GMR

Identification and cloning of *GOLDEN2-LIKE1* (*GLK1*), a transcription factor associated with chloroplast development in *Brassica napus* L.

Y.L. Pan[1]*, Y. Pan[1]*, C.M. Qu[2], C.G. Su[1], J.H. Li[1] and X.G. Zhang[1]

[1]Key Laboratory of Horticulture Science for Southern Mountainous Regions, Ministry of Education, College of Horticulture and Landscape Architecture, Southwest University, Chongqing, China
[2]Engineering Research Center of South Upland Agriculture, College of Agronomy and Biotechnology, Southwest University, Chongqing, China

*These authors contributed equally to this study.
Corresponding author: X.G. Zhang
E-mail: Zhangdupian@swu.edu.cn

Genet. Mol. Res. 16 (1): gmr16018942
Received July 4, 2016
Accepted December 19, 2016
Published February 16, 2017
DOI http://dx.doi.org/10.4238/gmr16018942

ABSTRACT. Photosynthesis is the process by which dry matter accumulates, which affects rapeseed yield. In this study, we identified *GOLDEN2-LIKE1* (*GLK1*), located on chromosome A07 and 59.2 kb away from the single nucleotide polymorphism marker SNP16353A07, which encodes a transcription factor associated with the rate of photosynthesis in leaves. We then identified 96 *GLK1* family members from 53 species using a hidden Markov model (HMM) search and found 24 of these genes, which were derived from 17 Brassicaceae species. Phylogenetic analysis showed that 24 Brassicaceae proteins were classified into three subgroups, named the Brassica family, *Adenium arabicum*, and *Arabidopsis*. Using homologous cloning methods, we identified four *BnaGLK1* copies; however, the coding sequences

were shorter than the putative sequences from the reference genome, probably due to splicing errors among the reference genome sequence or different gene copies being present in the different *B. napus* lines. In addition, we found that *BnaGLK1* genes were expressed at higher levels in leaves with more chloroplasts than were present in other leaves. Overexpression of *BnaGLK1a* resulted in darker leaves and siliques than observed in the control, suggesting that *BnaGLK1* might promote chloroplast development to affect the rate of photosynthesis in leaves. These results will help to elucidate the mechanism of chloroplast biogenesis by *GLK1* in *B. napus*.

Key words: *Brassica napus* L.; *GLK1*; Chloroplast development; Transcription factor

INTRODUCTION

Brassica napus L. (AACC, 2n = 38) is one of the most important oilseed crops in the world, particularly in Canada, Europe, and China (Hu et al., 2007). To date, the consumption of vegetable oil has risen sharply in order to meet the demands of population growth and the developing global economy (Ahmad et al., 2015). Therefore, increasing rapeseed oil production is one of the key yield parameters important for rapeseed breeding. Moreover, evidence has shown that photosynthesis is not only associated with yield, but also affects dry matter production and oil content (Hua et al., 2012); however, the regulation and relative contribution of photosynthesis to these outcomes is not well known in rapeseed.

In plants, both nuclear and plastid genomes contribute to photosynthesis (Martin et al., 2002). Plastids are a diverse group of organelles that have essential metabolic and signaling functions throughout the life cycle of the plant. Numerous photosystem reactions are catalyzed by the products of genes present in the plastid genome (Martin et al., 2002; Dyall et al., 2004; Nakamura et al., 2009; Yamori and Shikanai, 2016). In plants, chloroplasts are one of the most important organelles, as they facilitate photosynthesis and thereby provide the chemical energy needed for growth (Nakamura et al., 2009; Liu et al., 2016). However, plastid biogenesis relies on the import of nuclear-encoded plastid proteins. Moreover, many genes encoding products that positively regulate the expression of genes involved in photosynthesis or plastid differentiation have been identified. For example, the *Lhca* and *Lhcb* gene family encodes the light-harvesting chlorophyll a/b-binding proteins LHCI and LHCII, which capture light energy and transfer it to chlorophylls (Chls) in the core reaction centers of photosystems I and II (PSI and PSII) (Kobayashi et al., 2013). chlorophyll b binding proteins of the light-harvesting complexes (LHCB) and the small subunit of ribulose-1,5-bisphosphate carboxylase (SSU) modulate photosynthesis-related nuclear gene expression (Kakizaki et al., 2009), and HY5 and its homologs HYH and HFR1/REP1/RSF1 affect the expression of photosynthesis-related genes and proplastid differentiation (Fairchild et al., 2000; Spiegelman et al., 2000; McCormac and Terry, 2002; Wang et al., 2016), which is the key target of the COP1/SPA complex (Xu et al., 2015). Therefore, subunits of the photosystem reaction center are encoded by multiple genes in plant proplastids.

The GOLDEN2-LIKE (GLK) transcription factors directly influence the transcription of genes related to photosynthesis, and *GLK1* plays a crucial role in normal chloroplast

development in many plant species, including *Arabidopsis thaliana*, *Zea mays* (maize), *Oryza sativa* (rice), *Solanum lycopersicum* (tomato), and the moss *Physcomitrella patens* (Rossini et al., 2001; Waters et al., 2008, 2009; Nakamura et al., 2009; Kobayashi et al., 2013; Nguyen et al., 2014). *GLK* genes belong to the GARP (G, G2; AR, ARR; P, Psr1) superfamily, which contains two homologs of *GLK* genes (*GLK1* and *GLK2*) (Riechmann et al., 2000). In *Arabidopsis*, the pale green phenotype of the *glk1* and *glk2* double mutant is rescued by the transgenic expression of either of the two *GLK* genes (Waters et al., 2008, 2009). *ZmGLK1* is a homolog, and *OsGLK1* is an ortholog, of *GLK* with similar phenotypes, which act as transcriptional regulators of cell type differentiation (Hall et al., 1998; Rossini et al., 2001). Although Chl accumulation in photosynthetic tissues was reduced in response to a loss of GLK activity, overexpression of GLK induced Chl accumulation and chloroplast biogenesis (Waters et al., 2009; Kobayashi et al., 2013; Wang et al., 2013). Moreover, *GLK* genes redundantly promote photosynthetic development in C3 plants, but regulate dimorphic chloroplast differentiation in C4 plants (Wang et al., 2013). In addition, *GLK1* directly regulates the expression of a set of nuclear-encoded genes related to chloroplast function and plastid-encoded genes (Nakamura et al., 2009; Waters et al., 2009), suggesting that GLKs play important roles in chloroplast biogenesis during organ development. However, the precise function and regulatory mechanism of *GLK1* is unclear in *B. napus*.

In this study, we identified the GOLDEN2-LIKE transcription factor (GLK1) in *B. napus* which may be associated with chloroplast development and located in a Quantitative Trait Loci region for leaf photosynthesis rate and related physiological traits (Yan et al., 2015). Hence, we predicted that *GLK1* modulates the photosynthesis rate in rapeseed leaves. Subsequently, we cloned the full-length coding sequence (CDS) of *GLK1* from *B. napus*, and determined the phylogeny of *GLK1* genes in the context of the current plant genome sequence database, to provide insight into the evolutionary trajectory of *GLK1* gene function in plants. In addition, using quantitative real-time PCR (qRT-PCR), we investigated the expression pattern of *BnaGLK1* genes in different organs of four *B. napus* accessions. Additionally, the function of *BnaGLK1* was inferred by its overexpression in *B. napus*. Therefore, our results provide a foundation for the identification of *GKL1* function, and provide an initial genetic construct that can be used to study its effect on photosynthesis in *B. napus*.

