China.

#These authors contributed equally to this work

*Correspondence and requests for materials should be addressed to Wenmin Qiu at the following address, phone number, and email address:

Wenmin Qiu

Address: Research Institute of Subtropical Forestry, Chinese Academy of Forestry.

Fuyang, Hangzhou 311400, China

Tel: +086-571-63311860

E-mail: qiuwm05@163.com

Abstract

Basic/helix-loop-helix (bHLH) gene family plays pivotal roles in plant development and response to various stress in plants. Emerging evidences indicate that some members of the bHLH family are responsible for Cd stress, however little is known about the *bHLH* genes in the Cd-tolerant poplar species *Populus* × *canescens*. In our study, 170 *bHLH* genes were identified from the *P. canescens* genome, which were named based on their chromosomal location and classified into 22 subfamilies. The gene structure and motif compositions were considerably conserved among the subfamilies. Cis-acting element analysis of the *PcbHLH* promoters showed that a lot of essential cis elements are present in the promoter regions associated with plant hormones, development, and stress. qPCR results indicated that high concentrations of Cd differentially regulated the expression of PcbHLHs in various tissues. It is expected to study the mechanism of Cd tolerance of *P. canescens*, aiming to cultivate new *P. canescens* materials with high concentration and high tolerance to repair soil cadmium pollution.

Deleted: of the

Commented [SP1]: Please rephrase this sentence

Keywords

bHLH gene family, *Populus canescens*, Cadmium stress

Deleted: d

INTRODUCTION

The eukaryotic domain contains a broad distribution of bHLH (basic/helix-loop-helix) proteins. It is one of the largest transcription factor family after MYB in plants (Ledent and Vervoort, 2001, Riechmann and Ratcliffe, 2000). The bHLH proteins consists of conserved 60 amino acids which can be divided into two distinct functional regions (Jones, 2004). The basic region is distributed at the N-terminus of the domain and is composed of 17 residues which are related to DNA binding (Atchley et al., 1999). The helix-loop-helix region is located at the C-terminus of the domain and consists mainly of two hydrophilic α-helices separated by a loop of variable length and sequence (Nair and Burley, 2000). The HLH domain allows for the formation of homodimeric or heterodimeric complex and promotes protein-protein interactions (Ferre-D'Amare et al., 1994, Murre et al., 1989). The bHLH proteins regulate the expression of downstream genes by binding to specific motifs in target genes. The putative core recognition motif of bHLH proteins is a hexanucleotide E-box (5'-CANNTG-3'). There are different types of E-boxes depending on the two

Deleted: hydrophobic residues, forming

central bases in the sequence, the most common type of which is a G-box composed of a palindromic structure (5'-CACGTG-3'). Some conserved amino acids situated within the basic region of the domain are responsible for recognizing the core consensus sequence, whereas other residues determine the specificity for a particular E-box type (Robinson et al., 2000). The bHLH transcription factor family in eukaryotes has been categorized into six primary groups, distinguished by phylogenetic relationships, DNA-binding motifs, and functional characteristics (Atchley and Fitch, 1997, Ledent and Vervoort, 2001, Simionato et al., 2007). bHLHs are involved in a wide and diverse array of biological processes. Within yeast, bHLH proteins participate in chromosome segregation, enhancement of general transcription, and regulation of metabolic processes(Robinson and Lopes, 2000). In the realm of animals, bHLHs actively participate in sensing environmental signals, governing the cell cycle and circadian rhythms, and orchestrating a diverse array of essential developmental processes(Amoutzias et al., 2004, Atchley and Fitch, 1997, Ledent and Vervoort, 2001, Stevens et al., 2008). In plants, functional studies have shown the involvement of bHLHs in the regulation of growth, development, and stress responses(Ito et al., 2012, Kanaoka et al., 2008). Notably, bHLHs are known to perform key functions in mediating cadmium (Cd) tolerance of plants. For instance, AtbHLH38 and AtbHLH39 interact with AtbHLH29 to increase the tolerance of *A. thaliana* seedlings to Cd *via* decreased Cd transfer from roots to shoots and to improve the iron homeostasis and concentration of shoots(Wu et al., 2012). bHLH104 positively regulates four heavy metal detoxification-associated genes, *IREG2*, *MTP3*, *HMA3* and *nicotianamine synthase 4* (NAS4), which play roles in Cd sequestration and tolerance(Yao et al., 2018).

Cd is one of the most toxic heavy metals for animals and plants. It occurs in naturally high abundance in zinc and lead ores and phosphate fertilizers(Garrett and Robert, 2000, Mclaughlin and Singh, 1999). Cd can persist for years to decades in soils, which not only causes the irreversible influence in root and leaf development of crops, but also can be accumulated in the human body through the food chain to cause risks to human health(Sarwar et al., 2017). In 2010, according to the data of national soil surveys, more than 1/5 of farmland area in China was mainly polluted by Cd, which contributes to a reduction of 10^{10} kg in grain production and pollution of 1.2×10^{10} kg in grain(Rafiq et al., 2014, Tang et al., 2016). Therefore, Cd contamination in soil has drawn broad public attention and the remediation of soil contamination with Cd is urgent and imperative. Plants have evolved diverse defense mechanisms to cope with heavy metals, such as extrusion, chelation,

Deleted: B

Formatted: Font: Italic

Formatted: Font: Italic

Deleted: B

Deleted: While

Deleted: p

vacuolar sequestration and regulation of distribution(Clemens, 2006, Clemens et al., 2002, Lux et al., 2011, Verbruggen et al., 2009), and phytoremediation is an emerging technology for remediation of heavy metal-contaminated soil, which has a lot of advantages: efficient, eco-friendly and economic. This technology uses hyperaccumulators to remove heavy metals from soil(Kumar, 1995). Most hyperaccumulator species, however, have a small aboveground biomass, which limited their potential for large-scale remediation of contaminated soils.

