Comparative genomics analysis of bHLH genes in cucurbits identifies a novel gene regulating cucurbitacin biosynthesis

Running title: A new bHLH regulates cucurbitacin biosynthesis

Yuanchao Xu\textsuperscript{1}, Huimin Zhang\textsuperscript{2}, Yang Zhong\textsuperscript{1,3}, Naiyu Jiang\textsuperscript{2}, Xiaoyun Zhong\textsuperscript{1}, Qiqi Zhang\textsuperscript{1}, Sen Chai\textsuperscript{2}, Hongbo Li\textsuperscript{3} and Zhonghua Zhang\textsuperscript{2*}

\textsuperscript{1} Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China.

\textsuperscript{2} Engineering Laboratory of Genetic Improvement of Horticultural Crops of Shandong Province, College of Horticulture, Qingdao Agricultural University, Qingdao 266109, China.

\textsuperscript{3} Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Shenzhen Key Laboratory of Agricultural Synthetic Biology, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China.

*Correspondence to zhangzhonghua@qau.edu.cn; Tel: 0532-58957760; Fax: 0532-58957760

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. This is an Open Access article distributed under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract

The basic helix-loop-helix (bHLH) family of transcription factors (TFs) participate in a variety of biological regulatory processes in plants, and have undergone significant expansion during land plant evolution by gene duplications. In cucurbit crops, several bHLH genes have been found to be responsible for the agronomic traits such as bitterness. However, the characterization of bHLH genes across the genomes of cucurbit species has not been reported, and how they have evolved and diverged remains largely unanswered. Here we identified 1,160 bHLH genes in seven cucurbit crops and performed a comprehensive comparative genomics analysis. We determined orthologous and paralogous bHLH genes across cucurbit crops by syntenic analysis between or within species. Orthology and phylogenetic analysis of the tandem-duplicated bHLH genes in the Bt cluster which regulate the biosynthesis of cucurbitacins suggest that this cluster is derived from three ancestral genes after the cucurbit-common tetraploidization event. Interestingly, we identified a new conserved cluster paralogous to the Bt cluster that includes two tandem bHLH genes, and the
evolutionary history and expression profiles of these two genes in the new cluster suggest the involvement of one gene (Brp) in the regulation of cucurbitacin biosynthesis in roots. Further biochemical and transgenic assays in melon hairy roots support the function of Brp. This study provides useful information for further investigating the functions of bHLH TFs and novel insights into the regulation of cucurbitacin biosynthesis in cucurbit crops and other plants.

**Keywords:** bHLH TF, cucurbit crops, comparative genomics, cucurbitacin

**Introduction**

Basic helix-loop-helix (bHLH) transcription factors (TFs) are widespread in eukaryotes and constitute one of the largest TF families in plants\(^1\). Many studies suggest that the variety of bHLH genes in plants were derived from one or a few predecessors through a significant number of gene duplications\(^2-4\). Segmental and tandem duplications or whole-genome duplications (WGDs) should contribute to the expansion of copy numbers in the bHLH family\(^5,6\). In addition to the eudicot-common hexaploidy event, cucurbit crops also underwent an ancient cucurbit-common tetraploidization (CCT) event at ~90 million years ago (MYA)\(^7\), and cucurbita species have a recent WGD\(^8\). Which bHLH genes remained after these multiple duplication events and how they have evolved have not been investigated in cucurbit crops.

bHLH TFs can act as transcriptional activators or repressors and play essential roles in plant developmental and physiological processes, such as the regulation of flag leaf angle\(^9\) and shoot branching\(^10\), and responses to light\(^11\), phytohormones\(^12\) and low temperature\(^13\). Notably, many bHLH TFs regulate the biosynthesis of specialized secondary metabolites in plants\(^14-19\), including cucurbitacins in cucumber (Cucumis sativus), melon (Cucumis melo), and watermelon (Citrullus lanatus)\(^17,18\).
Cucurbitacins are triterpenoids that confer a bitter taste in Cucurbitaceae plants\textsuperscript{17}. The distinct cucurbitacins in cucurbits are structurally similar, and are synthesized mainly by conserved syntenic biosynthetic genes, which was reported to be regulated by a cluster (hereafter referred to as the Bt cluster) harboring several bHLHs\textsuperscript{17,18}. In cucumber, the Bt (Bitter fruit) and Bl (Bitter leaf), located in the Bt cluster, can regulate the accumulation of cucurbitacin C (CuC) in fruits and leaves, respectively\textsuperscript{17}. In melon and watermelon, the syntenic homologs of Bt cluster were found to exert similar function in diverse tissues by regulating cucurbitacin B (CuB) in melon and cucurbitacin E (CuE) in watermelon\textsuperscript{18}. However, the evolutionary history and functional divergence of genes within the Bt cluster are largely unknown.

In this study, we identified and characterized 1,160 bHLH genes and performed a comparative evolutionary analysis in the seven cucurbit crops: cucumber, melon, watermelon, bottle gourd (Lagenaria siceraria), wax gourd (Benincasa hispida), bitter gourd (Momordica charantia), and pumpkin (Cucurbita pepo). We determined orthologous and paralogous relationships among these bHLH genes and investigated features of bHLH tandem-duplicated genes (TDGs). Furthermore, we describe the evolutionary history and divergence of the Bt cluster and discovered a novel functional bHLH gene regulating cucurbitacin biosynthesis in a Bt paralogous cluster. These results enhance our understanding of the biosynthetic regulation of secondary metabolites through TFs.

Results

Identification and classification of bHLH genes in seven cucurbit crops
By searching the genomes of seven cucurbits we identified 149 non-redundant \textit{bHLH} genes in cucumber, 151 in melon, 154 in watermelon, 155 in bottle gourd, 150 in wax gourd, 150 in bitter gourd and 251 in pumpkin. All \textit{bHLH} genes were numbered according to their genomic coordinates, yielding \textit{CsabHLH001–149}, \textit{CmebHLH001–151}, \textit{ClabHLH001–154}, \textit{LsibHLH001–155}, \textit{BhibHLH001–150}, \textit{MchbHLH001–150} and \textit{CmabHLH001–251} (Table S1). The number of \textit{bHLH} genes was comparable among the cucurbit species investigated here, with the exception of pumpkin, whose genome encoded 1.67-fold more \textit{bHLH} family members than the average number observed in the other six cucurbit crops, possibly due to a lineage-specific recent WGD event\(^8\).