MATERIAL AND METHODS

Identification and isolation of candidate genes for *B. napus* photosynthesis

Based on a Recombinant Inbred lines (RIL) population, including 172 lines, several QTLs associated with leaf photosynthesis rates were identified in *B. napus* (Yan et al., 2015). Next, all Single Nucleotide Polymorphism (SNP) markers tightly linked to QTLs were selected and used to identify candidate genes associated with leaf photosynthesis rates. First, the SNP markers were mapped to the *B. napus* reference genome (http://www.genoscope.cns.fr/brassicanapus/data/) (Chalhoub et al., 2014). Then, DNA sequence information of the QTL regions was used to query the *B. napus* database in a BLAST search with 85% overlap and 98% identity. Eventually, candidate genes in the 200-kb flanking regions of the QTLs were predicted according to the *B. napus* CDSs.

Identification of a *GLK1* gene superfamily in plants

Multiple database searches were performed to identify members of the *GLK1* superfamily in a variety of plant species. Genome, proteome, gene prediction, and annotation data from 77 plant species were downloaded from the Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html#!search), PGDD (PLANT GENOME DUPLICATION DATABASE, http://chibba.agtec.uga.edu/duplication/index/files), and NCBI (http://www.ncbi.nlm.nih.gov/) databases. To identify *GLK1* genes and their homologs, BLAST (BLASTN and TBLASTX) searches with an e-value threshold of ≤ 1e-10 were performed in these databases using the *A. thaliana* GLK1 (AtGLK1, AT2G20570) protein sequence as query, which was constructed using Geneious Pro 4.8.5 (http://www.geneious.com; Biomatters, Auckland, New Zealand). Then, the cDNA, genomic DNA, and amino acid sequences corresponding to putative *GLK1* genes were downloaded and filtered. Three-letter acronyms followed by *GLK1* were used as the species gene name, with the first letter indicating genus and the following two letters indicating species (e.g., *AthGLK1* for *A. thaliana GLK1*). Letters and numbers were added after the taxon names to represent individual gene copies (**Table S1**).

Phylogenetic analysis

Total sequences of GLK1 were retrieved and used for phylogenetic tree construction. Multiple alignment analysis was performed using BioEdit 7.0 with shading >80% threshold (Cheng et al., 2016). Using the neighbor-joining (NJ) method, a phylogenetic tree based on GLK1 amino acid sequences was constructed using the biosoftware Geneious Pro 4.8.5. The reliability of the phylogenetic analysis was assessed by bootstrap analysis with 1000 replicates. Finally, the phylogenetic trees were visualized using FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Determination of gene structure, transmembrane domains, and conserved motifs

By comparing the CDSs and their corresponding genomic sequences, the exon/intron structure of the *BnaGLK1* genes was analyzed and displayed using Gene Structure Display Serve 2.0 (http://gsds.cbi.pku.edu.cn) (Guo et al., 2007). The potential transmembrane domains of GLK1 were predicted using TMHMM (http://www.cbs.dtu.dk/services/TMHMM-2.0/) (Krogh et al., 2001). Functional motifs of GLK1 proteins were identified using multiple expectation maximization (MEME) (http://meme.nbcr.net/meme/cgi-bin/meme.cgi) (Bailey and Elkan, 1994). The parameters used were as follows: distribution of motifs was any number of repetitions, optimum motif width was set from 3 to 300 amino acids, and the maximum number of motifs detected was 100. Each motif was individually checked so that only those with an e-value of <1e-10 were retained for motif detection in *Arabidopsis* and *B. napus* GLK1 proteins (Lu et al., 2015).

Identification and confirmation of *GLK1* gene members in *B. napus*

The *B. napus* accessions L256, L257, ZY821, and Sophia, which have different chloroplast contents (Figure 1), were grown under normal field conditions in Chongqing, China. Normal agronomic procedures for field management were followed. Leaf chlorophyll

content was measured rapidly in triplicate using the SPAD-502Plus chlorophyll meter. Chlorophyll content was compared among varieties with a one-way ANOVA. Total RNA was isolated from 100 mg of young leaves using the RNAprep Pure Plant Kit according to the manufacturer protocol (Tiangen, Beijing China). Contaminating genomic DNA (gDNA) was immediately eliminated using DNase I, and first-strand cDNA was synthesized following the manufacturer protocol. To determine the gene sequence of *B. napus GLK1*, the full-length CDS of *BnaGLK1* was re-amplified from *B. napus* using the degenerate primer pair *BnaGLK1*_F: 5'-ATGTTAGCTCTCTCTCCGGC/AAAGGAAC-3' and *BnaGLK1*_R: 5'-TCAGGCACAAGA/GCGCGGT/CT/CGGAGG-3', which was designed based on the alignment results. Then, the CDS of *BnaGLK1* was ampified in a 50-μL volume including 1 μL template (~100 ng), 1 pmol each primer, 0.5 mM dNTP mix, 5X TransStart *Taq* reaction buffer (with 20 mM Mg^{2+}), and 2.5 U *Pfu-Taq* DNA polymerase (*TransGen*, Beijing, China). Amplifications were performed on a PTC-200 thermo cycler with the following cycling parameters: 94°C for 4 min, and then 35 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 1 min, followed by elongation at 72°C for 10 min. The PCR products were separated on 1% agarose gel electrophorsis and then purified using gel extraction kit (TIANGEN, Beijing, China). The amplified fragments were cloned into the pGEM-T vector according to the manufacturer protocol (Promega). Sequencing confirmed that *BnaGLK1* had been cloned. Three positive colonies were detected. All primer sequences are listed in Table 1.

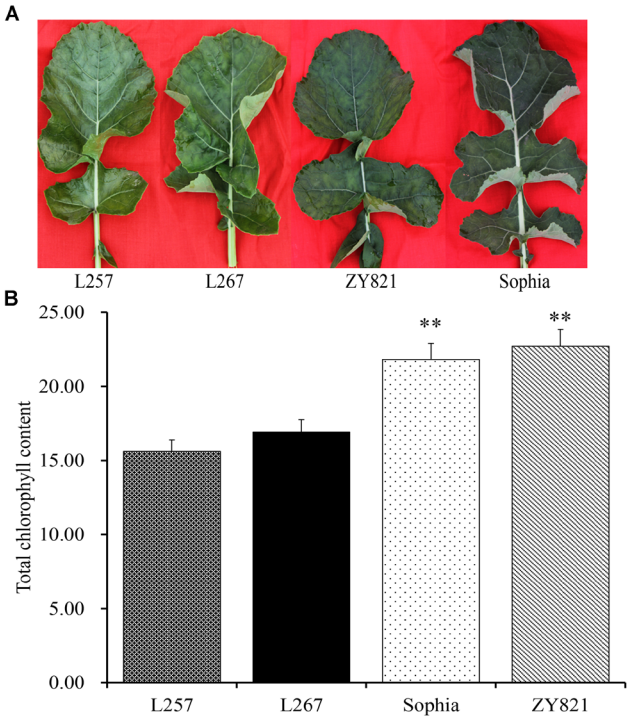

Figure 1. Phenotypes of *Brassica napus* accessions used in this study. **A.** Middle leaves of *B. napus* lines during flourishing florescence. **B.** Means and standard error of total chlorophyll contents in leaves of four *B. napus* accessions. *F* ratios and probabilities from one-way ANOVAs are given (d.f. = 3). Among varieties, columns with the capital letters are significantly different (Tukey's tests, $P < 0.01$).