Poplar are seen as candidates for removing heavy metal contamination from polluted soil because of their high growth rates, large aboveground biomass, deep and highly branched root systems, and efficient Cd uptake and translocation to the aerial parts(De Oliveira, 2018, Song et al., 2019). While most *Populus* species are not considered hyperaccumulators of Cd, specific poplar genotypes have demonstrated a remarkable ability to accumulate substantial amounts of Cd in their aboveground tissues. One such example is *Populus* × *canescens* (*P. canescens*), which accumulates Cd levels exceeding 100 μg g^{-1} dry mass in both leaf and bark tissues, surpassing the threshold for Cd hyperaccumulation in plants(He et al., 2011, Sun et al., 2009). Hence, *P. canescens* stands out as a prospective candidate for remediating Cd-polluted soil. The physiological and molecular mechanisms associated with Cd accumulation in *P. canescens* have been partially unveiled(He et al., 2011, Schützendübel et al., 2002). However, it remains uncertain how TFs, especially bHLH TF family members, influence the physiological functions of *P. canescens* in response to Cd stress.

In this study, 170 *bHLH* family genes were identified and characterized in the *P. canescens* genome, which were further classified in 22 subgroups and could be distributed over 19 chromosomes. The analysis has encompassed their gene structures, conserved motifs, DNA-binding capability, cis-acting elements in promoters, multiple sequence alignment, gene duplications, synteny examination, and phylogenetic relationships.

In addition, the tissue-specific and Cd induced expression profiles analysis of some *PcbHLH* genes were conducted. This study lays a solid foundation for subsequent research on the functions and regulatory mechanisms of *PcbHLH* genes in response to Cd stress. It also identifies potential candidate genes associated with Cd stress resistance in *P. canescens*.

MATERIALS AND METHODS

Identification of the basic/helix-loop-helix family genes in *P. canescens*

Deleted: .

Deleted:

Deleted: A

Commented [SP2]: Is a reference available for this statement?

Commented [SP3]: Please briefly summarize the recent studies done to unravel the cd effect in P. canescens.

P. canescens genomic sequence data was downloaded from https://www.aspendb.org/. *Arabidopsis thaliana* bHLHs (AtbHLHs) were obtained from Lorenzo Carretero-Paulet(Carretero-Paulet et al., 2010). The information and sequence of *Populus trichocarpa* (*P. trichocarpa*) bHLHs (PtbHLHs) were retrieved from https://www.ncbi.nlm.nih.gov/. The Hidden Markov Model (HMM) files of bHLH domain (PF0010) were downloaded from https://www.ebi.ac.uk/interpro/, which were used as queries to hmmsearch the bHLH proteins of *P. canescens* (PcbHLHs) with SPDE software. AtbHLHs were also utilized for BLASTp searches against the *P. canescens* genomic sequence data. To further confirm the presence of the bHLH domain (E-value < $1e^{-5}$), CD-search (https://www.ncbi.nlm.nih.gov/) and SMART (http://smart.embl-heidelberg.de/) were used to analyze the obtained PcbHLH sequences. The molecular weights (kDa) and isoelectric points (pI) of the PcbHLHs were calculated by SPDE.

Chromosomal locations of *PcbHLHs*

To determine the physical locations of *PcbHLH* genes, the starting and ending positions of all *bHLH* genes on each chromosome were obtained from the *P. canescens* database. The diagram was drawn using TBtools.

Multiple sequence alignments of PcbHLHs

Multiple sequence alignments were performed using ClustalX in MEGA7 and then visualized using Jalview (version 1.8.3). Variable sequences at the N- and C-terminal regions were removed, while the conserved domains located in the central part were preserved. The sequence logos for bHLHs were generated through the submission of multiple alignment sequences to a dedicated online platform (https://weblogo.berkeley.edu/logo.cgi). Using the criteria described by Massari and Murre (Massari and Murre, 2000), the PcbHLH proteins were classified into two main categories, relying on sequence information within the N-terminal region of the bHLH domains: DNA binding and Non-DNA binding (less than four basic amino acids). The DNA binding bHLHs were further divided into two groups: E-box binding and Non-E-box binding (only Glu12 or Arg15). E-box binding can therefore be subdivided further into two subgroups: G-box binding (contains His/Lys8, Glu12 and Arg16) and Non-G-box binding (only Glu12 and Arg16).

Phylogenetic analysis of PcbHLHs

To investigate the evolutionary relationships among *P. canescens* and *Arabidopsis* bHLHs, multiple sequences alignment of 170 complete PcbHLH sequences and 167 complete AtbHLH sequences

Commented [SP4]: P. canescens genomic data could not be found in the website mentioned. So please check again and provide the exact source from where the data was retrieved.

Deleted: ere

Commented [SP5]: Please provide the external link for this software.

Deleted: d

Commented [SP6]: Please provide the external link to access the database or a reference.

Commented [SP7]: Please provide a reference

Deleted: as

was generated using MEGA7. A Maximum Likelihood phylogenetic tree was constructed using MEGA7 with a bootstrap value set to 1,000 replicates. Itol (https://itol.embl.de/) was employed for visualization of the phylogenetic tree.

Gene structure and protein motifs analysis

The exon/intron organization and splicing phase of the predicted *PcbHLHs* were investigated based on the Gff3 annotation files of *P. canescens* genome and then graphically displayed by TBtools. Conserved motifs within PcbHLH proteins were determined using MEME (https://meme-suite.org/meme/tools/meme, Version 5.5.4) under default parameters, with the following adjustments: (1) the optimal motif width ranged from ⩾10 to ⩽100, and (2) a maximum of ten motifs were identified. The resulting phylogenetic trees, gene structures, and conserved motifs were visualized using TBtools.

