Cao et al. BMC Genomics (2016) 17:201
DOI 10.1186/s12864-016-2519-3

BMC Genomics

RESEARCH ARTICLE Open Access

Evidence for the role of transposons in the recruitment of *cis*-regulatory motifs during the evolution of C4 photosynthesis

Chensi Cao, Jiajia Xu, Guangyong Zheng and Xin-Guang Zhu*

Abstract

Background: C_4 photosynthesis evolved from C_3 photosynthesis and has higher light, water, and nitrogen use efficiencies. Several C_4 photosynthesis genes show cell-specific expression patterns, which are required for these high resource-use efficiencies. However, the mechanisms underlying the evolution of *cis*-regulatory elements that control these cell-specific expression patterns remain elusive.

Results: In the present study, we tested the hypothesis that the *cis*-regulatory motifs related to C_4 photosynthesis genes were recruited from non-photosynthetic genes and further examined potential mechanisms facilitating this recruitment. We examined 65 predicted bundle sheath cell-specific motifs, 17 experimentally validated cell-specific *cis*-regulatory elements, and 1,034 motifs derived from gene regulatory networks. Approximately 7, 5, and 1,000 of these three categories of motifs, respectively, were apparently recruited during the evolution of C_4 photosynthesis. In addition, we checked 1) the distance between the acceptors and the donors of potentially recruited motifs in a chromosome, and 2) whether the potentially recruited motifs reside within the overlapping region of transposable elements and the promoter of donor genes. The results showed that 7, 4, and 658 of the potentially recruited motifs might have moved via the transposable elements. Furthermore, the potentially recruited motifs showed higher binding affinity to transcription factors compared to randomly generated sequences of the same length as the motifs.

Conclusions: This study provides molecular evidence supporting the hypothesis that transposon-driven recruitment of pre-existing *cis*-regulatory elements from non-photosynthetic genes into photosynthetic genes plays an important role during C_4 evolution. The findings of the present study coincide with the observed repetitive emergence of C_4 during evolution.

Keywords: *cis*-regulatory elements, Motif recruitment, Transposons, Binding affinity

Background

C_4 photosynthesis differs from C_3 photosynthesis by possessing a CO_2 concentrating mechanism, which enables C_4 plants to achieve higher light, water, and nitrogen use efficiencies [1, 2]. The higher photosynthetic efficiency in C_4 plants is achieved by elevating the concentration of CO_2 around ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO). Extensive efforts have been made to engineer a C_4 photosynthetic machinery into C_3 plants such as rice and wheat [3, 4]. Elucidation of the molecular mechanism underlying the evolution of the key components of the concentrating process in C_4 plants and identifying its molecular regulators, either as *cis*-regulatory elements or *trans*-factors and controlling C_4 photosynthetic features [5, 6] are necessary to successfully perform C_4 photosynthesis [5, 7]. To date, despite the establishment of the biochemical and anatomical features of C_4 photosynthesis, our understanding of the genetic control of various C_4 properties such as the reduction in interveinal distance, increased number of chloroplasts within bundle sheath (BS) cells, extensive differentiation of M and BS chloroplast proteomes, and higher plasmodesmata abundance for transport between M and BS cells [3] is limited. More efforts to identify regulatory elements required for the establishment of cell-specific expression patterns of C_4 photosynthesis-related genes are warranted.

* Correspondence: zhuxinguang@picb.ac.cn
CAS Key Laboratory for Computational Biology, CAS-MPG Partner Institute for Computational Biology, Chinese Academy of Sciences, Room 102, Physiology Building, 320 Yueyang Road, Shanghai 200031, China

BioMed Central

© 2016 Cao et al. **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Various approaches, including both forward genetics and reverse genetics approaches, have been utilized to address this question [8].

To date, the *cis*-regulatory motifs controlling the cell-specific expression patterns of C_4-related genes have been mainly discovered through experimental approaches such as deletion analysis [4, 5]. Recent technological advances in computational biology have facilitated the identification of motifs [9–11]. Such computational analyses usually start with clustering genes from a transcriptomic data set into different clusters, followed by prediction of motifs in genes from each cluster [9–11]. One underlying assumption of this approach is that genes within the same cluster are potentially regulated by common *cis*-regulatory elements. However, there are circumstances where this assumption is violated. For example, let us consider three genes, *A, B*, and *C*. Genes *A* and *B* are regulated by the same *cis*-regulatory elements, whereas *C* is regulated by *B* and hence shows the same expression pattern as that of *B*. If the expression pattern is used as the sole criterion in clustering these genes, then these three genes will be misclassified into the same gene cluster, which in turn can lead to the inaccurate detection of *cis*-regulatory elements. Gene regulatory networks constructed based on conditional mutual information can solve the issue of misclassifying genes into the same cluster because this algorithm only detects genes with direct regulatory relationships [12].