The basic region in the \textit{bHLH} domain determines its DNA-binding activity\(^20\). Using previously published criteria\(^5\), the \textit{bHLH} TFs were divided into one group of DNA-binding proteins and another group of non-DNA-binding proteins (25.1–32.3% of all \textit{bHLH} TFs). Furthermore, we subdivided the DNA-binding \textit{bHLH} proteins into two subcategories as a function of their predicted cognate DNA-binding motif: E-box and non-E-box-binding proteins (8.0–9.3% of all \textit{bHLH} TFs). E-box-binding \textit{bHLH} proteins were further subdivided into two groups: G-box-binding proteins (47.1–53.2% of all \textit{bHLH} TFs) and non-G-box-binding proteins (11.0–13.3% of all \textit{bHLH} TFs) (Fig. 1a).

Intron and exon structures can help elucidate the phylogenetic relationships of a gene family. We identified 12 different intron distribution patterns based on the number (zero to three) and relative positions of introns within the region encoding the \textit{bHLH} domain (designated I–XI and O, Fig. 1b). About 80% of the identified \textit{bHLH}s featured the three most common patterns: I, IX, and XI. Pattern IX was the most common, with only one intron in the loop region. Pattern I was characterized by three introns at three highly conserved positions and was the second most common pattern. Finally, pattern XI with no intron in the \textit{bHLH} domain-encoding region was the third...
most common pattern (Fig. 1b). The intron distribution patterns in cucurbit crops are like as in tomato (Solanum lycopersicum)\textsuperscript{21} and Arabidopsis (Arabidopsis thaliana)\textsuperscript{5}, suggesting that this might be a common feature among plant species.

Fig. 1 DNA-binding ability and conserved intron distribution patterns of \textit{bHLH} genes in seven cucurbit crops. a Predicted DNA-binding characteristics based on the amino-acid sequences of the bHLH domain. A species tree of seven cucurbits is shown in left. Csa: cucumber, Cme: melon, Cla: watermelon, Lsi: bottle gourd, Bhi: wax gourd, Cma: pumpkin and Mch: bitter
gourd. CCT: cucurbit-common tetraploidization; sWGD: specific whole-genome duplication. Blue numerical values show the estimated divergence time of each node or occurrence time of WGD (MYA, million years ago). b Intron distribution patterns within the bHLH domain of seven cucurbit crops. Position of introns are indicated by triangles and numbered (1 to 3) based on the bHLH region of *CsaPIF3* (*CsaV3_2G007370*), which is exhibited at the top. The count and percentage of bHLHs displaying each pattern in seven cucurbit crops are given in the right table.

**Evolution of bHLH genes in seven cucurbit crops**

To understand the phylogenetic relationships among *bHLHs* in cucurbit crops, we constructed an unrooted neighbor-joining phylogenetic tree using a multiple alignment of the bHLH domain. The 1,160 cucurbit *bHLH* genes are clustered into 28 subfamilies according to the classification of *Arabidopsis* *bHLHs* (Fig. 2a and Fig. S1). In most subfamilies, DNA-binding properties, intron pattern distributions and motif architectures of *bHLH* genes were relatively conserved (Fig. S1), suggesting the reliability of the classification and phylogenetic tree of cucurbit *bHLH* genes presented here.

To further investigate the evolutionary trajectory of the bHLH family, we constructed a syntenic map and generated a syntenic gene list of *bHLH* genes across the seven cucurbit crops (Fig. 2b and Table S2). Most highly conserved syntenic blocks were shared by the seven cucurbit genomes, with ~90% of all *bHLH* genes mapping to these orthologous blocks. For instance, >10 orthologous *bHLH* genes formed a superblock that was perfectly matched in all seven cucurbit crop genomes, as represented by the magenta lines in Fig. 2b. We next clustered and identified 151 *bHLH* orthologous groups (OGs) present in at least two cucurbit genomes (Table S2, OG001–151) and 13 species-specific *bHLH* OGs (Table S2, OG152–164). In 92% of all OGs, bHLH TFs from each cucurbit crop exhibited the same DNA-binding ability and their encoding genes showed the same intron distribution pattern (Table S2).
We then identified paralogous gene pairs: 28 pairs in cucumber, 28 pairs in melon, 31 pairs in watermelon, 25 pairs in bottle gourd, 14 pairs in wax gourd, 33 pairs in bitter gourd, and 129 pairs in pumpkin (Fig. S2 and Table S3). About 94% of these paralogous bHLH TFs shared the same DNA-binding ability and their encoding genes showed the same intron distribution pattern (Fig. S2). All paralogous bHLH gene pairs had non-synonymous substitutions/synonymous substitutions (Ka/Ks) values below 0.6 (Fig. S3), suggesting that the majority of paralogous bHLH genes has undergone purifying selection. These paralogous genes might be the extant product of WGDs during the evolutionary history of cucurbit crops. These results offer novel insights into the evolution of bHLH genes in cucurbit crops.
Fig. 2 Evolution of bHLH genes in seven cucurbit crops. a An unrooted neighbor-joining phylogenetic tree of 1,322 bHLHs from cucumber, melon, watermelon, bottle gourd, wax gourd, pumpkin, bitter gourd and Arabidopsis. bHLHs are divided into 28 subfamilies. The Roman numerals in brackets denote subfamilies, as defined in Arabidopsis. Detailed information of the phylogenetic tree is given in Fig. S1. b Synteny analysis of bHLH genes among seven cucurbit
crops. A species tree of seven cucurbits is shown in left. CCT: cucurbit-common tetraploidization; sWGD: specific whole-genome duplication. Blue numerical values show the estimated divergence time of each node or occurrence time of WGD (MYA, million years ago). Magenta numerals mean bHLH genes number in each cucurbit crop. Magenta and grey lines display the collinear bHLH genes among seven cucurbit crop genomes. The light grey lines denote collinear blocks.

**Characterization of bHLH TDGs**

Local gene duplication generates tandem duplicated genes (TDGs) and is ubiquitous during genome evolution\(^{23}\). To explore the features and potential functions of bHLH TDGs, we identified clusters of bHLH TDGs: 10 genes in cucumber (four clusters); 9 in melon (three clusters); 14 in watermelon (five clusters); 15 in bottle gourd (seven clusters); 10 in wax gourd (four clusters); 10 in bitter gourd (four clusters); and 8 in pumpkin (four clusters), based on chromosome localization of the genes and sequence similarity of the encoded proteins (Fig. S4 and Table1). TDG clusters were named according to their order in their genomic coordinates, such as “T1”, “T2” (Fig. S4). The numbers of bHLH TDGs varied across the seven cucurbit crops, with more bHLH TDGs in watermelon and bottle gourd than in the other five species. The \(K_s\) values for each TDG pair belonging to the same cluster were then calculated to trace their divergence time after duplication (Table S4). In watermelon and bottle gourd, we obtained two (T1 and T5) and three (T4, T5, and T6) gene pairs with \(K_s\) values lower than 0.32, suggesting that these TDGs may emerge recently after the divergence between the two species (Table S4). This might explain the higher number of bHLH TDGs in these two species. Pumpkin, which underwent a recent WGD event, had the smallest number of bHLH TDGs, despite having the largest bHLH gene family among the seven cucurbits (Fig. S4, Table1 and Table S4). Considering WGD events may accelerate the loss of TDGs\(^{24}\), we speculate that the evolution of TDGs in pumpkin might have been affected by the recent WGD event\(^8\).