Cis-Acting element analysis in PcbHLH promoters

TBtools was utilized to retrieve the 2000bp upstream sequences of *PcbHLH* genes. These sequences were regarded as promoter regions. PlantCARE (https://bioinformatics.psb.ugent.be/webtools/plantcare/html/) was employed to analyze the cis-acting elements of the promoters. A heat map representing the number of cis-acting elements was created using Rstudio (Version 4.2.2). TBtools was used for visualizing the cis-acting elements of the promoters.

Gene duplication and synteny analysis of *PcbHLH* genes

The duplication events and collinearity of the *bHLH* genes of poplar and other plants were analyzed by Multiple Collinearity Scan toolkit (MCScanX) software, and the diagram was drawn using TBtools. The estimation of the nonsynonymous (ka) and synonymous (ks) substitution rates of *bHLH* gene pairs were analyzed with TBtools software.

Plant growth conditions and Cd treatment

P. canescens seedlings were obtained through micropropagation and those well-rooted poplar plantlets showing uniform growth status were cultured in the following hydroponically cultivation at 25 ˚C under long days (16 h light/ 8 h dark each day) in an artificial climate chamber grew in the half-strength Hoagland nutrient medium for a month. $CdCl_2$ solution was prepared for Cd stress treatment. *P. canescens* seedlings were then placed in half- strength Hoagland nutrient medium with 10 μM of Cd and cultured for 168h. Roots, stems and leaves of

Deleted: ere

Deleted: s

Deleted: and

Commented [SP8]: Please provide a reference for the protocol used and revise the paragraph appropriately. The sentence formation in this paragraph needs to be revised.

196 individual *P. canescens* were sampled at 0, 6, 12, 24, 48, 96, and 168 h. All samples were collected
197 at the same time on each sampling day to eliminate the influences of plant rhythm.

198 **RNA extraction and qRT-PCR analysis**

199 Total RNA of *P. canescens* seedlings was extracted from root, stem, and leaf samples using the
200 RNAprep Pure Plant Plus Kit (TIANGEN, Beijing, China). First-strand cDNA was synthesized
201 using PrimeScript™ RT Master Mix (TaKaRa, Dalian, China). The quantitative primers were
202 designed using TBtools. The qPCR was performed with 2 × Q3 SYBR qPCR Master Mix
203 (TOLOBIO, Shanghai, China) on a 7300 Real-Time PCR System (Applied Biosystems, CA, United
204 States). The relative expression levels were calculated using the $2^{-\Delta\Delta CT}$ method using *EF1B* gene as
205 reference (Wildhagen et al., 2010).

206

207 ## RESULTS

208 **Identification and classification of basic/helix-loop-helix protein in *P. canescens***

209 A total of 179 predicted bHLH proteins were initially obtained in *P. canescens* using HMMER with
210 Hidden Markov Models and BLASTp. After filter of uncertain proteins with incomplete domains
211 through CD-Search and SMART, 170 putative bHLH proteins were finally identified (Table S1),
212 which were designated as *PcbHLH1* to *PcbHLH170* based on their position in the genome. Analysis
213 of the biochemical properties of these proteins (Table S2) revealed that their amino acid lengths
214 ranged from 90 (*PcbHLH43* and *PcbHLH154*) to 740 (*PcbHLH46*). The molecular weights varied
215 from 10.29 kDa to 71.95 kDa, and the predicted isoelectric points ranged from 4.63 to 10.18. All
216 the proteins had grand average of hydropathy values below zero, suggesting that they are all
217 hydrophilic proteins.

218 **Chromosomal localization of *PcbHLH* genes**

219 Distribution mapping results showed 169 out of 170 *PcbHLH* genes are distributed unevenly in
220 different chromosomes of *P. canescens* genome, while *PcbHLH170* was not mapped to the
221 chromosomes. Chromosome 2 contained the largest number (17) of *PcbHLH* genes, whereas only
222 four *PcbHLH* genes were present on chromosome 3 and 7.

223 **Multiple sequence alignments of PcbHLHs**

224 A multiple sequence alignment was conducted for the 170 PcbHLH protein sequences. The results
225 revealed that the bHLH domain comprises four conserved regions: a basic region, two helical

Commented [SP9]: Please mention about the replicates used for the experiment.

Commented [SP10]: It is really confusing of whether the bHLH genes were selected from P. canescens or P. trichocarpa because the IDs provided in the Table S1 are from P. trichocarpa which were renamed as PcbHLH's. Please clarify and provide the Pc gene ID's in the table

Commented [SP11]: What could be reason for this gene not getting mapped on the chromosome if it has been selected from the same genome?

Commented [SP12]: Please provide the figure reference

regions, and a loop region. The basic region is composed of His-5, Glu-9, Arg-10, Arg-12, and Arg-13 residues, while the first helix region consists of Ile-16, Leu-23, Leu-27, Val-28, and Pro-29 residues. The loop region primarily consists of Lys-35 and Asp-41 residues, and the second helix region is composed of Ala-43, Leu-46, Glu-48, Ala-49, Ile-50, Tyr-52, and Leu-56. As shown in Figure 2, we could clearly observe a high conservation of Leu-23 in the 170 PcbHLH amino acid sequences, suggesting its crucial role in promoting dimerization among PcbHLH proteins.

The basic region of the PcbHLH proteins play essential roles as DNA-binding region. According to Massari's classification, these proteins are categorized into two main groups based on sequence information of the basic region within bHLH domains. A large groups consist of proteins with DNA-binding domains expected to bind DNA, and a smaller group is composed of proteins with non-DNA-binding domains (Massari and Murre, 2000). In total, 165 PcbHLH proteins were identified as DNA-binding proteins, while the remaining 5 were classified as non-DNA-binding proteins (Table 1). The DNA-binding proteins can be categorized into two classes: E-box bind proteins (including G-box binder) and non-E-box proteins. According to the conservation of the residues, 144 PcbHLHs were predicted to bind E-box, which contain Glu-9 and Arg-13 at the basic region Additionally, among the 144 E-box- binding proteins, 109 belonged to G-box-binding proteins depending on whether His/Lys-5, Glu-9, and Arg-13 are present in the basic region.