In the present study, we examined the potential mechanism related to the formation of new *cis*-regulatory elements in the promoter region of C_4-related genes. To do this, we first identified the potentially recruited *cis*-regulatory elements using gene regulatory networks constructed based on conditional mutual information. Furthermore, we used three sets of motifs, i.e., network-derived motifs, experimentally identified *cis*-regulatory elements, and predicted bundle sheath specific motifs, to test whether these exist in the genes directly linked to the C_3 ortholog of C_4 genes in a C_3 gene regulatory network. Lastly, to explore the potential mechanisms responsible for the recruitment of these motifs, we explored whether the potentially recruited motifs reside in the overlapping regions between transposable elements (TEs) and promoter regions of the C_4 genes. We discussed all these results in light of the hypothesis that transposons play a role in the recruitment of these motifs during the emergence of C_4 photosynthesis.

Results

Identification of *cis*-regulatory elements that were potentially recruited during C_4 evolution

Based on the strategy shown in Fig. 1, 40 pairs of orthologs of the C_4 genes in maize and rice, including its promoter sequences, were obtained. We checked the distribution of BS cell-specific motifs in these 40 pairs (see Methods). We identified seven motifs that might have been potentially recruited during C_4 evolution based on the following criteria: a) these are differentially distributed in C_4 and C_3 orthologs; b) these existed in the neighboring genes of the C_3 orthologs (Table 1).

Evidence for potential involvement of transposon in motif recruitment

We next tested whether transposons played a role during this motif movement. Considering that a small chromosomal distance facilitates transposition, we first determined the distance between the donor gene to the acceptor gene relative to the total length of the chromosome, and identified candidate donor genes that were located <1/10 of the total length of the chromosome from the acceptor genes (Fig. 2; Additional file 1: Table S3). Furthermore, when TE transferred a motif to another locus, the particular motif should be within the overlapping region of TEs and the promoters of the donor gene. We hence aligned the sequences of TEs with those of the promoters of donor genes by using BLAST [13] and identified the overlapping regions. All seven motifs that were differentially distributed in C_4 and C_3 orthologs were indeed present in the overlapping regions of TEs and candidate donors of motifs (Table 2). These two pieces of evidence suggest that transposons may have played a role in the recruitment of these BS cell-specific motifs.

Recruited motifs are possible binding sites for transcription factors

Earlier reports have shown that TEs contribute to the formation of new TF binding sites [14, 15] and evolution of new regulatory mechanisms. Here we checked whether the recruited motifs are potential binding sites of transcription factors (TFs). We tested 124 TFs, for which the position weight matrix information is available from TRANSFAC (Additional file 1: Table S4). For the potentially recruited motifs and TFs, we calculated their binding possibilities (see Methods). Seven of these potentially recruited motifs showed higher (more than two-fold) binding affinities with TFs compared to the calculated affinity for random elements of the same length as the motif (Table 3; see Methods). Several of the TFs binding to these motifs were earlier identified to be potential regulators of photosynthesis such as Opaque-2 and GBF (Table 4) [16].

Occurrence of TE-driven motif recruitment in the experimentally validated motifs

To test whether TE-driven motif recruitment is a general phenomenon, we further examined whether TEs are involved in the recruitment of motifs that were