Furthermore, we investigated the distribution of TDGs in OGs and their potential functions. All TDGs were distributed in ten OGs (Table 1), including four (OG001,
OG002, OG003, and OG004) in which we detected TDG clusters in at least four cucurbit crops. Homologous genes in Arabidopsis for these TDG cluster genes in OG001 and OG002 were reported to be involved in the regulation of iron homeostasis. OG003 included the Bt cluster, which regulates cucurbitacin biosynthesis in different tissues in cucumber, melon and watermelon, whereas the function of genes in OG004 remains to be elucidated in cucurbits.

Table 1 List of orthologous gene group of tandem duplicated genes (TDGs) among seven cucurbit crops

| Orthologous group No. | Melon           | Cucumber        | Watermelon     | Wax gourd     | Bottle gourd | Bitter gourd | Pumpkin    |
|-----------------------|-----------------|-----------------|----------------|--------------|--------------|--------------|------------|
| OG001                 | CmehlH04*       | CsabHLH125*     | CmehlH09*      | BhbHLH041*   | LsibHLH01*   | MchbHLH004* | CmehlH030* |
|                       | CmehlH05*       | CsabHLH126*     | CmehlH010*     | BhbHLH042*   | LsibHLH038*  | MchbHLH005* | CmehlH031* |
|                       | CmehlH06*       | CsabHLH127*     | CmehlH011*     | BhbHLH043*   | LsibHLH009*  | MchbHLH006* | CmehlH087* |
|                       |                 |                 |                |              |              |              |            |
| OG002                 | CmehlH112       | CsabHLH087*     | CmehlH049*     | BhbHLH035*   | LsibHLH013*  | MchbHLH001 | CmehlH024  |
|                       | CmehlH088*      | CmehlH090*      | CmehlH049*     | BhbHLH022*   | LsibHLH014*  | MchbHLH091  |            |
|                       |                 |                 |                |              |              |              |            |
| OG003                 | CmehlH120*      | CsabHLH093*     | CmehlH101*     | BhbHLH135*   | LsibHLH012*  | MchbHLH077* | CmehlH017  |
|                       | CmehlH121*      | CsabHLH094*     | CmehlH042*     | BhbHLH136*   | LsibHLH012*  | MchbHLH078* | CmehlH019* |
|                       | CmehlH122*      | CsabHLH095*     | CmehlH033*     | BhbHLH122*   | LsibHLH013*  | MchbHLH017  | CmehlH019* |
|                       |                 |                 |                |              |              |              |            |
| OG004                 | CmehlH082*      | CsabHLH065*     | CmehlH064*     | BhbHLH008*   | LsibHLH071*  | MchbHLH128  | CmehlH031  |
|                       | CmehlH065*      | CsabHLH065*     | CmehlH009*     | BhbHLH009*   | LsibHLH072*  | MchbHLH197  |            |
|                       |                 |                 |                |              |              |              |            |
| OG005                 | CmehlH138       | CsabHLH111      | CmehlH079*     | BhbHLH145*   | LsibHLH132*  | MchbHLH146  | CmehlH210  |
|                       |                 |                 |                |              |              |              |            |
| OG006                 | CmehlH037       | CsabHLH117      | CmehlH113      | BhbHLH058*   | LsibHLH109*  | MchbHLH088  | CmehlH054  |
|                       |                 |                 |                |              |              |              |            |
| OG007                 | CmehlH079       | CsabHLH061      | CmehlH061      | BhbHLH005*   | LsibHLH110*  | MchbHLH148  | CmehlH148  |
|                       |                 |                 |                |              |              |              |            |
| OG008                 | CmehlH039       | CsabHLH049      | CmehlH115      | BhbHLH057*   | LsibHLH113*  | MchbHLH091* | CmehlH057  |
|                       |                 |                 |                |              |              |              |            |
| OG009                 | CmehlH066       | CsabHLH036      | CmehlH150      | BhbHLH083*   | LsibHLH148*  | MchbHLH0105*| No gene    |
|                       |                 |                 |                |              |              |              |            |
| OG010                 | CmehlH141       | CsabHLH016      | CmehlH039      | BhbHLH148*   | LsibHLH025*  | No gene     | CmehlH224* |
|                       |                 |                 |                |              |              |              |            |

Number of TDGs (clusters): 9 (three) 10 (four) 14 (five) 10 (four) 15 (seven) 10 (four) 8 (four)

Notes: Asterisk (*) mark the tandem duplicated genes.
Evolution and divergence of the Bt cluster

In cucumber, melon and watermelon, the Bt cluster regulates cucurbitacin biosynthesis in different tissues.\textsuperscript{17,18} To explore the evolution and divergence of this cluster across more cucurbit crops, we performed a comprehensive comparative genomic analysis by assessing local synteny of paralogous or orthologous genes, as well as constructing a phylogenetic tree and analyzing gene expression patterns. Syntenic paralogous gene analysis detected a paralogous cluster (OG004 in Table 1) of the Bt cluster in cucurbit crops (Fig. 3a, Fig. S5 and Table S5). The $K_s$ values of each paralogous gene pair mapping to the Bt cluster were much smaller than those of all gene pairs between clusters (Fig. 3b and Fig. S6). The phylogenetic tree showed that the evolutionary distance between genes in the Bt cluster is closer than that of genes between clusters (Fig. 3b). These results suggested that the Bt cluster in cucurbits might have arisen after the CCT event\textsuperscript{7} (Fig. 3b and Fig. S6). By local syntenic orthologous gene analysis, we found that the Bt cluster is collinearly distributed among the Cucurbitaceae crops, as is its paralogous cluster (Fig. 3c and 3d). In the paralogous cluster, the number of genes appear to be conserved, with two genes in each species (Fig. 3d). However, the gene numbers in the Bt cluster were much more flexible, with three genes in cucumber, five genes in melon, five genes in watermelon, four genes in bottle gourd and two genes in wax gourd in the Bt cluster (Fig. 3c and Fig. S6). To infer the number of ancestral genes in the Bt cluster and their evolution, we divided $bHLH$ genes from the Bt cluster and its paralogous cluster from the seven cucurbit species into five major clades (Clade I–V) based on the phylogenetic tree (Fig. 3e). The genes from the Bt cluster were distributed to Clades I, II, and III (Fig. 3e), suggesting three ancestral genes in the Bt cluster after the CCT event. In cucumber, three Bt cluster genes remained in Clades I, II, and III. However, new members appeared in Clade I for watermelon, melon and bottle gourd after
species diversification, based on the intraspecies and interspecies $K_s$ values and phylogenetic analysis (Fig. 3e, Fig. S6 and S7). For example, the $K_s$ values of each pair among CmebHLH121, CmebHLH122 and CmebHLH123 in the CmeBt cluster were smaller than interspecies $K_s$ values, suggesting that these three genes arose in melon after its divergence from the other species (Fig. S6 and S7). In addition, several members in Clade II or Clade III were lost in some cucurbit crops: wax gourd lacked a gene in Clade III (Fig. 3e). Finally, we inferred the evolutionary correspondence of Bt cluster genes among the seven cucurbit species (Table S6).