Table 1. Primers used in this study.

Primers	Sequences (5'-3')	Length (bp)	Tm (°C)
BnaGLK1aF	TCTCCGGCAAGGAACTCCACAAGA	204	64
BnaGLK1aR	GATCTCAGGGTCCATCTCCAAGTC		
BnaGLK1bF	TGGGTCTCGGATACTCCCTACTGG	91	63
BnaGLK1bR	CAGGCGGTGTCGTAAACCTCGTAG		
BnaGLK1cF	ATGGATCGCACCGGCACCCACTATA	209	65
BnaGLK1cR	GAGTATCGGAGACCCAAAAAGGTGG		
BnaGLK1dF	CTAATAAAAAGGGTATTTCGGGAGA	182	60
BnaGLK1dR	CTTCCCTTCGTTGCTATTGTTCTTA		
OvBnaGLK1F	CGGGATCCCGATGTTAGCTCTCTCTCCGGCAAGGAAC	1251	65
OvBnaGLK1R	CCGAGCTCGTCAGGCACAAGGCGCGGTTGGAGG		
BarstaF	CGACATCCGCCGTGCCACCGA	528	60
BarstaR	TAGATCTCGGTGACGGGCAGG		

Expression profile analysis of *BnaGLK1* genes

To decipher the function of *BnaGLK1* genes, their spatial expression patterns were determined in seeds, flowers, stems, buds, leaves, and silique pericarps by qRT-PCR analysis of four *B. napus* lines with different backgrounds. Specific primers were designed based on the alignment results (Table 1). qRT-PCR was performed with a typical 20-μL PCR mixture including 10 μL SYBR® Premix Ex *Taq*™ II, 100 ng template cDNA, 0.8 μM each PCR primer, and ddH_2O to a final volume of 20 μL. The amplification protocol was performed on a Bio-rad CFX96 Real Time System (USA) with optimal cycling conditions (95°C for 2 min, followed by 40 cycles of 95°C denaturation for 10 s, and annealing and extension at 60°C for 20 s). Melting curves were obtained under the following conditions: 95°C for 10 s followed by a constant increase in temperature from 65 to 95°C at an increment of 0.5°C/cycle. The relative expression of *BnGLK1* was calculated by the $2^{-\Delta\Delta Ct}$ method using the expression of *BnACTIN7* (EV116054) and *BnUBC21* (EV086936) as the internal controls (Fricker et al., 2007). Three biological replicates for each sample were used for real-time PCR analysis and three technical replicates were analyzed for each biological replicate. All samples were amplified in triplicate from the same total RNA preparation and the mean value was used for further analysis.

Generation of transgenic plants and phenotypic analysis

Full-length *BnaGLK1a* cDNA was amplified with the primers *OvBnaGLK1aF* and *OvBnaGLK1aR*, using cDNA from *B. napus* (ZY821) leaves as a template. The fragments were cloned into the pGEM-T vector (Promega) and sequenced. Subsequently, the confirmed fragments were sub-cloned into the pCAMBIA-2301 vector with a CaMV 35S promoter, and then introduced into *Agrobacterium tumefaciens* LBA4404, which was introduced into *B. napus* (ZY821) via floral dip transformation (Tan et al., 2011). Positive transgenic lines were selected by PCR analysis with two insertion-specific primer pairs *OvBnaGLK1aF* + *OvBnaGLK1aR* and *BarstaF* + *BarstaR* (Table 1). The phenotypic characteristics were analyzed under a stereoscope, at 1X magnification (Olympus, MVX10, Japan), and the chlorophyll content in leaves was also measured rapidly in triplicate using the SPAD-502Plus chlorophyll meter.

RESULTS

Identification and isolation of *GLK1* genes in *B. napus*

Several QTLs associated with leaf photosynthesis rate have been detected in *B. napus* (Yan et al., 2015). DNA sequence information of a QTL region flanked by SNP markers has been mapped to the *B. napus* database (Chalhoub et al., 2014), and the 200-kb sequences flanking the QTL region were extracted from the abovementioned databases. Among these QTL regions, we found the *GLK1* (BnaA07g00410D) was located on chromosome A07, 59.2 kb away from the SNP marker SNP16353A07 (Figure 2). Furthermore, two additional homologs of *GLK1* were identified and found to be located on chromosomes C07 (BnaC07g00300D) and C08 (BnaC08g36330D), respectively (Figure 2). Although the copy of *GLK1* (BnaC07g00300D) was more than 25 Mb away from the marker SNP28360C07 on chromosome C07, no markers were associated with *GLK1* (BnaC08g36330D) on chromosome C08, possibly due to the low density of markers on that chromosome. We, therefore, inferred that *GLK1* might influence the leaf photosynthesis rate in *B. napus*, and have vital roles during biomass accumulation in *B. napus*.

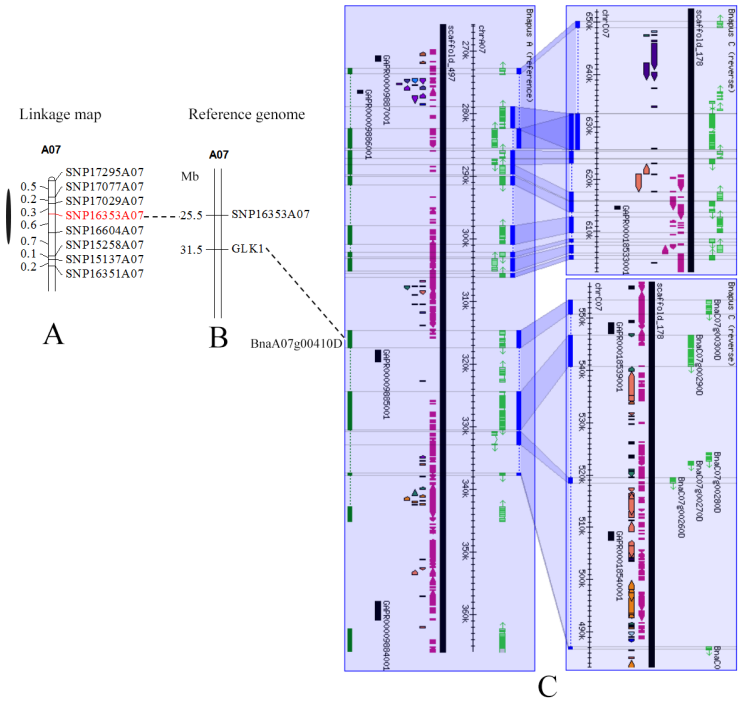

Figure 2. Synteny analysis of candidate quantitative trait loci (QTL) regions of A07 in *Brassica napus*. **A.** Linkage map constructed by single nucleotide polymorphism (SNP) markers. The black oval indicates the QTL for net photosynthesis rate. The genetic distance (cM) and SNP markers are listed on the left and right of the bar, respectively. The SNP marker tightly linked to the QTL is denoted in red. **B.** Chromosome A07 originated from the *B. napus* reference genome (Chalhoub et al., 2014). The physical distance (Mb) and SNP marker and candidate gene are listed on the left and right of the bar, respectively. **C.** The collinearity comparison of QTL flanking sequences among Brassicaceae species.

Identification and phylogenetic analysis of *GLK1* genes

To identify putative GLK1 proteins in different plant species, we conducted a HMM search of various genome databases (Krogh et al., 2001), followed by a comparative genetic analysis of the GLK1 family using AthGLK1as a reference sequence. We identified 96 proteins from 53 plant species, which we classified as members of the Myb_DNA binding family (Pfam: 00249) using InterPro and Pfam analyses (**Table S1**).