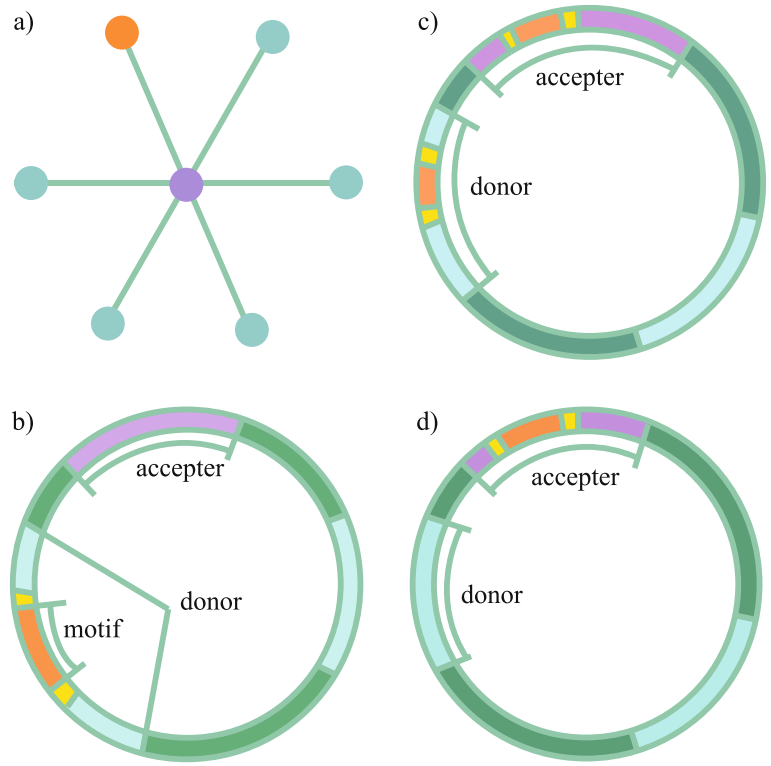

Fig. 1 Distribution of motifs in genes before and after recruitment. **a** Before the recruitment, the donor (*orange dot*) is located in a neighboring gene of the acceptor (*purple dot*) in a gene regulatory network (GRN). **b** Before the recruitment, the donor contains the motif (*orange block*) and the acceptor (*purple block*) lacks the motif. **c** After a copy-and-paste recruitment, the acceptor (*purple block*) contains the motif (*orange block*, whereas the donor also maintains the motif. **d** After a cut-and-paste recruitment, the acceptor (*purple block*) recruits the motif (*orange block*), whereas the donor loses the motif

experimentally identified to be related to their host gene' cell-specific gene expression pattern [5] and also the predicted *cis*-regulatory motifs based on the genes in the same gene community in a gene regulatory network.

Of the 17 experimentally validated motifs, five were identified as potentially recruited motifs (Additional file 1: Table S5). Of these five potentially recruited motifs, four resided in the overlapping region of TEs and their candidate donors (Table 5; Fig. 3). In addition, the donors of these four motifs were proximally located to the acceptors in their residing chromosome (Additional file 1: Table S6). Similar to the analysis of the BS cell-specific motifs, the five putative recruited and validated motifs showed higher binding affinity to TFs (Additional file 1: Table S7).

We also obtained similar results in the analysis of network-derived motifs. There were 1,034 motifs differentially distributed in the C_4 and C_3 orthologs, and 1,000 of these were identified as potential recruited motifs (Additional file 2). A total of 658 of the 1,000 potentially recruited motifs were present in the overlapping region of TEs and candidate donors (Additional file 3), whereas the donors were situated proximal to the acceptors (Additional file 4). We also calculated the binding capacity of network-derived motifs (Additional file 5).

Discussion

Evidence supporting the recruitment of pre-existing *cis*-regulatory elements from non-photosynthetic genes into C_4 genes

In the present study, we evaluated the possibility that a motif may have been recruited if it satisfies two criteria. First, this motif pre-exists in the neighboring genes surrounding an ortholog of a C_4 photosynthetic gene in the rice genetic regulatory network; however, the orthologs of the C_4 photosynthesis gene does not contain this motif. Second, this motif appears in the C_4 ortholog of this photosynthesis gene. Based on these two criteria, we identified 7 out of the 65 bundle-sheath specific motifs to have been recruited during C_4 emergence (Table 1). In addition to these predicted BS-specific motifs, we further examined the recruitment of *cis*-regulatory elements previously identified through experimental approaches to be associated with cell-specific expression [5], as well as motifs derived from gene regulatory networks. The results of these analyses also suggested a large-scale recruitment of pre-existing *cis*-regulatory elements from non-photosynthetic genes into C_4 photosynthetic genes during C_4 evolution (Additional file 1: Tables S5, S6 and S7; Additional files 2, 3, 4 and 5).