The expression patterns of genes in the Bt cluster and its paralogous cluster showed high degree of tissue specificity. For example, Cla400, BhibHLH135, LsibHLH120, CsabHLH095 (CsBt), Cme003 and MchbHLH077 were predominantly expressed in fruits, while CsabHLH093 (CsBl) exhibited a leaf-specific expression pattern and CmabHLH017, ClabHLH004, LsibHLH123, CsabHLH064, and LsibHLH071 were mainly expressed in roots (Fig. 3f). These expression patterns imply their potential functions of the respective bHLH genes in specific tissues, echoing the expression of CsBt and CsBl in cucumber fruits and leaves and their role in regulating cucurbitacin biosynthesis, respectively\(^\text{17,18}\). Notably, most genes in Clade IV were highly expressed in roots (Fig. 3f), suggesting their potential roles in regulating cucurbitacin biosynthesis in this tissue.
Fig. 3 Evolution and divergence of the Bt cluster. a The syntenic block between the Bt cluster and its paralogous cluster in cucumber. The magenta numbers are $K_s$ values. b Evolution of the Bt cluster and its paralogous cluster in cucumber. The magenta numbers are $K_s$ values. c, d Syntenic regions of the Bt cluster (c) and its paralogous cluster (d) among cucumber (Csa) melon (Cme) and watermelon (Cla). e, f A gene tree (e) and expression heat maps (f) of genes in Bt cluster and its paralogs. Grey box indicates missing value. The magenta ‘W’ means genes only expressing in wild material fruits. The magenta abbreviations are the same as reference 18. Triangles mean premature translational termination.
The CDS of CmBt and CmBt' are identical. One gene in the Bt paralogous cluster, exhibiting high expression in the roots of cucurbit crops, is designated Brp gene.

A novel functional bHLH gene of the Bt paralogous cluster regulates cucurbitacin biosynthesis in roots

We found that one gene in the Bt paralogous cluster of each cucurbit crop, belonging to Clade IV, exhibited a high-level expression in roots (designated Brp, Fig. 3e and f). Given its evolutionary history, we hypothesized that Brp might be involved in regulating the biosynthesis of cucurbitacins in the roots of cucurbit crops. To explore the function of Brp in cucurbits, we conducted heterologous transient expression experiments in Nicotiana benthamiana leaves, using a reporter construct consisting of the promoter from the Bitter (Bi) locus of melon and watermelon, driving the expression of the firefly luciferase (LUC) gene. Here Bi gene encodes oxidosqualene cyclase which catalyzes the first step of cucurbitacin biosynthesis (Fig. S8). Compared to the control, the normalized transcriptional activity (LUC/REN) of proBi:LUC was significantly promoted by transient overexpression of Brp from melon (CmebHLH082) and watermelon (ClabHLH064) (Fig. 4a–c).

To better understand the underlying function of Brp in cucurbit crops, we generated CmeBrp-overexpression (CmeBrp-OE) transgenic hairy melon roots with CmeBrp expression driven by the constitutive cauliflower mosaic virus (CaMV) 35S promoter (Fig. 4d). We also produced transgenic hairy melon roots of CRISPR/Cas9 (cr)-mediated CmeBrp knockout mutants (CmeBrp-cr) (Fig. 4d and 4e). We then quantified cucurbitacin B (CuB) contents in hairy roots of the wild-type (WT), CmeBrp-cr and CmeBrp-OE lines, which revealed a drastic increase in CuB contents in CmeBrp-OE lines and a significant decrease in CmeBrp-cr lines compared to the WT (Fig. 4f and Fig. S9). RT-qPCR analysis indicated that the expression of CmeBrp and four key biosynthetic genes of CuB (CmeBi, Cme890, Cme180 and CmeACT, Fig.
were markedly upregulated in the CmeBrp-OE lines compared to the WT (Fig. 4g-k), whereas the expression of four key biosynthetic genes of CuB were substantially lower in the CmeBrp-cr lines relative to the WT (Fig. 4h-k). In summary, these results suggest that CmeBrp may regulate the biosynthesis of cucurbitacin B in melon roots.
**Fig. 4** **Brp** participates in regulating cucurbitacin biosynthesis. a Schematic diagram shows the constructs used in the transient transcriptional activity assays. b, c **Brp** activates **Bi** expression in a tobacco leaf assay system for melon (b) and watermelon (c). Values are denoted as mean ± standard deviation (SD) from four biological replicates. d GFP signals are detected in hairy roots of CmeBrp over-expression and CmeBrp knockout lines. Bar=1mm. e Sequence analysis of CmeBrp in two knockout lines. A total of 25 colonies are sequenced in CmeBrp-cr lines. The mutant types and number are showed at right. f GFP signal are detected in hairy roots of CmeBrp over-expression and CmeBrp knockout lines. Bar=1mm. e Sequence analysis of CmeBrp in two knockout lines. A total of 25 colonies are sequenced in CmeBrp-cr lines. The mutant types and number are showed at right. f Content of cucurbitacin B in WT, CmeBrp knockout and CmeBrp over-expression lines. g Relative expression of CmeBrp in WT, CmeBrp knockout and CmeBrp over-expression lines. h-k Relative expression of four key CuB biosynthetic genes, CmeBi (h), Cme890 (i), Cme180 (j) and CmeACT (k), in WT, CmeBrp knockout and CmeBrp over-expression lines. Transcript levels are measured by Real-time quantitative polymerase chain reaction (RT-qPCR). Values are represented as mean ± SD from three biological replicates. Statistical significance is calculated using Student’s t-test: **, P < 0.01

**Discussion**

Comparative genomics analysis can be powerful approaches to provide novel insights into gene function and evolution. In this study, we performed comparative genomics analysis of **bHLH** genes in seven cucurbit species and inferred evolution and divergence of the **Bt** cluster. Together with evolutionary relationships and expression profiles, these results empowered the discovery of a novel gene regulating cucurbitacin biosynthesis. The methodology applied in this research presents an example of how gene family analysis facilitates functional gene studies.