Phylogenetic analysis of *PcbHLH* genes

To investigate the evolutionary relationships between the PcbHLHs and the AtbHLHs, a maximum-likelihood phylogenetic tree was constructed. Analysis of the clade support values and the topology of the tree resulted in the division of the tree into 22 subfamilies. Figure 3 illustrates that each subfamily includes representatives from both *A. thaliana* and *P. canescens*, suggesting extreme conservation in the bHLH domains of AtbHLHs and PcbHLHs during evolution. The largest subfamily, Clade III, comprises 37 members, while the smallest subfamily, Clade XIV, consists of only 5 members.

Analysis of gene structure and conserved motif of bHLH family

To further understand the conservation and diversity of the PcbHLH proteins, motif analysis was conducted using MEME, 10 types of putative conserved protein motifs were identified. As depicted in Figure 4, members within the same subfamily exhibit a similar motif structure, indicating that they share similar structural and functional features. Gene structure analysis can help reveal the

Deleted: demonstrated the

Commented [SP13]: The table can be revised with details of the bHLH's that are grouped in each class rather than just providing the numbers which are already mentioned in the text.

Deleted: A

Commented [SP14]: Figure legends can be further improved.

evolutionary relationships of PcbHLH gene family members. Out of all 170 members, only 11 lacked introns and clustered closely in two subfamilies. The remaining 159 members containing introns displayed significant similarities in gene structure within the same subfamily, further suggesting that the members in the same subfamily have close evolutionary relationships.

Analysis of cis-acting elements in PcbHLH promoters

Cis-acting elements within the promoter regions serve as essential indicators for the classification of subfamilies and the functional characterization of family members. 2000bp upstream of the transcriptional start sites were identified as the promoter regions, from which several important cis-acting elements were identified for gene function analysis. These elements were categorized into three groups based on their functional categories: TC-rich repeats (cis-acting element involved in defense and stress responsiveness), LTR (cis-acting element involved in low-temperature responsiveness), MBS (MYB binding site involved in drought-inducibility) belong to stress-responsive. GCN4_motif (cis-regulatory element involved in endosperm expression), CAT-box (cis-acting regulatory element related to meristem expression), O2-site (cis-acting regulatory element involved in zein metabolism regulation) belong to plant development-related. AuxRR-core (cis-acting regulatory element involved in auxin responsiveness), P-box (gibberellin-responsive element), CGTCA/TGACG-motif (cis-acting regulatory element involved in the MeJA-responsiveness), ABRE (cis-acting element involved in the abscisic acid responsiveness), SARE (cis-acting element involved in salicylic acid responsiveness) belong to phytohormone responsive. Analysis depicted in Figure 5 revealed that most members contained phytohormone-related cis-acting elements, suggesting the possible involvement of these genes in responses to different abiotic stressors.

Gene duplication of PcbHLHs and comparative syntenic analysis across species

To understand the predominant driving force in the evolutionary process of *PcbHLH* genes, the gene duplication events within *P. canescens* were examined using TBtools. As shown in Figure 6, we obtained 92 segmental duplication gene pairs, while there was one tandem duplication gene pair (*PcbHLH48*/*PcbHLH58*) in *P. canescens*. These results indicated that *PcbHLH* genes were mainly generated by gene duplication during evolution. To disclose the direction of evolution, the Ka/Ks ratio of *PcbHLH* duplicated genes were calculated. As indicated in Table S3, the Ka/Ks ratio for all gene pairs was below 0.8. The maximal was 0.72 and the minimum was 0.08, suggesting that the

Commented [SP15]: Are these elements grouped only into 3 categories? But the figures shows more than 3 so please revise this section appropriately.

Commented [SP16]: Please check the spelling in the figure

Deleted: 7

bHLH genes have undergone predominately purifying selection.

In order to further illuminate the phylogenetic mechanisms of PcbHLH family, we constructed a comparison of the syntenic map of *P. canescens* related to *P. trichocarpa* and *Arabidopsis*, respectively. A total of 144 orthologous *bHLH* gene pairs were present between *P. canescens* and P. *trichocarpa* and 93 orthologs between *P. canescens* and *A. thaliana*. The number of orthologous gene pairs between *P. canescens* and *P. trichocarpa* was much greater than that between *P. canescens* and *A. thaliana*, suggesting that *P. canescens* was more closely related to *P. trichocarpa* than to *A. thaliana*.

***PcbHLH* genes expression patterns in response to Cd stress**

In previous studies, the members of clade IVc and Ib have been characterized as key transcription factors regulating the Cd stress response in *Arabidopsis* (Hao et al., 2021). According to the phylogenetic analyses of AtbHLHs and PcbHLHs proteins, we could preliminarily explore the functions of PcbHLH proteins. To pinpoint the *PcbHLH* genes crucial for Cd stress responses, we meticulously screened 14 *bHLH* genes in *P. canescens*, considering their confirmed functional homologs in the same subfamily. We then examined their expression patterns under Cd treatments. As depicted in Figure 7A, in the roots, *PcbHLH148*, *PcbHLH98*, *PcbHLH10*, *PcbHLH122*, *PcbHLH168*, *PcbHLH167*, *PcbHLH169*, *PcbHLH96*, *PcbHLH162* exhibited upregulated expression after Cd treatment, with *PcbHLH19*, *PcbHLH50*, *PcbHLH47*, *PcbHLH164* and *PcbHLH61*, showing downregulated expression. In the stems, the expression of *PcbHLH162* showing a decreased expression, while the most remaining genes exhibited upregulation until 6h, the rest *PcbHLH50*, *PcbHLH122*, *PcbHLH47* exhibited upregulation until 48h. In the leaves, all selected genes, except for *PcbHLH61*, exhibited upregulated expression. *PcbHLH162* upregulated until 6h, and the rest genes upregulated until 168h. This indicates a distinct positive regulatory response of *PcbHLH* genes to Cd threats, with most genes being highly expressed in the leaves (Figure 7B); Only *PcbHLH96* exhibited high expression in the roots.