Table 1 Predicted bundle sheath cells specific motifs. The potentially recruited motifs are shown in red color.

GTCGAATAAG	TCGAAACG	AAACGT	CATTCGAT	ATCTCG
CCGTTT	TCGATAGG	AGACGTGT	TATTCG	GGAATC
ATTCGATACG	GTACGTGT	TCGAATCG	GTGCGATT	CAACGT
AGATCCGTGT	GTACGAGT	TTCCGT	GTTTCG	GATACG
ACGTACGAGT	CGTTTACG	TGACGT	GTACGA	CCGTAT
ACTCGATACG	TAACGCTA	TGCGAT	GATTCG	ACGCTA
ATCCGTGT	CGTTCGAT	CAAACGTG	CATTCG	GATAAG
AATCCGTGTG	ACGTGT	TCGTAT	AATACG	ACTCAC
ACCCGTTTGG	ATCCGT	GTCCTA	ACACGT	GATAGC
ACTCGTTAGT	TCGATTCG	TCACGT	TCGAGA	TCGTTA
TAGGCT	CCGTGT	ACGATA	GCTAGT	ACGAGT
CGTTTCGA	CGTTTGAC	ACGAAA	GTAAGA	TCGTGA
TGACGTGG	TCGAAT	TCGATA		

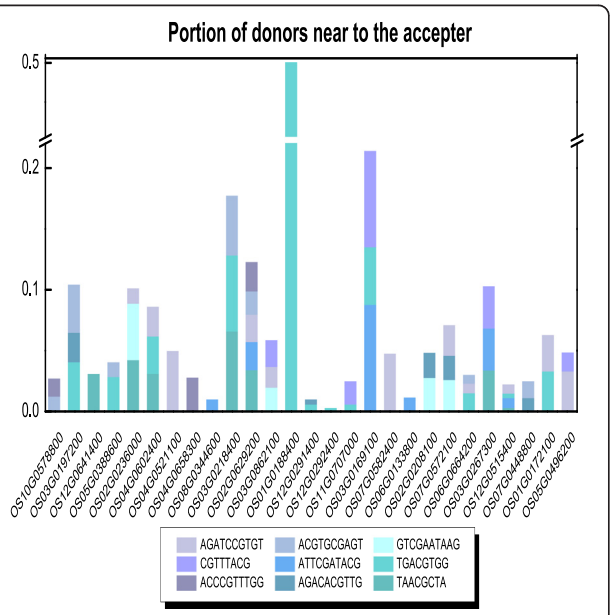

Fig. 2 Proportion of donors that were proximal to the acceptor. For each acceptor (X axis) and each potentially recruited bundle sheath-specific motif (colored block), the proportion of donors (colored segments along the Y axis) that were near to the acceptor in the chromosome where the acceptor gene resides is indicated

Evidence for the potential role of transposable element in the recruitment of C_4-specific motifs

TEs contribute to the interactions among various gene regulatory networks and the control the expression of genes [17–20] and lncRNA [21]. These can potentially contain binding sites for TFs [20]. Earlier work has suggested that about half of TF-binding sites are derived from TEs in human and mouse [14]. TEs may therefore contribute to the evolution of species-specific regulatory functions and phenotypes. In the present study, all seven putative recruited BSC-specific motifs were detected within the overlapping region of TEs and the promoter regions of candidate donor gene (Table 2). Furthermore, these donors are located near the acceptors in one chromosome, suggesting that TEs may have played an important role in the recruitment process. Similar results were obtained for the experimentally validated motifs and network-derived motifs (Additional file 1: Tables S5, S6 and S7; Additional files 2, 3, 4, 5 and 6). Similar to the function of TEs in human and mouse [14], the recruited motifs showed higher binding affinity to TFs. We hence propose that these putative recruited motifs might have contributed to the formation of new TF-binding sites and consequently modified the interactions among various gene regulatory networks in rice and maize. Not all of the putative recruited motifs were presented within the overlapping regions of TEs and the

Table 2 Predicted bundle sheath cell specific motifs that might have been recruited into C_4 related enzymes through transposable elements