Using *PpaGLK1a* and *PpaGLK1b* from *P. patens* as an outgroup, we classified orthologs of *GLK1* genes into different sub-groups, such as Brassicaceae (light pink), and Poales subgroup (dark blue), and Solanaceae subgroups (yellow) (Figure 3). In addition, we found that 24 GLK1 members from 17 species of Brassicaceae were clustered together in a single group (Figure 3, light pink), in accordance with the general consensus that they share a common ancestor (Yang et al., 1999; Chalhoub et al., 2014). However, 24 *GLK1* members from 17 Brassicaceae species were divided into three subgroups, which were named the Brassica family, *A. arabicum*, and *Arabidopsis*, with *BnaGLK1* being more closely related to *BraGLK1* and *BolGLK1* in the Brassicaceae subgroup (Figure 3). Additionally, we aligned and compared the CDSs and their corresponding genomic sequences (**Figures S1** and **S2**) to decipher the evolutionary relationships amongst the Brassicaceae species (Figure 4). We identified five introns for each *GLK1* gene sequence that were conserved in terms of exon/intron structure, excluding *CruGLK1*, *LalGLK1*, and *BolGLK1b*, which contained four, four, and seven introns, respectively. Moreover, three categories of intron phase (0, 1, and 2) were previously described (Pan et al., 2015), and the intron phase pattern (0, 0) was found to be strikingly conserved in all the *GLK1* sequences (Figure 4).

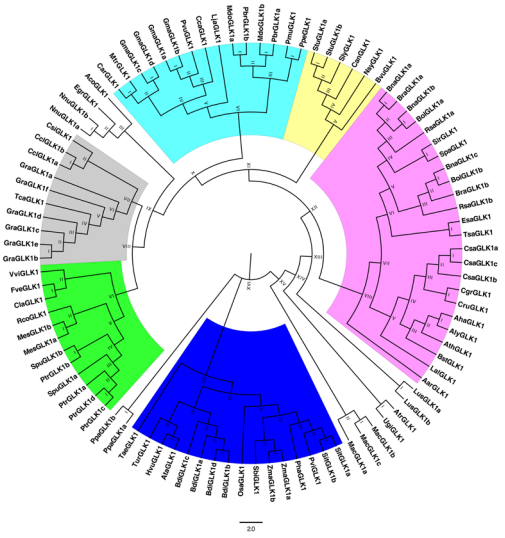

Figure 3. Phylogenetic relationships of plant GLK1 proteins. The phylogenetic tree was generated using the Neighbor-Joining (NJ) method with bootstrap analysis (1000 replicates). The tree was displayed using FigTree v1.4.2. Bootstrap values >50% are denoted at the nodes. Roman numerals represent the branch length. The same color indicates genes classified into the same phylogenetic sub-group (Brassicaceae - light pink; Poales - dark blue; Solanaceae - yellow; Fabidae - light blue; Malvidae - light green; Citrus - gray). The scale bar indicates the average number of amino acid substitutions per site.

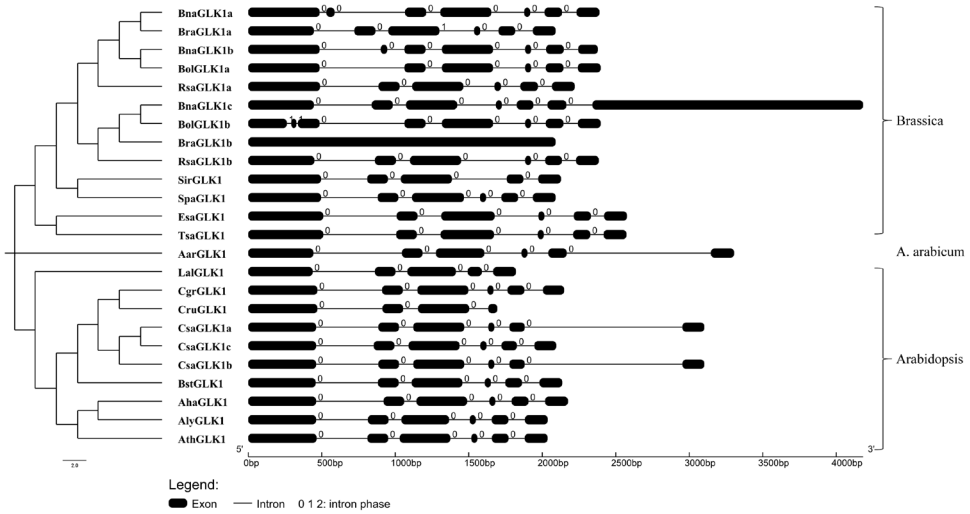

Figure 4. Phylogenetic analysis and gene structure of *GLK1* genes in Brassicaceae species. We aligned 24 full-length amino acid sequences using ClustalW2. The phylogenetic tree (left panel) was constructed using Geneious Pro 4.8.5 and the NJ method (1000 bootstrap replicates), and displayed using FigTree v1.4.2. Gene structure is shown in the right panel. Exons and introns are denoted by black boxes and horizontal lines, respectively. Introns in phases 0, 1, and 2 are represented by the numbers 0, 1, and 2, respectively. The scale bar represents 2.0 kb.

Conservation of the GLK1 sequences in Brassicaceae species

In the present study, no putative transmembrane domains were predicted in GLK1 proteins using the TMHMM program (Krogh et al., 2001), in accordance with the known properties of GLK1 proteins (Hall et al., 1998; Rossini et al., 2001). Based on the alignment results, the identity of the HLH domain was found to be 96.5%, showing high conservation among Brassicaceae species (Figure 5 and **Figure S3**). Furthermore, the putative helix-loop-helix (HLH) DNA-binding domain in *GLK* gene families always consists of two helices that are separated by a 22-amino acid loop (Hall et al., 1998; Rossini et al., 2001), and Brassicaceae species contain a glutamic acid (E) residue within this 22-amino acid loop that is replaced with aspartic acid (D) in *Arabidopsis* species (Figure 5). An aspartic acid (D) residue was found in *OsGLK1* and *ZmGLK1* (Rossini et al., 2001), suggesting that the sequences are highly conserved in related species. In addition, all copies of GLK1 proteins were highly conserved based on amino acid sequence alignments, especially in closely related species (Figure 5). In addition, some differences in amino acid residues were detected among Brassicaceae species, which may be directly associated with the specific functions of this gene in different species (Figure 5). In addition, we analyzed motifs in GLK1s from Brassicaceae species using the MEME suite 4.11.1 (http://meme-suite.org/tools/meme). We identified 19 distinct motifs in GLK1 protein sequences (**Figure S3**), four of which were widely distributed in all GLK1 sequences (**Table S2** and Figure 6). Furthermore, the motif TIGR01557 for myb_SHAQKYF (Motif 1) and PLN03162 for the Golden-2 like transcription factor (Motif 3) are putative motifs in the GLK1 family (Fitter et al., 2002). Identification of amino acid differences within these motifs in Brassicaceae species will provide insight into the specific functions of GLK1 in different species.