Deleted: transcriptional alterations

Commented [SP17]: Can you please explain how this heat map was generated for all these genes? Was it only by qPCR? Please provide the raw data for the expression analysis performed for this study.

DISCUSSION

Various plants have been extensively studied to identify their bHLH families, revealing that the bHLH transcription factor family, the second largest in eukaryotes(Ledent and Vervoort, 2001, Riechmann and Ratcliffe, 2000), is particularly noteworthy. The abundance of plant genome data

has played a pivotal role, providing indispensable sequence resources to facilitate the comprehensive genome-wide identification of *bHLH* genes in plants. Currently, the bHLH family has been defined and mined in many plants, however, there has been no related work on its distribution and function in Cd-tolerant Poplar, *P. canescens*. In this study, 170 *bHLH* genes were identified and characterized in *P. canescens* (Figure 3), which is more than *A. thaliana* (167) and tomato (159) but less than rice (177) and *P. trichocarpa* (183) (Carretero-Paulet et al., 2010). Based on the genome size, the ratio to *PcbHLH* genes in the *P. canescens* genome was less than *Arabidopsis* but more than tomato. These findings suggest that the count of *bHLH* genes varies among different plant species and is not directly proportional to the size of their genomes. Based on phylogenetic analysis, we classified 170 *PcbHLH* genes into 22 subfamilies. Each subfamily contains both AtbHLHs and PcbHLHs, suggesting that the *PcbHLH* genes had diversified before these two species evolved. There were only six subfamilies of bHLH in animal genomes. But in plants, the bHLH gene family was divided into 24 subfamilies in tomato(Wang et al., 2015), 17 in *Ginkgo biloba*(Zhou et al., 2020), 20 in *Camellia sinensis*(Cui et al., 2018), 21 in *A. thaliana*(Toledo-Ortiz et al., 2003). This suggest that the classification of bHLH gene family in plants is very different from that of animals.

Analyzing conserved protein motifs and gene structures (Figure 4) clearly indicates that members belonging to the same subfamily typically have a shared evolutionary origin and engage in similar physiological activities(Ke et al., 2020). Moreover, modifications in gene structure have been uncovered as potential factors leading to variations in gene functions. These variations are influenced by three key mechanisms: The three primary mechanisms influencing variations in gene functions are exon/intron gain or loss, exonization or pseudo-exonization, and insertion or deletion. In addition, we observed a remarkable uniformity in both the quantity and arrangement of exons/introns among members belonging to the same subfamily. It is worth highlighting the previous reports that have indicated that the number of exons/introns in sesame *bHLH* genes varies from 0 to 10(Kazemitabar et al., 2020). Likewise, *Andrographis paniculata* exhibits a range of 0 to 14 for the number of exons/introns(Xu et al., 2022), and *P. canescens* displays variability from 0 to 14 in terms of exons/introns. These findings parallel the earlier results mentioned.

The expansion of gene family may be due to gene duplication events, including tandem, fragment, whole-genome duplication, and transposition events(Flagel and Wendel, 2009, Zhang, 2003). Two

Deleted: .

Deleted: W

Deleted: ,

Deleted: at

or more genes on the same chromosome are often related to tandem duplication, while segmental duplication often occurs on different chromosomes(Schlueter et al., 2007). Our results indicate that the expansion of gene family in *P. canescens* may be caused by tandem duplication and segmental duplication (Figure 6A). The two comparative syntenic maps of the association between *P. canescens* and *P. trichocarpa* or *A. thaliana*, which further clarify the origin and evolution of the PcbHLH gene family (Figure 6B), showed that PcbHLHs had a syntenic relationship with 144 and 93 genes in *P. trichocarpa* and *A. thaliana*, respectively. These indicate that *P. canescens* is phylogenetically closer to *P. trichocarpa* than to *A. thaliana*. Furthermore, the analysis of the Ka/Ks ratio revealed that *PcbHLH* gene pairs experienced purifying selection (Table S3). Key residues within the basic region play a crucial role in distinguishing variations in the specific hexanucleotide sequence core at the promoter of target genes. This capacity enables the classification of plant bHLHs into distinct DNA-binding categories. Using the criteria described by Massari and Murre(Massari and Murre, 2000), the PcbHLH proteins were divided into several categories (Table 1). The proportion of G-box binding were 64% which was the maximal DNA binding. Non-G-box binding were 21%, Non-E-box binding were 12%, Non-DNA binding were 3%. Compared with *A. thaliana* the Non-DNA binding were 18% far more than Non-DNA binding of PcbHLHs(Toledo-Ortiz et al., 2003). Yet, it is essential to confirm whether these non-DNA binding sequences still possess DNA-binding activity despite having a low basic region.