Identified potentially recruited motifs	Transposable elements potentially involved	Potential donors
ACCCGTTTGG	18 retrotransposons	OS02G0610400
ACCCGTTTGG	160 MITEs	OS04G0617600
AGATCCGTGT	ORSgTERT00100016 (retrotransposon)	OS02G0264800
AGATCCGTGT	8 MITEs	OS02G0594100
AGATCCGTGT	7 MITEs	OS02G0656500
AGATCCGTGT	ORSgTEMT02200003 (MITE)	OS02G0672600
AGATCCGTGT	ORSgTEMT00101222 (MITE)	OS04G0640700
AGATCCGTGT	ORSgTEMT00900860, ORSgTEMT00900099, ORSgTEMT00901345 (MITEs)	OS07G0602100
AGATCCGTGT	ORSgTEMT00100961, ORSgTEMT00100458, ORSgTEMT00100087 (MITEs)	OS07G0603500
AGATCCGTGT	15 MITEs	OS07G0618600
ATTCGATACG	69 MITEs	OS02G0602600
ATTCGATACG	12 MITEs	OS03G0225500
ATTCGATACG	9 MITEs	OS03G0272300
CGTTTACG	9 transposons	OS03G0265900
CGTTTACG	ORSgTETNOOT00014 (transposon)	OS03G0819700
GTCGAATAAG	659 MITEs	OS02G0194000
GTCGAATAAG	ORSgTEMT00502552 (retrotransposon)	OS02G0195500
GTCGAATAAG	ORSgTEMT01900013, ORSgTEMT01900038, ORSgTEMT01900021 (MITEs)	OS02G0266500
TAACGCTA	7 MITEs	OS08G0523600
TAACGCTA	ORSgTETNOOT01500 (transposon)	OS08G0554000
TGACGTGG	ORSgTEMT00100324 (MITE)	OS03G0183300

promoter regions of candidate donor genes (Table 1), thereby suggesting other mechanisms for the emergence of these motifs during C_4 evolution.

Implications of transposon-driven recruitment of *cis*-regulatory elements to the evolution of C_4 photosynthesis

C_4 photosynthesis differs from C_3 photosynthesis in various aspects, including recruitment of new decarboxylation enzymes, re-adjustment of nitrogen metabolism, starch metabolism, and partitioning of the photosynthetic enzymes or proteins into bundle sheath and mesophyll cells [22]. Given these large number of differences between C_3 and C_4 photosynthesis, it is remarkable that C_4 photosynthesis has independently emerged in more than 60 lineages [23]. Furthermore, the emergence of C_4 photosynthesis occurred within a relatively short geological period. This is because 40 million years ago, the global atmospheric CO_2 concentration dropped, and C_4 photosynthesis started to show its competitive advantage over C_3 photosynthesis [1], thereby eventually resulting in C_4 photosynthesis 20 million years later [24]. How can such a complex trait have evolved in such a short timeframe? This study provides new sequence-based evidences that recruitment of pre-existing motifs might have been a mechanism for C_4 evolution, which in turn may have contributed to the rapid evolution of C_4 photosynthesis.

Furthermore, we showed that the new regulatory mechanism involving C_4 photosynthesis might have been created through transposon-mediated motif movements. Genome duplication has been regarded as a major mechanism responsible for the creation of material for neofunctionalization or creation of new genes during C_4 emergence [25]. However, several recent analyses have shown that the copy number of C_4-related genes are not necessarily higher than those in C_3 species [26]. Transposon-driven creation of new genes hence might have been used as an alternative mechanism for the creation of novel regulatory mechanisms for C_4-related genes. Furthermore, considering that during C_4 evolution, not only those motifs from the promoter regions, but also those in the coding sequences were potentially recruited [26–28]. Therefore, transposons were utilized as ideal mechanism for the recruitment of regulatory motifs because these can mobilize elements without particular location preferences [21]. Considering that there are relatively a lower number of whole genome duplication events during the evolution of land plants, this transposon-driven emergence of new genes might have been the predominant mechanism that has substantially contributed to the rapid evolution of new functions during the evolution of C_4 photosynthesis of the low-CO_2 oligocene period [24]. Additional experimental evidence is needed to test this potential mechanism.