Tandem duplications are a widespread phenomenon in plant genomes and play significant roles in evolution and adaptation to changing environments. Compared to WGD-derived duplicate genes, TDGs provide a continuous supply of variants available for adaptation to continuously changing environments. Even though the number of **bHLH** genes is the highest in pumpkin compared to that in other cucurbit crops, it has the fewest **bHLH** TDGs. Both TDGs and WGDs can lead to the expansion of gene families and increase gene diversity. During pumpkin evolution,
WGD may have provided additional genetic material and increased gene diversity, and thus the fraction of lost tandem arrays was significantly larger than that of lost non-tandem genes\textsuperscript{24}. Therefore, we hypothesize that TDGs in pumpkin were dramatically lost after the recent WGD, which might be resulted from gene dosage effect.

Compared to the metabolic biosynthetic gene clusters\textsuperscript{36-39}, the transcription factor (TF) clusters, governing these metabolic enzymes, are less identified and characterized\textsuperscript{40}. Unlike the metabolic biosynthetic clusters that include genes encoding various classes of metabolic enzymes\textsuperscript{41,42}, TFs in the same cluster may have overlapping or distinct regulatory functions. In Madagascar periwinkle (\textit{Catharanthus roseus}), three clustered \textit{bHLH} genes (named \textit{BIS1} [\textit{bHLH iridoid synthesis 1}], \textit{BIS2}, and \textit{BIS3}) regulate iridoid biosynthesis in the terpenoid indole alkaloid biosynthetic pathway\textsuperscript{43-45}. However, the origin, copy number, and evolution of TF clusters are largely unknown\textsuperscript{45}. In this study, we determined that gene number within the \textit{Bt} cluster is variable across species, and a novel \textit{bHLH} cluster consisting of two genes that is relatively conserved and paralogous to the \textit{Bt} cluster, also regulates cucurbitacin biosynthesis. These \textit{bHLH} genes belonged to 15(Ib3) subfamily (Fig. S1). Although the function of \textit{AtbHLH}s belonging to 15(Ib3) subfamily were unknown in Arabidopsis, these \textit{AtbHLH}s clustered with \textit{Brps} into a clade and highly expressed in roots. Therefore, we speculated that the \textit{Bt} paralogous cluster is an ancestral cluster, as evidenced by the higher \textit{Ks} values of each pair genes in the \textit{Bt} paralogous cluster (Fig. S3b and Fig. S6). These findings will help understand the function and evolution of TF clusters.

Materials and Methods

Data collection and identification of \textit{bHLH} genes in seven cucurbit crops
Melon genome sequence data was obtained from the Melonomics database (www.melonomics.net), bitter gourd genome sequence data was obtained from the https://db.cngb.org/search/assembly/CNA0000004/, and the other five cucurbit crops (cucumber, watermelon, wax gourd, bottle gourd and pumpkin) genome sequence data were obtained from the http://cucurbitgenomics.org. The information and sequences of *A. thaliana* bHLHs (*AtbHLHs*) were retrieved from https://www.arabidopsis.org. The bHLH proteins of seven cucurbit crops (*CsabHLHs*, *CmebHLHs*, *ClabHLHs*, *BhibHLHs*, *LsibHLHs*, *CmabHLHs* and *MchbHLHs*) were predicted using the HLH hidden Markov model (HMM) profile obtained from Pfam (http://pfam.xfam.org, PF00010) and used as queries to search the bHLH proteins from each cucurbit crop sequences with HMMER (version 3.1b2) software (http://hmmer.janelia.org). We also performed a BLASTP search against the AtbHLHs database. Redundant protein sequences were removed by searching in the NCBI database (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) and the SMART database (http://smart.embl-heidelberg.de). In addition, some bHLH genes were re-annotated, due to error existing in raw annotation, by ab initio prediction, transcript mapping and evidence from other version reference genome.

**Multiple alignments and Phylogenetic analysis**

Multiple sequence alignments of identified bHLH domains of 162 Arabidopsis (*AtbHLHs*), 149 cucumber (*CsabHLH*), 151 melon (*CmebHLH*), 154 watermelon (*ClabHLH*), 150 wax gourd (*BhibHLH*), 155 botter gourd (*LsibHLH*), 251 pumpkin (*CmabHLH*) and 150 bitter gourd (*MchbHLH*) bHLH proteins were carried out by muscle software with default parameters. To visualize the conserved motifs, the sequences were analyzed with WEBLOGO programs (http://weblogo.berkeley.edu) based on result of each cucurbit crop multiple sequence alignment. Based on result of all seven cucurbit crops and Arabidopsis multiple sequence alignment, a neighbor-joining tree was constructed using MEGA 7.048 using a bootstrap test with
1000 replicates based on Jones-Taylor-Thornton (JTT) model and 80% Partial deletion for gap treatment. The phylogenetic tree was visualized in MEGA 7.0.

**DNA-binding ability analysis**

Firstly, based on multiple sequence alignments of bHLH domains proteins, we identified conserved amino acid residues in the bHLH domains and four conserved motifs including one basic, two amphipathic α-helices and one loop which linked the two amphipathic α-helices (Figure S10 and Table S7). Then we defined proteins with more than five basic amino acid residues in the basic region as DNA-binding proteins based on the criteria developed by Toledo-Ortiz et al. The bHLH proteins were divided into a group of DNA-binding proteins and a group of non-DNA-binding proteins. Furthermore, the DNA-binding proteins were subdivided into two subcategories, E-box (based on the presence of E-2 and R-4) and non-E-box-binding proteins (without the simultaneous presence of E-2 and R-4). Then E-box-binding proteins were further subdivided into two groups, including G-box-binding proteins (H/K-1, E-2, and R-5 were required) and non-G-box-binding proteins. Consensus amino acids and Number, such as the E-2, were described in Table S7.

**Gene structure analysis and protein motif detection**

The exon/intron organization and splicing phase of the predicted bHLHs were investigated based on the GFF/GTF annotation files of seven cucurbit crop genomes. Furthermore, to discover the intron distribution pattern, we did alignment analysis with the coding sequence of the bHLH domain and genome sequences using Blat software. Then exon/intron structures, splicing phase and different regions of bHLH domain graphically displayed by the Gene Structure Display Server (GSDS, http://gsds.cbi.pku.edu.cn/). To analyze other conserved motifs in complete amino acid sequence of bHLHs, the protein sequences of candidate bHLHs were analyzed using MEME (Version 5.1.1) software (http://meme-suite.org/tools/meme). The
parameters settings are: number of motifs to find, 15; minimum width of motifs, 6; and maximum width of motifs, 50.