Cd is toxic pollutants that is detrimental to living organisms. Contamination with heavy metals such as Cd is a serious threat to human health. Cd can cause human diseases, including kidney disorders, neurotoxicity and osteoporosis(Jarup and Akesson, 2009, Satarug et al., 2010). Even moderate levels of heavy metals can be damaging to living organisms when exposed for an extended period. Consumption of crop plants cultivated on contaminated sites or irrigated with polluted groundwater may result in long-term heavy metal absorbance and accumulation in the human body(Shimbo et al., 2001, Zahir et al., 2005) and controlling the heavy metal pollution in soil has become a key problem. Phytoremediation is a relatively new technology, with limited research conducted in recent years. The majority of hyperaccumulators employed in phytoremediation are herbaceous plants with small biomass and underdeveloped root systems. Additionally, large woody plants with well-developed roots, such as *P. canescens*, a rapidly growing poplar species, are considered suitable for phytoremediation. Prior research has unveiled the crucial involvement of

Formatted: Font: Italic

Deleted: at

Deleted: ure

Deleted: occurs over

Deleted: . So, how to

bHLH proteins in the plant's response to stress induced by heavy metals. For instance, heterologous expression of soybean *GmORG3*, a *bHLH* family gene, enhanced Cd tolerance in yeast. Furthermore, the overexpression of *GmORG3* in soybean mosaic seedlings using a hairy root system showed that overexpressing plants enhanced Cd tolerance and stabilized Fe homeostasis(Xu et al., 2017). In *Arabidopsis*, several closely related *bHLH* genes from the same clades have been characterized as key transcription factors regulating the Cd stress response. *FIT/bHLH38* or *FIT/bHLH39* heterodimers can respond to Cd tolerance by affecting the transcriptional expression of downstream genes(Wu et al., 2012, Yuan et al., 2008). *BHLH104* loss-of-function mutants were sensitive to Cd stress, and the Cd tolerance was enhanced upon overexpression of *bHLH104*(Tissot et al., 2019). Based on the phylogenetic tree constructed with bHLH proteins in *P. canescens* and *A. thaliana*, PcbHLH proteins from Clade XV (*FIT/bHLH38* and *FIT/bHLH39*), Clade XVI, Clade XIV and Clade XIII (neighboring subfamily) were clustered together with well-defined AtbHLH proteins known for their association with the plant's response to heavy metal stress. This clustering offers a valuable reference for predicting the biological functions of *PcbHLH* genes. We screened these 14 *PcbHLH* genes to investigate the molecular mechanism of PcbHLH proteins during Cd enrichment in *P. canescens*, RT-qPCR was performed to analyze the transcript levels of these *PcbHLHs* under Cd treatment to verify whether the poplar *bHLH* family could respond to Cd stress. The results showed that these orthologous genes responded to Cd stress indicating they might be the essential regulators involved in the Cd stress (Figure 7A). The expression level increased to a maximum in leaves within 168 hours and in stems within 6 hours suggesting that these *PcbHLH* genes were involved in the long- and short-term regulation of Cd stress. Most of *PcbHLH* genes were highly expressed in leaves thereby confirming tissue-specific expression (Figure 7B). To investigate the regulatory mechanisms of the *PcbHLH* genes, we analyzed their cis-elements from the transcriptional start site to the –2000 bp upstream region. The cis-acting element analysis revealed that PcbHLH family genes were widely involved in a variety of physiological processes, including plant growth and development, hormone responses and stress responses. In addition, several stress response-related elements exist widely in PcbHLH gene promoter regions, suggesting that PcbHLHs play pivotal roles in mediating the plant's response to abiotic stress.

The bHLH family, one of the largest transcription factor families, is inadequately studied in plants, with many family members' functions remaining unexplored. Analyzing the *PcbHLH* genes will

Deleted: ,

Deleted: were

Deleted: ion

Deleted: plant

Deleted: ,

Deleted: plant

enhance our fundamental understanding and provide a theoretical basis for breeding new Cd pollution-resistant germplasm, aiming to cultivate new *P. canescens* materials with high concentration and high tolerance to repair soil Cd pollution.

Supplementary Materials: Table S1: The detail information of bHLH members in *Populus canescens* and *Arabidopsis thaliana*, Table S2: The analysis of physicochemical properties of PcbHLHs, Table S3: Ka/Ks ratio of putative orthologous and paralogous pairs in *P. canescens*. Table S4: qPCR primer.

Funding: This work was supported by the National Nonprofit Institute Research Grant of CAF (No. RISF2021YZ01 and RISFZ-2021-01).

Author Contributions: W.Q. conceived and designed the experiments; Data curation, Y.Y. and Z.H.; Formal analysis, Y.Y.; Investigation, Y.Y., X.L., and H.L.; Software, Y.Y., X.H., and J.X.; Validation, Z.H. and W.Q.; Writing—original draft, Y.Y. and X.L.; Writing—review & editing, W.Q. and R.Z. All authors read and approved the final manuscript.

Declaration of competing interest: The authors declare that the study was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

References

Amoutzias, G. D., Robertson, D. L., Oliver, S. G., Bornberg-Bauer, E., 2004, Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. EMBO Rep, 5: 274-9.

Atchley, W. R., Fitch, W. M., 1997, A natural classification of the basic helix-loop-helix class of transcription factors. Proc Natl Acad Sci U S A, 94: 5172-6.

Atchley, W. R., Terhalle, W., Dress, A., 1999, Positional dependence, cliques, and predictive motifs in the bHLH protein domain. J Mol Evol, 48: 501-16.

Carretero-Paulet, L., Galstyan, A., Roig-Villanova, I., Martinez-Garcia, J. F., Bilbao-Castro, J. R., Robertson, D. L., 2010, Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in Arabidopsis, poplar, rice, moss, and algae. Plant Physiol, 153: 1398-412.

Clemens, S., 2006, Toxic metal accumulation, responses to exposure and mechanisms of tolerance in plants. Biochimie, 88: 1707-19.

Clemens, S., Palmgren, M. G., Kramer, U., 2002, A long way ahead: understanding and engineering plant metal accumulation. Trends Plant Sci, 7: 309-15.