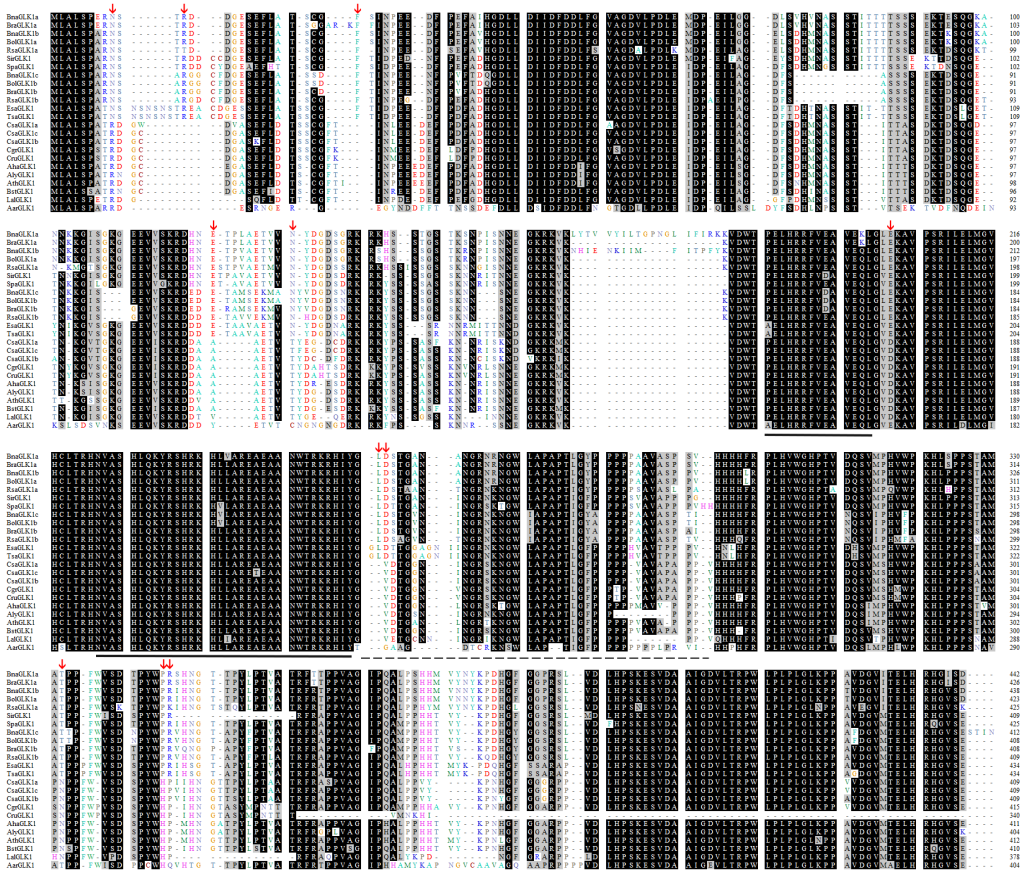

Figure 5. Multiple alignment analysis of 24 GLK1 proteins in Brassicaceae species. Full-length GLK1 proteins were aligned using the default settings of Muscle in Bioedit. Identical sequences (numbered on the right) are highlighted in black and similar residues are shaded in gray (shading >80% threshold). Differences in the amino acid sequence are indicated in color. The putative HLH DNA-binding domain is indicated by solid horizontal bars, indicating the predicted α-helix segments conserved in all Brassicaceae GLK1 proteins, and dashed horizontal bars indicate the maximum extension of the second α-helix. Red arrows indicate the mutated amino acid loci as observed during phylogenetic analysis of GLK1 among Brassicaceae species.

Triplication analysis of *GLK1* genes

Brassicaceae species have a propensity to undergo genome duplications and mergers during evolution (Chalhoub et al., 2014). The allotetraploid species *B. napus*, which was derived from the interspecific hybridization of two diploid species, *B. rapa* and *B. oleracea*, is an ideal model for the study of gene evolution and function. In addition, excessive gene loss is typical following polyploid formation in eukaryotes (Qu et al., 2013). In this study, all copies of *GLK1* identified were verified using the homology cloning method, and each of the orthologous blocks corresponding to ancestral blocks was identified using collinearity between Brassicaceae orthologs (**Table S3**).

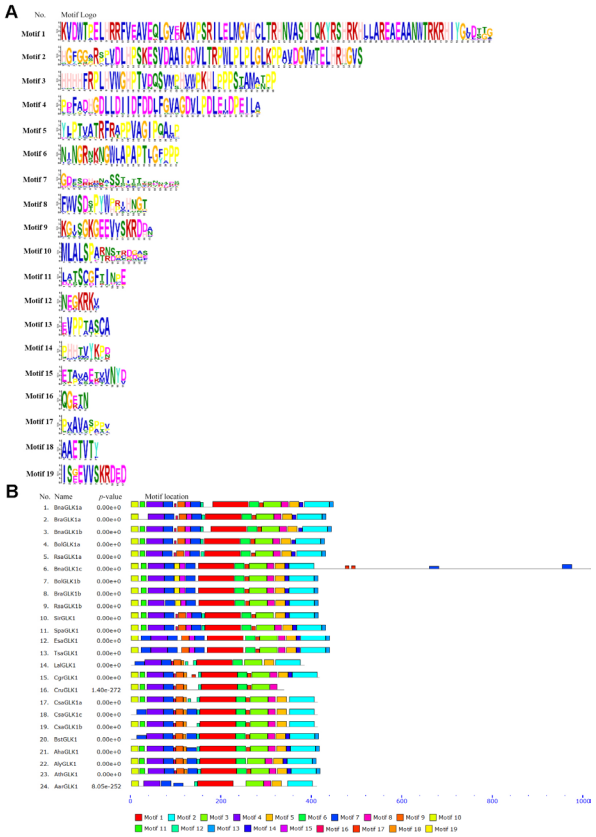

Figure 6. Combined motif diagram and sequence logos of GLK1 in Brassicaceae species by Multiple Expectation Maximization for Motif Elicitation (MEME) analysis. **A.** Amino acid residues are indicated by one-letter abbreviations. The height of letters is proportional to the relative frequency of the corresponding amino acid in the aligned column. **B.** The blocks show the order and spacing of non-overlapping motif matches on the protein sequences. Only the most significant motif is shown when two motifs match. All of the motifs derived from the MEME program are represented by different colored boxes. Scale bar indicates the number of amino acid residues.

Three copies were identified in the least fractionated blocks (LF), the medium fractionated blocks (MF1), and the most fractionated blocks (MF2) in the *B. napus* and *O. sativa* genomes. In the *B. rapa* and *B. oleracea* genomes, two copies of *GLK1* were detected and located on the LF and MF2 blocks, respectively; however, only one copy was identified in the remaining species (**Table S3**). These results provide an opportunity to study *GLK1* gene retention in the triplicated genomes of Brassicaceae species.

Cloning and confirmation of members of the *BnaGLK1* gene family

Using the homology cloning method, 16 *GLK1* sequences, which varied in length from 1242 to 1251 bp, were isolated from *B. napus* (**Figure S4**). We named these sequences *BnaGLK1_1* to *BnaGLK1_16* (**Table S4**), respectively. Phylogenetic analysis showed that they were divided into four homologous genes cloned from *B. napus* (Figure 7A), and that the

amplified sequences were shorter than three putative sequences from the reference genome (**Table S4**). Those differences between the four copies of the *GLK1* gene were probably due to splicing errors of the reference genome sequence or to different gene copies being present in different accessions (Figure 7B), named *BnGLK1a* to *BnGLK1d* (**Table S4**).