Identification of ortholog genes among seven cucurbit crops and paralog genes in each cucurbit crop

Gene syntenic analysis by MCscanX (http://chibba.agtec.uga.edu/duplication/mcscan/) with default parameters. Based on the syntenic blocks, each inter-genomic bHLH orthologous genes among seven cucurbit crops were identified. Then we merge the orthologous gene pairs information by python script. Orthologous genes visualized image was generated using jcvi (https://github.com/tanghaibao/jcvi). Similarly, intra-genomic bHLH paralogous genes were identified. The paralogous genes graphically displayed by circos figure.

Chromosomal locations and tandem duplicated gene detection.

The chromosomal positions of bHLH loci were obtained from each cucurbit crop GFF/GTF annotation files. The distribution of bHLH genes on chromosomes in each cucurbit crop were drawn using TBtools software. Tandem genes were detected using SynOrths and MCscanX (http://chibba.agtec.uga.edu/duplication/mcscan/) softwares with default parameters.

Ka, Ks and Ka/Ks analysis

The non-synonymous substitutions(Ka), synonymous substitutions (Ks) and Ka/Ks values of paralogous gene pairs and tandem array genes (between any two genes in a same tandem duplicated gene cluster) were calculated using KaKs Calculator.

Gene expression analyses of bHLH genes in seven cucurbit crops
RNA-seq data for the *bHLH* genes were obtained from previous studies of differential gene expression in organs and tissues in seven cucurbit crops\(^8,27-32\). All clean RNA-seq reads, from each sample, were mapped onto the corresponding genome sequences, using TopHat2\(^54\) (version 2.1.0) with default parameters. The generated BAM format alignments, together with the gene GTF annotation file, the fragments per kilobase of exon model per million reads mapped (FPKM) values were computed using stringtie\(^54\) (v1.3.4d). Finally, the log\(_2\)-transformed (FPKM+1) values were used to generate heat map by R.

**Dual-Luciferase (Dual-LUC) Assay**

The full-length coding sequences (CDS) of *Brp* was inserted into the pBI121 plasmid to generate *Brp*-pBI121 effector, while the 2000-bp sequences upstream of the translation initiation start sites of *Bi* gene was cloned into the pGreen II 0800-LUC to generate p*Bi*-LUC double-reporter vector. The *Brp*-pBI121 effector and p*Bi*-LUC reporter vector were transformed into *Agrobacterium tumefaciens* strain GV3101 and GV3101(pSoup-p19), respectively. The reporter and effector were co-infiltrated into *N. benthamiana* leaves by a volume ratio of 9:1. The empty pBI121 vector was used as control. The leaf samples were collected within 60 hours after injection for measuring luciferase activities using the Dual-Luciferase Reporter Assay System (Promega, United States, E1910) by GloMax 20/20 Luminometer (Promega, United States) according to the manufacturer’s instructions (Promega, United States). The relative reporter gene expression levels were expressed as the ratio of Firefly luciferase to Renilla luciferase (LUC/REN). Four independent transformations for each sample were performed.

**Agrobacterium rhizogenes-mediated hairy root transgenic system in melon**

To construct *CmeBrp* over-expression vector, the full-length coding sequences (CDS) of *CmeBrp* was cloned into the pCAMBIA1305.4 binary vector, using an In-Fusion
HD Cloning Kit (Clontech); To generate CRISPR/Cas9 editing vector, the 19-bp sgRNA fragment of CmeBrp was assembled into the vector pBSE402 by using the golden gate cloning method. The individual read frame green fluorescent protein (GFP) was used as a reporter gene by the constitutive cauliflower mosaic virus (CaMV) 35S promoter in CmeBrp-OE and CmeBrp-cr vectors. These vectors were transformed into Agrobacterium rhizogenes strain Ar. Qual (CAT#: AC1060).

Melon peeled seeds were sterilized by 75% (v/v) ethanol for 30 s, followed by 0.3% (v/v) sodium hypochlorite solution for 15 min. The sterilized seeds were germinated on MS30 medium at 28°C for two days in darkness before seedling were grown at 25°C for one week under a 16 h light/ 8 h dark photoperiod until cotyledon fully expanded. Cotyledons were cut off at the basal and tip end, and soaked in diluted A. rhizogenes strain containing the binary vector at 28°C for 20 min, then co-cultivated on MS solid medium for two days in darkness at 23°C. To regenerate roots, explants were transferred to MS solid medium containing 100 mg/L Timentin for two weeks at 25°C under a 16 h light/ 8 h dark photoperiod. The positive roots with GFP fluorescence at approx. 2 cm length were collected for subsequent experiments. In order to meet the dosage and consistency of samples in subsequent experiments, we had to mix 4-5 roots as a single biological replication.

**UPLC analysis of cucurbitacins B from melon**

The ultra-performance liquid chromatography (UPLC) analysis of cucurbitacins B as described previously by Zhou et al., (2016). Samples were flash frozen in liquid nitrogen and ground to powder. The resultant powder (0.1 g) was added to methanol (1 mL) and homogenized for 15 min, followed by centrifugation at 10,000 g at 4°C for 10 min. The solution was filtered through 0.22 μm membrane prior to injection and then analyzed on a HPLC system.

**Real-time quantitative PCR**
Total RNA was isolated from hairy roots of wild type, *CmeBrp-cr* and *CmeBrp-OE* lines using a Quick RNA isolation Kit (Huayueyang), and samples of 1 μg were reverse-transcribed using a GoScript™ Reverse Transcription Mix, Oligo(dT) (Promega, A2791, USA) according to the instruction manual. Then a quantitative PCR assay was performed on an ABI 7900 (Applied Biosystems) machine using GoTaq® qPCR Master Mix Kit (Promega, A6001, USA) according to the manufacturer’s instructions. Three independent biological replicates were performed. Relative gene expression was performed using the $2^{-\Delta\Delta CT}$ method and the *ubiquitin* gene (*MELO3C009513*) was as the reference gene$^{55}$. Primers are listed in Supplemental Table S8.

**Acknowledgements**

This work was supported by National Natural Science Foundation of China (32130093 to Z.Z., 31772304 to Z.Z.) and the National Key R&D Program of China (2016YFD0100307). This work was also supported by the “Taishan Scholar” Foundation of the People’s Government of Shandong Province and the Chinese Academy of Agricultural Science (ASTIP-CAAS and CAAS-XTCX2016001), the Shenzhen municipal (The Peacock Plan KQTD2016113010482651), and Central Public-interest Scientific Institution Basal Research Fund (No. Y2017PT52).

**Conflict of interests**

The authors declare that they have no conflict of interest.

**Author contributions**

YX participated in design of the research, performed the data analysis, and wrote the manuscript. HZ designed the experiments. YX, YZ and NJ performed the experiments. XZ and QZ collected genome and RNA sequencing data. HL and SC revised the
manuscript. ZZ conceived, designed the research and revised the manuscript. All authors read and approved the manuscript.

Availability of data

All data supporting the results of this article are included within the article and its additional files.