457 Cui, X., Wang, Y. X., Liu, Z. W., Wang, W. L., Li, H., Zhuang, J., 2018, Transcriptome-wide identification
458 and expression profile analysis of the bHLH family genes in Camellia sinensis. Funct Integr
459 Genomics, 18: 489-503.
460 De Oliveira, V. H. T., 2018, Tolerance, toxicity and transport of Cd and Zn in Populus trichocarpa.
461 Environmental experimental botany, 155.
462 Ferre-D'Amare, A. R., Pognonec, P., Roeder, R. G., Burley, S. K., 1994, Structure and function of the
463 b/HLH/Z domain of USF. EMBO J, 13: 180-9.
464 Flagel, L. E., Wendel, J. F., 2009, Gene duplication and evolutionary novelty in plants. New Phytol, 183:
465 557-564.
466 Garrett, Robert, G., 2000, Natural Sources of Metals to the Environment. Human and Ecological Risk
467 Assessment, 6: 945-963.
468 Hao, Y., Zong, X., Ren, P., Qian, Y., Fu, A., 2021, Basic Helix-Loop-Helix (bHLH) Transcription Factors
469 Regulate a Wide Range of Functions in Arabidopsis. International Journal of Molecular Sciences,
470 22: 7152.
471 He, J., Qin, J., Long, L., Ma, Y., Li, H., Li, K., Jiang, X., Liu, T., Polle, A., Liang, Z., Luo, Z. B., 2011, Net
472 cadmium flux and accumulation reveal tissue-specific oxidative stress and detoxification in
473 Populus x canescens. Physiol Plant, 143: 50-63.
474 Ito, S., Song, Y. H., Josephson-Day, A. R., Miller, R. J., Breton, G., Olmstead, R. G., Imaizumi, T., 2012,
475 FLOWERING BHLH transcriptional activators control expression of the photoperiodic flowering
476 regulator CONSTANS in Arabidopsis. Proc Natl Acad Sci U S A, 109: 3582-7.
477 Jarup, L., Akesson, A., 2009, Current status of cadmium as an environmental health problem. Toxicol
478 Appl Pharmacol, 238: 201-8.
479 Jones, S., 2004, An overview of the basic helix-loop-helix proteins. Genome Biol, 5: 226.
480 Kanaoka, M. M., Pillitteri, L. J., Fujii, H., Yoshida, Y., Bogenschutz, N. L., Takabayashi, J., Zhu, J. K., Torii, K.
481 U., 2008, SCREAM/ICE1 and SCREAM2 specify three cell-state transitional steps leading to
482 arabidopsis stomatal differentiation. Plant Cell, 20: 1775-85.
483 Kazemitabar, S. K., Faraji, S., Najafi-Zarrini, H., 2020, Identification and in silico evaluation of bHLH genes
484 in the Sesamum indicum genome: Growth regulation and stress dealing specially through the
485 metal ions homeostasis and flavonoid biosynthesis - ScienceDirect. Gene Reports, 19.
486 Ke, Y. Z., Wu, Y. W., Zhou, H. J., Chen, P., Wang, M. M., Liu, M. M., Li, P. F., Yang, J., Li, J. N., Du, H., 2020,
487 Genome-wide survey of the bHLH super gene family in Brassica napus. BMC Plant Biol, 20: 115.
488 Kumar, A., 1995, Flood control and rural settlement planning, pp. 192 p., Mohit Publications, New Delhi.
489 Ledent, V., Vervoort, M., 2001, The basic helix-loop-helix protein family: comparative genomics and
490 phylogenetic analysis. Genome Res, 11: 754-70.
491 Lux, A., Martinka, M., Vaculik, M., White, P. J., 2011, Root responses to cadmium in the rhizosphere: a
492 review. J Exp Bot, 62: 21-37.
493 Massari, M. E., Murre, C., 2000, Helix-loop-helix proteins: regulators of transcription in eucaryotic
494 organisms. Mol Cell Biol, 20: 429-40.
495 Mclaughlin, M. J., Singh, B. R., 1999, Cadmium in Soils and Plants || The Environmental Chemistry of
496 Cadmium. 10.1007/978-94-011-4473-5: 11-37.
497 Murre, C., McCaw, P. S., Baltimore, D., 1989, A new DNA binding and dimerization motif in
498 immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell, 56: 777-83.
499 Nair, S. K., Burley, S. K., 2000, Recognizing DNA in the library. Nature, 404: 715, 717-8.
500 Rafiq, M. T., Aziz, R., Yang, X., Xiao, W., Rafiq, M. K., Ali, B., Li, T., 2014, Cadmium phytoavailability to rice

(Oryza sativa L.) grown in representative Chinese soils. A model to improve soil environmental quality guidelines for food safety. Ecotoxicol Environ Saf, 103: 101-7.

Riechmann, J. L., Ratcliffe, O. J., 2000, A genomic perspective on plant transcription factors. Curr Opin Plant Biol, 3: 423-34.

Robinson, K. A., Koepke, J. I., Kharodawala, M., Lopes, J. M., 2000, A network of yeast basic helix-loop-helix interactions. Nucleic Acids Res, 28: 4460-6.

Robinson, K. A., Lopes, J. M., 2000, SURVEY AND SUMMARY: Saccharomyces cerevisiae basic helix-loop-helix proteins regulate diverse biological processes. Nucleic Acids Res, 28: 1499-505.

Sarwar, N., Imran, M., Shaheen, M. R., Ishaque, W., Kamran, M. A., Matloob, A., Rehim, A., Hussain, S., 2017, Phytoremediation strategies for soils contaminated with heavy metals: Modifications and future perspectives. Chemosphere, 171: 710-721.

Satarug, S., Garrett, S. H., Sens, M. A., Sens, D. A., 2010, Cadmium, environmental exposure, and health outcomes. Environ Health Perspect, 118: 182-90.