Table 3 Potentially recruited bundle sheath cells specific motifs have higher binding affinity with TFs compared to randomly generated sequence of the same length

TFBS_ID	TF Name	Relative MR	TFBS_ID	TF_Name	Relative MR
M00819	Knox3	76.92	M00506	LIM1	166.67
M00937	TGA1a	125.00	M00443	Opaque-2	66.67
M00442	ABF	43.48	M00653	OCSBF-1	20.41
M00660	RITA-1	28.57	M01186	STF1	58.82
M00788	EmBP-1b	250.00	M00936	HBP-1a	-
M00441	GBF	1000.00	M00697	HBP-1b	166.67
M00366	EmBP-1	27.03	M01156	BZR1	8.00
M00654	OSBZ8	200.00	M01065	ABZ1	100.00
M00700	ROM	-	M00401	ABF1	71.43

Conclusions

The present study has provided sequence-based evidence that suggests that transposon-mediated movement of motifs might have played a role in the formation of new *cis*-regulatory elements during the evolution of C_4 photosynthesis. More experiments are needed to test this possibility. However, if this is true, then this may serve as a possible mechanism for the rapid emergence of C_4 photosynthesis within a relatively short geological period during the Oligocene [24].

Methods

The whole analysis pipeline was composed of three sessions: 1) analysis of the possibility of motif recruitment; 2) analysis of whether TE-mediated motif movement served as the mechanism responsible for the observed motif recruitment; 3) analysis of whether the recruited motifs served as binding targets of TFs. The pipeline of the analysis is shown in Fig. 4, and the details are described in the following sections.

Gene regulatory network reconstruction and motif prediction based on the network

The rice and maize GRNs were built using a PCA-CMI algorithm [12] using rice and maize transcriptomics data. With the constructed maize gene regulatory network, we classified the genes into communities with Markovian clustering algorithm [29] (MCL, http://micans.org/mcl/), and in each community of genes, we predicted motifs by using the Weeder2.0 software [30] (http://159.149.160.51/modtools/). We obtained a total of 54 communities and 1,649 motifs (hereby defined as network-derived motifs).

De novo prediction of BS-cell specific motifs

We downloaded transcriptomics data for both the BS cells and mesophyll cells [31]. A total of 1,045 genes that showed relatively higher expression levels in BS cells were classified into clusters by K-mean clustering, with the number of clusters selected based on Figure of merits (FOM) using the R package clValid (http://cran.r-project.org/web/packages/clValid/index.html). The motifs of genes of each cluster were predicted by using Weeder2.0 using the sequence 3 kb upstream of the transcription start site (TSS). We obtained 65 motifs in BS-specific genes (Table 1). These motifs were annotated as BS cell-specific motifs.

Distribution of motifs in promoter regions of C_4-related genes of maize and its orthologs in rice

When a particular motif (orange block, Fig. 1) is recruited into an acceptor gene (purple block/dot, Fig. 1) from a neighboring gene (orange dot, Fig. 1) in the gene regulatory network (GRN, Fig. 1), it is necessary that the promoter region of the donor gene contains this motif and the acceptor lacks this motif prior to the recruitment event (Fig. 1). In the present study, we focused on motif recruitment into 78 C_4 genes [31] (Additional file 1: Table S8).

As illustrated in Fig. 1, we first scanned the sequence 3 kb upstream of the TSS (downloaded from Ensembl Plant, http://plants.ensembl.org/index.html) in the C_4 genes and their orthologous genes in rice to check whether a particular motif was present or not. For a *cis*-regulatory element validated experimentally to be involved in C_4 photosynthesis, we aligned the element to the promoter sequences of C_4 genes and their orthologous genes in rice. For predicted motifs based on gene regulatory network, we used

Table 4 The identified transcription factors which showed high binding affinities to BS cell specific motifs and brief description of the TFs