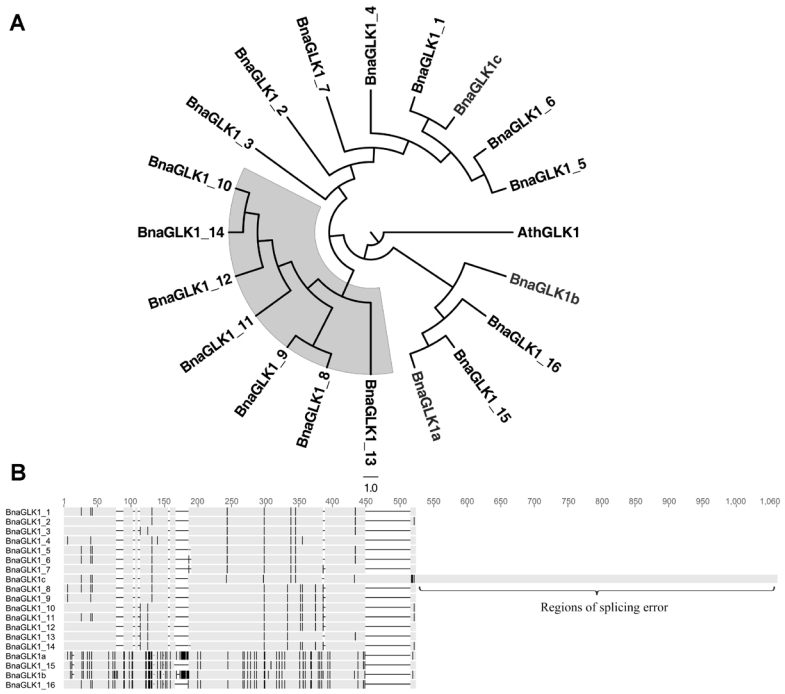

Figure 7. Classification of *BnaGLK1* family members in *Brassica napus*. **A.** Phylogenetic relationships of BnaGLK1 proteins from different *B. napus* accessions. Gray indicates the new copy of *BnGLK1* named *BnaGLK1d*. **B.** Alignment analysis of BnaGLK1 proteins from different *B. napus* accessions. Gray indicates the conserved sequence regions, and black denotes amino acid differences.

Expression profiles of *BnaGLK1* genes

GLK1 is essential for normal chloroplast development in many plant species (Rossini et al., 2001; Waters et al., 2008, 2009; Nakamura et al., 2009; Nguyen et al., 2014). Results of one-way ANOVAs showed there were significant differences in chlorophyll content among varieties (Tukey's tests, $P < 0.01$, Figure 1). Therefore, we collected seeds, flowers, stems, buds, leaves, and silique pericarps from the ZY821 and Sophia lines of *B. napus,* which have a relatively high chloroplast content, and L257 and L267, which have a low chloroplast content, and analyzed *BnaGLK1* expression using qRT-PCR analysis. We found that the expression levels of *BnaGLK1* varied among different tissues and organs in each sample. *BnaGLK1* gene expression was minimal in the seeds, flowers, and buds of all tested samples, and was higher in the leaves (Figure 8). The expression levels of *BnaGLK1* in leaves were 311-fold (P value < 0.01) higher than in the other tissues. Moreover, all *BnaGLK1* genes were expressed at higher levels in ZY821 and Sophia, which had a higher chloroplast content than L257 and

L267, and a 2.87-fold (P value < 0.01) higher content than in ZY821 and Sophia accessions compared to L257 and L267 (Figures 1 and 8). Those results show that *BnaGLK1* genes are specifically expressed in the leaves of *B. napus*, and thus might be associated with chloroplast development in leaves, in accordance with previously published findings (Yan et al., 2015). In addition, stems and silique pericarps contribute to biomass accumulation in the later stages of oilseed development. However, the expression levels of *BnaGLK1* genes were lower in the stems and silique pericarps than in the leaves (Figure 8).

To identify the function of *BnaGLK1a*, we overexpressed 35S-*BnaGLK1a* in *B. napus* (ZY821). Here, the T_1 generation of transgenic rapeseed was generated by overexpressing *BnaGLK1a*, which showed darker leaves than the control (Figure 9). The chloroplast content was 2.5-fold (P value < 0.01) higher in transgenic lines compared to the control (data not shown), suggesting that *BnaGLK1a* might promote chloroplast development and thus affect the photosynthesis rate in *B. napus*. These results provide a foundation for the further study of *BnaGLK1*.

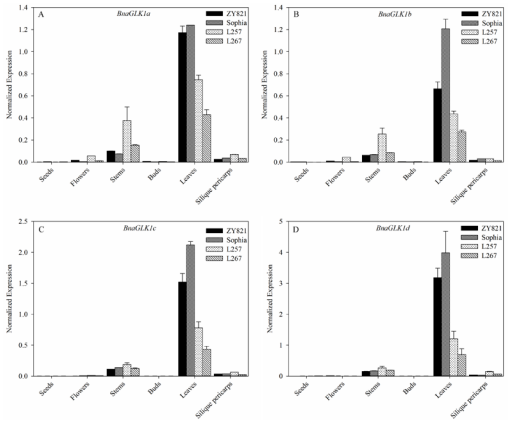

Figure 8. Expression profiles of *BnaGLK1* genes in *Brassica napus* with different backgrounds. **A.** *BnaGLK1a*; **B.** *BnaGLK1b*; **C.** *BnaGLK1c*; **D.** *BnaGLK1d*. The expression levels were normalized using the expression of *BnActin7* and *BnUBC21* genes as a reference. Data indicate a mean value of three repeats from three independent qRT-PCR assays. Error bars represent the SE for three independent experiments. The primers are listed in Table 1.

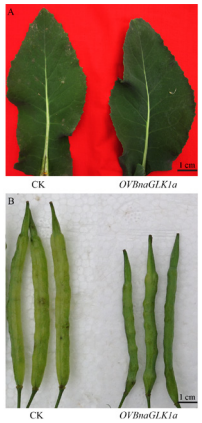

Figure 9. Phenotypic characteristics of the control (ZY821) and T_1 generations. **A.** Phenotypic characteristics of leaves. **B.** Phenotypic characteristics of Siliques. Scale bar is 1.0 cm.