Schlueter, J. A., Lin, J. Y., Schlueter, S. D., Vasylenko-Sanders, I. F., Deshpande, S., Yi, J., O'Bleness, M., Roe, B. A., Nelson, R. T., Scheffler, B. E., Jackson, S. A., Shoemaker, R. C., 2007, Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. BMC Genomics, 8: 330.

Schützendübel, A., Nikolova, P., Rudolf, C., Polle, A., 2002, Cadmium and H2O2-induced oxidative stress in Populus × canescens roots. 40: 577-584.

Shimbo, S., Zhang, Z. W., Watanabe, T., Nakatsuka, H., Matsuda-Inoguchi, N., Higashikawa, K., Ikeda, M., 2001, Cadmium and lead contents in rice and other cereal products in Japan in 1998-2000. Sci Total Environ, 281: 165-75.

Simionato, E., Ledent, V., Richards, G., Thomas-Chollier, M., Kerner, P., Coornaert, D., Degnan, B. M., Vervoort, M., 2007, Origin and diversification of the basic helix-loop-helix gene family in metazoans: insights from comparative genomics. BMC Evol Biol, 7: 33.

Song, J., Finnegan, P. M., Liu, W., Li, X., Yong, J. W. H., Xu, J., Zhang, Q., Wen, Y., Qin, K., Guo, J., Li, T., Zhao, C., Zhang, Y., 2019, Mechanisms underlying enhanced Cd translocation and tolerance in roots of Populus euramericana in response to nitrogen fertilization. Plant Sci, 287: 110206.

Stevens, J. D., Roalson, E. H., Skinner, M. K., 2008, Phylogenetic and expression analysis of the basic helix-loop-helix transcription factor gene family: genomic approach to cellular differentiation. Differentiation, 76: 1006-22.

Sun, Y., Zhou, Q., Wang, L., Liu, W., 2009, Cadmium tolerance and accumulation characteristics of Bidens pilosa L. as a potential Cd-hyperaccumulator. J Hazard Mater, 161: 808-14.

Tang, X., Li, Q., Wu, M., Lin, L., Scholz, M., 2016, Review of remediation practices regarding cadmium-enriched farmland soil with particular reference to China. J Environ Manage, 181: 646-662.

Tissot, N., Robe, K., Gao, F., Grant-Grant, S., Boucherez, J., Bellegarde, F., Maghiaoui, A., Marcelin, R., Izquierdo, E., Benhamed, M., Martin, A., Vignols, F., Roschzttardtz, H., Gaymard, F., Briat, J. F., Dubos, C., 2019, Transcriptional integration of the responses to iron availability in Arabidopsis by the bHLH factor ILR3. New Phytol, 223: 1433-1446.

Toledo-Ortiz, G., Huq, E., Quail, P. H., 2003, The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell, 15: 1749-70.

Verbruggen, N., Hermans, C., Schat, H., 2009, Molecular mechanisms of metal hyperaccumulation in plants. New Phytol, 181: 759-776.

Wang, J., Hu, Z., Zhao, T., Yang, Y., Chen, T., Yang, M., Yu, W., Zhang, B., 2015, Genome-wide analysis of

545 bHLH transcription factor and involvement in the infection by yellow leaf curl virus in tomato
546 (Solanum lycopersicum). BMC Genomics, 16: 39.
547 Wildhagen, Durr, Ehlting, PHYSIOL, R. J. T., 2010, Seasonal nitrogen cycling in the bark of field-grown
548 Grey poplar is correlated with meteorological factors and gene expression of bark storage
549 proteins. 2010,30(9): 1096-1110.
550 Wu, H., Chen, C., Du, J., Liu, H., Cui, Y., Zhang, Y., He, Y., Wang, Y., Chu, C., Feng, Z., Li, J., Ling, H. Q., 2012,
551 Co-overexpression FIT with AtbHLH38 or AtbHLH39 in Arabidopsis-enhanced cadmium
552 tolerance via increased cadmium sequestration in roots and improved iron homeostasis of
553 shoots. Plant Physiol, 158: 790-800.
554 Xu, J., Xu, H., Zhao, H., Liu, H., Xu, L., Liang, Z., 2022, Genome-wide investigation of bHLH genes and
555 expression analysis under salt and hormonal treatments in Andrographis paniculata. Industrial
556 Crops and Products, 183: 114928.
557 Xu, Z., Liu, X., He, X., Xu, L., Huang, Y., Shao, H., Zhang, D., Tang, B., Ma, H., 2017, The Soybean Basic
558 Helix-Loop-Helix Transcription Factor ORG3-Like Enhances Cadmium Tolerance via Increased
559 Iron and Reduced Cadmium Uptake and Transport from Roots to Shoots. Front Plant Sci, 8:
560 1098.
561 Yao, X., Cai, Y., Yu, D., Liang, G., 2018, bHLH104 confers tolerance to cadmium stress in Arabidopsis
562 thaliana. J Integr Plant Biol, 60: 691-702.
563 Yuan, Y., Wu, H., Wang, N., Li, J., Zhao, W., Du, J., Wang, D., Ling, H. Q., 2008, FIT interacts with AtbHLH38
564 and AtbHLH39 in regulating iron uptake gene expression for iron homeostasis in Arabidopsis.
565 Cell Res, 18: 385-97.
566 Zahir, F., Rizwi, S. J., Haq, S. K., Khan, R. H., 2005, Low dose mercury toxicity and human health. Environ
567 Toxicol Pharmacol, 20: 351-60.
568 Zhang, J., 2003, Evolution by gene duplication: an update. Trends in Ecology Evolution, 18: 292-298.
569 Zhou, X., Liao, Y., Kim, S. U., Chen, Z., Xu, F., 2020, Genome-wide identification and characterization of
570 bHLH family genes from Ginkgo biloba. Scientific Reports, 10.
571