Supplementary information

Additional Supporting Information may be found in the online version of this article.

**Figure S1.** Phylogenetic relationships and characters information of *bHLHs* in seven cucurbits and Arabidopsis.

**Figure S2.** Paralogous gene pairs in cucumber, melon, watermelon, bottle gourd, wax gourd, bitter gourd and pumpkin.

**Figure S3.** The *Ka/Ks* values of paralogous gene pairs in seven cucurbits.

**Figure S4.** Chromosomal localizations and tandem duplicatation of *bHLH* genes.

**Figure S5.** Syntenic block between *Bt* cluster and its paralogous cluster.

**Figure S6.** *Ks* values between (among) *Bt* cluster and its paralogous cluster genes.

**Figure S7.** *Ks* values of interspecies for *Bt* cluster genes.

**Figure S8.** Biosynthetic pathway of cucurbitacin B and its regulating genes in melon.

**Figure S9.** The ultra-performance liquid chromatography (UPLC) analysis of CuB in hairy roots of WT, *CmeBrp* knockout and *CmeBrp* over-expression lines.

**Figure S10.** Conserved amino acid analysis of bHLH domains.

**Table S1.** List of *bHLH* genes in seven cucurbit crops.

**Table S2.** List of orthologous groups among seven cucurbit crops.

**Table S3.** List of paralogous gene pairs in each cucurbit crop.

**Table S4.** *Ka, Ks* and *Ka/Ks* value of each pair TDGs in a same TDGs cluster.

**Table S5.** Evidence of the paralogous cluster of the *Bt* cluster in wax gourd.

**Table S6.** The relationships of *Bt* cluster and its paralogous genes among seven cucurbit species.
Table S7. Conserved amino acids of bHLHs in seven cucurbit crops.

Table S8. Primers used in this study.

References

1 Riechmann, J. L. et al. Arabidopsis Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes. Science 290, 2105 (2000).

2 Feller, A., Machemer, K., Braun, E. L. & Grotewold, E. Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. Plant J 66, 94-116, doi:10.1111/j.1365-313X.2010.04459.x (2011).

3 Carretero-Paulet, L. et al. Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in Arabidopsis, poplar, rice, moss, and algae. Plant Physiol 153, 1398-1412, doi:10.1104/pp.110.153593 (2010).

4 Pires, N. & Dolan, L. Origin and diversification of basic-helix-loop-helix proteins in plants. Mol Biol Evol 27, 862-874, doi:10.1093/molbev/msp288 (2010).

5 Toledo-Ortiz, G., Huq, E. & Quail, P. H. The Arabidopsis basic/helix-loop-helix transcription factor family. Plant Cell 15, 1749-1770, doi:10.1105/tpc.013839 (2003).

6 Li, X. et al. Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. Plant Physiol 141, 1167-1184, doi:10.1104/pp.106.080580
Wang, J. et al. An Overlooked Paleotetraploidization in Cucurbitaceae. *Mol Biol Evol* **35**, 16-26, doi:10.1093/molbev/msx242 (2018).

Sun, H. et al. Karyotype Stability and Unbiased Fractionation in the Paleo-Allotetraploid Cucurbita Genomes. *Mol Plant* **10**, 1293-1306, doi:10.1016/j.molp.2017.09.003 (2017).

Dong, H. et al. Genome-wide association studies reveal that members of bHLH subfamily 16 share a conserved function in regulating flag leaf angle in rice (*Oryza sativa*). *PLoS Genet* **14**, e1007323, doi:10.1371/journal.pgen.1007323 (2018).

Yang, F., Wang, Q., Schmitz, G., Muller, D. & Theres, K. The bHLH protein ROX acts in concert with RAX1 and LAS to modulate axillary meristem formation in Arabidopsis. *Plant J* **71**, 61-70, doi:10.1111/tpj.1365-313X.2012.04970.x (2012).

Paik, I., Kathare, P. K., Kim, J.-I. & Huq, E. Expanding Roles of PIFs in Signal Integration from Multiple Processes. *Mol Plant* **10**, 1035-1046, doi:10.1016/j.molp.2017.07.002 (2017).

Radoeva, T. et al. A Robust Auxin Response Network Controls Embryo and Suspensor Development through a Basic Helix Loop Helix Transcriptional Module. *The Plant Cell* **31**, 52-67, doi:10.1105/tpc.18.00518 (2019).

Ren, Y., Zhao, Q., Zhao, X., Hao, Y. & You, C. Expression Analysis of the MdClbHLH1 Gene in Apple Flower Buds and Seeds in the Process of Dormancy. *Horticultural Plant Journal* **2**, 61-66, doi:10.1016/j.hppj.2016.06.007 (2016).
Zhang, H. B., Bokowiec, M. T., Rushton, P. J., Han, S. C. & Timko, M. P. Tobacco transcription factors NtMYC2a and NtMYC2b form nuclear complexes with the NtJAZ1 repressor and regulate multiple jasmonate-inducible steps in nicotine biosynthesis. *Mol Plant* **5**, 73-84, doi:10.1093/mp/ssr056 (2012).

Shen, Q. *et al.* The jasmonate-responsive AaMYC2 transcription factor positively regulates artemisinin biosynthesis in Artemisia annua. *New Phytol* **210**, 1269-1281, doi:10.1111/nph.13874 (2016).

Ribeiro, B. *et al.* A Seed-Specific Regulator of Triterpene Saponin Biosynthesis in *Medicago truncatula*. *Plant Cell* **32**, 2020-2042, doi:10.1105/tpc.19.00609 (2020).

Shang, Y. *et al.* Biosynthesis, regulation, and domestication of bitterness in cucumber. *Science* **346**, 1084-1088, doi:10.1126/science.1259215 (2014).

Zhou, Y. *et al.* Convergence and divergence of bitterness biosynthesis and regulation in Cucurbitaceae. *Nat Plants* **2**, 16183, doi:10.1038/nplants.2016.183 (2016).

Xi, H., He, Y. & Chen, H. Functional Characterization of SmbHLH13 in Anthocyanin Biosynthesis and Flowering in Eggplant. *Horticultural Plant Journal* **7**, 73-80, doi:10.1016/j.hpj.2020.08.006 (2021).

Murre, M. E. M. a. c. Helix-loop-helix proteins: regulators of transcription in eukaryotic organisms. *Mol Cell Biol*, 429-440 (2000).

Sun, H., Fan, H. J. & Ling, H. Q. Genome-wide identification and characterization of the bHLH gene family in tomato. *BMC Genomics* **16**, 9, doi:10.1186/s12864-014-1209-2 (2015).
Heim, M. A. et al. The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. *Mol Biol Evol* **20**, 735-747, doi:10.1093/molbev/msg088 (2003).