DISCUSSION

In plants, chloroplasts play a crucial role in photosynthesis, and represent one of the most important organelles in green plant cells. Previous studies have shown that *GLK1* is essential for normal chloroplast development in many plant species, and directly influences the leaf photosynthesis rate in *Arabidopsis*, maize, and rice (Rossini et al., 2001; Waters et al., 2008; Nakamura et al., 2009; Nguyen et al., 2014). Moreover, several QTLs associated with the leaf photosynthesis rate have been reported in *B. napus* (Yan et al., 2015). Based on that research, we identified *GLK1* (BnaA07g00410D) on chromosome A07, 59.2 kb away from the SNP marker SNP16353A07 (Figure 2). We then sought to investigate *GLK1* gene evolution and diversity in *B. napus* by surveying 77 different plant species. The results showed that 96 sequences from 53 plant species were clearly classified into subgroups through phylogenetic analysis using *PpaGLK1* as an outgroup (Figure 3), revealing the evolutionary relationship amongst these proteins and the relative time since divergence from a common ancestor. As Poales and Brassicaceae formed two different subgroups, we propose that *GLK1* is associated with different biological functions in these plants (Nakamura et al., 2009; Waters et al., 2009; Wang et al., 2013; Nguyen et al., 2014). In addition, 24 members of the *GLK1* family from 17 Brassicaceae species were divided into three subgroups, namely the Brassica family, *A. arabicum*, and *Arabidopsis*, with *BnaGLK1* being more closely related to *BraGLK1* and *BolGLK1* in the Brassicaceae subgroup (Figure 3). This is in accordance with the allotetraploid *B. napus* being formed by the fusion of the diploids *B. rapa* (AA) and *B. oleracea* (CC) through large-scale homologous recombination between the A and C genomes (Chalhoub et al., 2014). Based on the *B. napus* genome sequence, we identified three *GLK1* homologs located in the rapeseed A or C sub-genome, and four copies of *GLK1* were isolated by manual cloning and sequencing (Figure 7). Therefore, the *B. rapa* and *B. oleracea* genomes contain the same number of *GLK1* genes as the *B. napus* genome. We presumed that *GLK1* had not undergone gene loss during allopolyploid formation, which is consistent with previous reports that *GLK* genes are present in more than one copy in plants (Rossini et al., 2001). However, the full-length CDS of the *GLK1* genes in *B. napus* was shorter than the putative sequence, probably due to errors in splicing of the reference genome sequence or different gene copies being present in different *B. napus* accessions. In addition, most Brassicaceae *GLK1* genes share three major structural features at both the nucleic acid and amino acid levels. First, all *GLK1* genes had remarkably conserved exon/intron structures (six exons/five intron) and intron phase patterns (0, 0, 0, 0, and 0), indicating that the *GLK1* gene structure was established and retained following divergence of the Brassicaceae (Figure 4). Second, the amino acid sequence of the HLH domain was highly conserved, and the domain was composed of two helices, as reported for *OsGLK1* and *ZmGLK1* (Hall et al., 1998; Rossini et al., 2001). Alignment analysis revealed a higher level of conservation amongst *GLK1* genes in Brassicaceae than in rice and maize (Figure 5) (Rossini et al., 2001). Thus, it seems that *GLK1* did not undergo functional divergence in Brassicaceae species during evolution. Importantly, amino acid mutant loci were identified in a subgroup of Brassicaceae *GLK1* genes (Figure 5), which could be used for allelic-specific PCR or for developing specific gene chips to distinguish the species. Third, 21 motifs with e-values of < 1e-10 were detected among the 24 Brassicaceae GLK1 proteins (**Table S2**). Motifs 1 to 5 were present in all input Brassicaceae GLK1 proteins, and motif 1 (80 aa) possessed the conserved Myb_DNA binding family domain (**Table S2** and Figure 6), which could be used as a candidate target region for analyzing the functional and structural

divergence among the GLK1 proteins of different clades. Therefore, we could define the features of *GLK1* gene family members based on the presence of HLH and GCT boxes.

The present study showed that *GLK1* in maize, rice, *Arabidopsis*, and *P. patens* is related to chloroplast development and regulation of plastid-encoded genes (Hall et al., 1998; Fitter et al., 2002; Waters et al., 2008, 2009; Yasumura et al., 2005). In *Arabidopsis*, Chlorophyll accumulation was reduced in photosynthetic tissues following the loss of *GLK* activity, and was increased in non-foliar tissues by overexpression of GLKs (Waters et al., 2009). In tomato, overexpression of either *GLK1* or *GLK2* resulted in the development of dark green tomato fruit with high chlorophyll and chloroplast levels, and co-suppression of *GLK1* led to the development of pale leaves with reduced chlorophyll content (Nguyen et al., 2014). Moreover, our results also showed that *BnaGLK1* genes are expressed at higher levels in lines with a high chloroplast content (i.e., ZY821 and Sophia) than in those with a low chloroplast content (L257 and L267) (Figure 8), suggesting that *BnaGLK1* is involved in normal chloroplast development. Importantly, rapeseed plants overexpressing *BnaGLK1a* tend to have darker leaves and siliques than control plants (Figure 9), with higher chloroplast contents (data not shown). However, the main source of biomass accumulation during the later stages of oilseed development is photosynthesis in the stems and silique pericarps. Therefore, oilseed production could be enhanced by preferentially increasing *BnaGLK1a* expression in stems and silique pericarps. In addition, *GLK* genes are involved in the key steps of chlorophyll biosynthesis, which promote the transcription of genes and increase flux through the pathway (Waters et al., 2009). Moreover, *GLK1* has been reported to be involved in retrograde signaling essential for coordinating plastid protein import and nuclear gene expression, particularly those that are involved in chloroplast development and are induced by environmental and hormonal signals (Nakamura et al., 2009; Kakizaki and Inaba, 2010). In the present study, we confirm that *BnGLK1* could be involved in chloroplast development in *B. napus*, which provides a foundation for understanding the molecular mechanism of *GLK1* function in the photosynthesis of *B. napus*.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (#31471885, #31101546) and the Chongqing Natural Science Foundation (#cstc2014jcyjA80026).

REFERENCES

Ahmad S, Sadaqat HA, Tahir MHN and Awan FS (2015). An insight in the genetic control and interrelationship of some quality traits in *Brassica napus*. *Genet. Mol. Res.* 14: 17941-17950. http://dx.doi.org/10.4238/2015.December.22.19

Bailey TL and Elkan C (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press 2: 28-36.

Chalhoub B, Denoeud F, Liu S, Parkin IA, et al. (2014). Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science* 345: 950-953. http://dx.doi.org/10.1126/science.1253435

Cheng Y, Yao ZP, Ruan MY, Ye QJ, et al. (2016). *In silico* identification and characterization of the WRKY gene superfamily in pepper (*Capsicum annuum* L.). *Genet. Mol. Res.* 15: gmr.15038675.