Kane, J., Freeling, M. & Lyons, E. The evolution of a high copy gene array in *Arabidopsis*. *J Mol Evol* **70**, 531-544, doi:10.1007/s00239-010-9350-2 (2010).

Fang, L., Cheng, F., Wu, J. & Wang, X. The Impact of Genome Triplication on Tandem Gene Evolution in *Brassica rapa*. *Front Plant Sci* **3**, 261, doi:10.3389/fpls.2012.00261 (2012).

Wang, N. et al. Requirement and functional redundancy of Ib subgroup bHLH proteins for iron deficiency responses and uptake in *Arabidopsis thaliana*. *Mol Plant* **6**, 503-513, doi:10.1093/mp/sss089 (2013).

Cui, Y. et al. Four IVa bHLH Transcription Factors Are Novel Interactors of FIT and Mediate JA Inhibition of Iron Uptake in *Arabidopsis*. *Mol Plant* **11**, 1166-1183, doi:10.1016/j.molp.2018.06.005 (2018).

Li, Z. et al. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics* **12**, 540, doi:10.1186/1471-2164-12-540 (2011).

Yano, R., Nonaka, S. & Ezura, H. Melonet-DB, a Grand RNA-Seq Gene Expression Atlas in Melon (*Cucumis melo* L.). *Plant Cell Physiol* **59**, e4, doi:10.1093/pcp/pcx193 (2018).

Guo, S. et al. Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat Genet* **51**, 1616-1623.
Wu, S. et al. The bottle gourd genome provides insights into Cucurbitaceae evolution and facilitates mapping of a Papaya ring-spot virus resistance locus. Plant J 92, 963-975, doi:10.1111/tpj.13722 (2017).

Jiang, B., Xie, D., Liu, W., Peng, Q. & He, X. De novo assembly and characterization of the transcriptome, and development of SSR markers in wax gourd (Benicaca hispida). PLoS One 8, e71054, doi:10.1371/journal.pone.0071054 (2013).

Cui, J. et al. Whole-genome sequencing provides insights into the genetic diversity and domestication of bitter gourd (Momordica spp.). Hortic Res 7, 85, doi:10.1038/s41438-020-0305-5 (2020).

Varaud, E. et al. AUXIN RESPONSE FACTOR8 regulates Arabidopsis petal growth by interacting with the bHLH transcription factor BIGPETALp. Plant Cell 23, 973-983, doi:10.1105/tpc.110.081653 (2011).

Yu, J. et al. PTGBase: an integrated database to study tandem duplicated genes in plants. Database (Oxford) 2015, doi:10.1093/database/bav017 (2015).

Qi, X. et al. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. Genome Biol 20, 38, doi:10.1186/s13059-019-1650-2 (2019).

Frey, M. et al. Analysis of a chemical plant defense mechanism in grasses. Science 277, 696-699, doi:10.1126/science.277.5326.696 (1997).

Qi, X. et al. A gene cluster for secondary metabolism in oat: implications for the
evolution of metabolic diversity in plants. *Proc Natl Acad Sci U S A* **101**, 8233-8238, doi:10.1073/pnas.0401301101 (2004).

38 Hen-Avivi, S. *et al.* A Metabolic Gene Cluster in the Wheat W1 and the Barley Cer-cqu Loci Determines beta-Diketone Biosynthesis and Glaucousness. *Plant Cell* **28**, 1440-1460, doi:10.1105/tpc.16.00197 (2016).

39 Winzer, T. *et al.* A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science* **336**, 1704-1708, doi:10.1126/science.1220757 (2012).

40 Shoji, T. & Yuan, L. ERF Gene Clusters: Working Together to Regulate Metabolism. *Trends Plant Sci* **26**, 23-32, doi:10.1016/j.tplants.2020.07.015 (2021).

41 Nutzmann, H. W., Huang, A. & Osbourn, A. Plant metabolic clusters - from genetics to genomics. *New Phytol* **211**, 771-789, doi:10.1111/nph.13981 (2016).

42 Nutzmann, H. W. *et al.* Active and repressed biosynthetic gene clusters have spatially distinct chromosome states. *Proc Natl Acad Sci U S A* **117**, 13800-13809, doi:10.1073/pnas.1920474117 (2020).

43 Van Moerkercke, A. *et al.* The bHLH transcription factor *BIS1* controls the iridoid branch of the monoterpenoid indole alkaloid pathway in *Catharanthus roseus*. *Proc Natl Acad Sci U S A* **112**, 8130-8135, doi:10.1073/pnas.1504951112 (2015).

44 Van Moerkercke, A. *et al.* The basic helix-loop-helix transcription factor *BIS2* is essential for monoterpenoid indole alkaloid production in the medicinal plant *Catharanthus roseus*. *Plant J* **88**, 3-12, doi:10.1111/tpj.13230 (2016).

45 Singh, S. K. *et al.* BHLH IRIDOID SYNTHESIS 3 is a member of a bHLH gene cluster
regulating terpenoid indole alkaloid biosynthesis in Catharanthus roseus. *Plant Direct* **5**, e00305, doi:10.1002/pld3.305 (2021).

46 Maja, D., Mavengahama, S. & Mashilo, J. Cucurbitacin biosynthesis in cucurbit crops, their pharmaceutical value and agricultural application for management of biotic and abiotic stress: A review. *South African Journal of Botany*, doi:10.1016/j.sajb.2021.08.044 (2021).

47 Polturak, G. & Osbourn, A. The emerging role of biosynthetic gene clusters in plant defense and plant interactions. *PLoS Pathog* **17**, e1009698, doi:10.1371/journal.ppat.1009698 (2021).

48 Kumar, S., Stecher, G. & Tamura, K. MEGAS: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol* **33**, 1870-1874, doi:10.1093/molbev/msw054 (2016).

49 Hu, B. *et al.* GSDS 2.0: an upgraded gene feature visualization server. *Bioinformatics* **31**, 1296-1297, doi:10.1093/bioinformatics/btu817 (2015).

50 Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **34**, W369-373, doi:10.1093/nar/gkl198 (2006).

51 Chen, C. *et al.* TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol Plant* **13**, 1194-1202, doi:10.1016/j.molp.2020.06.009 (2020).

52 Cheng, F., Wu, J., Fang, L. & Wang, X. Syntenic gene analysis between *Brassica rapa* and other Brassicaceae species. *Front Plant Sci* **3**, 198,
53 Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics* **8**, 77-80, doi:10.1016/S1672-0229(10)60008-3 (2010).

54 Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc* **11**, 1650-1667, doi:10.1038/nprot.2016.095 (2016).

55 Wang, S. *et al.* A Rare SNP Identified a TCP Transcription Factor Essential for Tendril Development in Cucumber. *Mol Plant* **8**, 1795-1808, doi:10.1016/j.molp.2015.10.005 (2015).