Dyall SD, Brown MT and Johnson PJ (2004). Ancient invasions: from endosymbionts to organelles. *Science* 304: 253-257.
Fairchild CD, Schumaker MA and Quail PH (2000). HFR1 encodes an atypical bHLH protein that acts in phytochrome A signal transduction. *Genes Dev.* 14: 2377-2391.
Fitter DW, Martin DJ, Copley MJ, Scotland RW, et al. (2002). GLK gene pairs regulate chloroplast development in diverse plant species. *Plant J.* 31: 713-727. http://dx.doi.org/10.1046/j.1365-313X.2002.01390.x
Fricker M, Messelhäusser U, Busch U, Scherer S, et al. (2007). Diagnostic real-time PCR assays for the detection of emetic Bacillus cereus strains in foods and recent food-borne outbreaks. *Appl. Environ. Microbiol.* 73: 1892-1898. http://dx.doi.org/10.1128/AEM.02219-06
Guo AY, Zhu QH, Chen X and Luo JC (2007). GSDS: a gene structure display server. *Yi Chuan* 29: 1023-1026. http://dx.doi.org/10.1360/yc-007-1023
Hall LN, Rossini L, Cribb L and Langdale JA (1998). GOLDEN 2: a novel transcriptional regulator of cellular differentiation in the maize leaf. *Plant Cell* 10: 925-936. http://dx.doi.org/10.1105/tpc.10.6.925
Hu S, Yu C, Zhao H, Sun G, et al. (2007). Genetic diversity of *Brassica napus* L. Germplasm from China and Europe assessed by some agronomically important characters. *Euphytica* 154: 9-16. http://dx.doi.org/10.1007/s10681-006-9263-8
Hua W, Li RJ, Zhan GM, Liu J, et al. (2012). Maternal control of seed oil content in *Brassica napus*: the role of silique wall photosynthesis. *Plant J.* 69: 432-444. http://dx.doi.org/10.1111/j.1365-313X.2011.04802.x
Kakizaki T, Matsumura H, Nakayama K, Che FS, et al. (2009). Coordination of plastid protein import and nuclear gene expression by plastid-to-nucleus retrograde signaling. *Plant Physiol.* 151: 1339-1353. http://dx.doi.org/10.1104/pp.109.145987
Kakizaki T and Inaba T (2010). New insights into the retrograde signaling pathway between the plastids and the nucleus. *Plant Signal. Behav.* 5: 196-199. http://dx.doi.org/10.4161/psb.5.2.11107
Krogh A, Larsson B, von Heijne G and Sonnhammer EL (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305: 567-580. http://dx.doi.org/10.1006/jmbi.2000.4315
Kobayashi K, Sasaki D, Noguchi K, Fujinuma D, et al. (2013). Photosynthesis of root chloroplasts developed in *Arabidopsis* lines overexpressing *GOLDEN2-LIKE* transcription factors. *Plant Cell Physiol.* 54: 1365-1377. http://dx.doi.org/10.1093/pcp/pct086
Liu TJ, Zhang CY, Yan HF, Zhang L, et al. (2016). Complete plastid genome sequence of *Primula sinensis* (Primulaceae): structure comparison, sequence variation and evidence for *accD* transfer to nucleus. *PeerJ* 4: e2101. http://dx.doi.org/10.7717/peerj.2101
Lu K, Guo W, Lu J, Yu H, et al. (2015). Genome-wide survey and expression profile analysis of the mitogen-activated protein kinase (MAPK) gene family in *Brassica rapa*. *PLoS One* 10: e0132051. http://dx.doi.org/10.1371/journal.pone.0132051
Martin W, Rujan T, Richly E, Hansen A, et al. (2002). Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Natl. Acad. Sci. USA* 99: 12246-12251. http://dx.doi.org/10.1073/pnas.182432999
McCormac AC and Terry MJ (2002). Light-signalling pathways leading to the co-ordinated expression of HEMA1 and Lhcb during chloroplast development in *Arabidopsis thaliana*. *Plant J.* 32: 549-559. http://dx.doi.org/10.1046/j.1365-313X.2002.01443.x
Nakamura H, Muramatsu M, Hakata M, Ueno O, et al. (2009). Ectopic overexpression of the transcription factor *OsGLK1* induces chloroplast development in non-green rice cells. *Plant Cell Physiol.* 50: 1933-1949. http://dx.doi.org/10.1093/pcp/pcp138
Nguyen CV, Vrebalov JT, Gapper NE, Zheng Y, et al. (2014). Tomato GOLDEN2-LIKE transcription factors reveal molecular gradients that function during fruit development and ripening. *Plant Cell* 26: 585-601. http://dx.doi.org/10.1105/tpc.113.118794
Pan X, Peng FY and Weselake RJ (2015). Genome-wide analysis of *PHOSPHOLIPID:DIACYLGLYCEROL ACYLTRANSFERASE* (*PDAT*) genes in plants reveals the eudicot-wide PDAT gene expansion and altered selective pressures acting on the core eudicot *PDAT* paralogs. *Plant Physiol.* 167: 887-904. http://dx.doi.org/10.1104/pp.114.253658
Qu C, Fu F, Lu K, Zhang K, et al. (2013). Differential accumulation of phenolic compounds and expression of related genes in black- and yellow-seeded *Brassica napus*. *J. Exp. Bot.* 64: 2885-2898. http://dx.doi.org/10.1093/jxb/ert148
Riechmann JL, Heard J, Martin G, Reuber L, et al. (2000). *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290: 2105-2110. http://dx.doi.org/10.1126/science.290.5499.2105
Rossini L, Cribb L, Martin DJ and Langdale JA (2001). The maize golden2 gene defines a novel class of transcriptional regulators in plants. *Plant Cell* 13: 1231-1244. http://dx.doi.org/10.1105/tpc.13.5.1231

Spiegelman JI, Mindrinos MN, Fankhauser C, Richards D, et al. (2000). Cloning of the *Arabidopsis* RSF1 gene by using a mapping strategy based on high-density DNA arrays and denaturing high-performance liquid chromatography. *Plant Cell* 12: 2485-2498. http://dx.doi.org/10.1105/tpc.12.12.2485

Tan H, Yang X, Zhang F, Zheng X, et al. (2011). Enhanced seed oil production in canola by conditional expression of *Brassica napus LEAFY COTYLEDON1 and LEC1-LIKE* in developing seeds. *Plant Physiol.* 156: 1577-1588. http://dx.doi.org/10.1104/pp.111.175000

Wang H, Wu G, Zhao B, Wang B, et al. (2016). Regulatory modules controlling early shade avoidance response in maize seedlings. *BMC Genomics* 17: 269. http://dx.doi.org/10.1186/s12864-016-2593-6

Wang P, Fouracre J, Kelly S, Karki S, et al. (2013). Evolution of GOLDEN2-LIKE gene function in C(3) and C (4) plants. *Planta* 237: 481-495. http://dx.doi.org/10.1007/s00425-012-1754-3

Waters MT, Moylan EC and Langdale JA (2008). GLK transcription factors regulate chloroplast development in a cell-autonomous manner. *Plant J.* 56: 432-444. http://dx.doi.org/10.1111/j.1365-313X.2008.03616.x

Waters MT, Wang P, Korkaric M, Capper RG, et al. (2009). GLK transcription factors coordinate expression of the photosynthetic apparatus in *Arabidopsis*. *Plant Cell* 21: 1109-1128. http://dx.doi.org/10.1105/tpc.108.065250

Xu X, Paik I, Zhu L and Huq E (2015). Illuminating progress in phytochrome-mediated light signaling pathways. *Trends Plant Sci.* 20: 641-650. http://dx.doi.org/10.1016/j.tplants.2015.06.010

Yan XY, Qu CM, Li JN, Li C, et al. (2015). QTL analysis of leaf photosynthesis rate and related physiological traits in *Brassica napus*. *J. Integr. Agric.* 14: 1261-1268. http://dx.doi.org/10.1016/S2095-3119(14)60958-8

Yang YW, Lai KN, Tai PY and Li WH (1999). Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages. *J. Mol. Evol.* 48: 597-604. http://dx.doi.org/10.1007/PL00006502

Yamori W and Shikanai T (2016). Physiological functions of cyclic electron transport around photosystem I in sustaining photosynthesis and plant growth. *Annu. Rev. Plant Biol.* 67: 81-106. http://dx.doi.org/10.1146/annurev-arplant-043015-112002

Yasumura Y, Moylan EC and Langdale JA (2005). A conserved transcription factor mediates nuclear control of organelle biogenesis in anciently diverged land plants. *Plant Cell* 17: 1894-1907. http://dx.doi.org/10.1105/tpc.105.033191

Supplementary material

Table S1. Gene taxa IDs of protein sequences used in this study. Taxa IDs were collected from public databases, as described in Materials and Methods.

Table S2. Motifs screened in the GLK1 proteins of 24 Brassicaceae species by Multiple Expectation Maximization for Motif Elicitation MEME analysis.

Table S3. Gene copies of *GLK1* in *Arabidopsis* and Brassicaceae species.

Table S4. Summary of *BnaGLK1* genes from different *Brassica napus* accessions.

Figure S1. Genome sequences of *GLK1* in Brassicaceae species for this study.

Figure S2. Full-length coding sequences (CDSs) of *GLK1* in Brassicaceae species used in this study.

Figure S3. Amino acid sequences of *GLK1* in Brassicaceae species used in this study.

Figure S4. Full-length CDSs of *GLK1* from *Brassica napus* used in this study.