TFs	*Binding motifs*	*Description of TFs*
TGA1b, TGA1a	TGACGTGG	The components of PSI, PET and PSII parts are co-regulated by TGA1b [37].
OSBZ8	TGACGTGG	TF highly expressed in salt tolerance cultivars [38].
EmBP-1	TGACGTGG,	Can activate transcription from a truncated promoter containing a pentamer of the Opaque-2 site in yeast cells.
GBF	TGACGTGG	GBF regulates photosynthetic electron transfer, C_4 genes, enzymes involved in the Calvin cycle, PSII and PSI [37].
RITA-1	TGACGTGG	Its ortholog in Arabidopsis plays an important role in regulation of light-induced genes (UniproKB, http://www.uniprot.org)
OCSBF-1	TGACGTGG,	The basal portion of a leaf has a 40-fold to 50-fold higher level of OCSBF-1 transcript than the apical portion of a leaf [39].
Opaque-2	TGACGTGG	Opaque-2 controls the expression of a cytosolic form of pyruvate orthophosphate dikinase-1 (*cyPPDK1*) [40].
STF1	TGACGTGG	STF1 can replace HY5 in photomorphogenesis and hormone signaling [41].
ROM	TGACGTGG	*ROM is related to seed storage [42].*
HBP-1a, HBP-1b	TGACGTGG	HBP-1b binds to the hexamer motif in the promoter of the 35S RNA gene of cauliflower mosaic virus [43].
ABZ1	TGACGTGG	A leucine zipper transcription factor involved in stress response [44].
ABF1	TGACGTGG	A transcription factor involved in stress response [45].
Knox3	TGACGTGG	May play a role in cytokinin biosynthesis/activation [46].
LIM1	TGACGTGG	LIM1 is related to apomixes in *Boechera* species [47].

Table 5 Transposable elements that may have played a crucial role in mediating transfer of experimentally validated motifs from the candidate donors to a C_4 acceptor gene

Motif	TEs	Candidate donors
AAAAAGG	ORSgTEMT01600223, ORSgTEMT01600962 (MITEs)	OS07G0541900
AAAAAGG	ORSiTERT00200148, ORSiTERT00200074 (retrotransposons)	OS07G0556200
AAAAAGG	ORSiTERT00200147 (retrotransposons)	OS07G0564000
AAAAAGG	ORSiTERTOOT00378, ORSiTERT00200082 (retrotransposons)	OS07G0577600
AAAAAGG	ORSgTEMT00101022 (MITEs)	OS07G0583200
AAAAAGG	ORSiTERTOOT00060 (retrotransposons)	OS07G0623100
CCCCCTCTCC	ORSiTETNOOT00105 (transposons)	OS03G0819700
GATAAG	ORSiTERT00200080 (retrotransposons)	OS01G0563500
GATAAG	ORSiTERTOOT00022, ORSgTERTOOT00073, ORSgTERTOOT00096 (retrotransposons)	OS01G0566100
GATAAG	ORSgTEMT00901457 (MITEs)	OS02G0219200
GATAAG	13 retrotransposons	OS02G0264800
GATAAG	ORSgTEMT03800107, ORSgTEMT03800095, ORSgTEMT03800034, ORSgTEMT03800059	OS04G0538800
GATAAG	9 transposons	OS06G0132400
GATAAG	ORSgTEMT00400014 (MITEs)	OS06G0190800
GATAAG	ORSgTEMT03000026, ORSgTEMT03000052 (MITEs)	OS07G0539700
GATAAG	ORSgTEMT03000026, ORSgTEMT03000050, ORSgTEMT01601523, ORSgTEMT03000052 (MITEs)	OS07G0542400
GATAAG	ORSiTERT00200079 (retrotransposons)	OS07G0556200
GATAAG	26 MITEs	OS07G0563350
GATAAG	ORSiTERT00200147 (retrotransposons)	OS07G0564000
GATAAG	23 MITEs	OS07G0602100
GATAAG	ORSiTERTOOT00060 (retrotransposons)	OS07G0623100
GATAAG	ORSiTERTOOT00060 (retrotransposons)	OS10G0542800
GATAAG	ORSgTEMT03800044 (MITEs)	OS10G0563400
GATAAG	ORSiTETNOOT00111, ORSiTETNOOT00133 (transposons)	OS10G0572000
GATAAG	19 MITEs (transposons)	OS10G0572300
GATAAG	ORSiTETNOOT00123, ORSiTETNOOT00106 (transposons)	OS11G0132501
GATAAG	ORSiTETNOOT00119 (transposons)	OS11G0132700
GATAAG	ORSgTEMT00100550, ORSgTEMT00100547 (MITEs)	OS11G0150100
GATAAG	27 MITEs	OS11G0160700
GATAAG	11 MITEs	OS11G0180300
GATAAG	ORSiTEMT01900003, ORSgTEMT01900052, ORSgTEMT01900021 (MITEs)	OS11G0182500
GATAAG	5 retrotransposons	OS11G0202000
GATAAG	50 retrotransposons	OS11G0209200
GATAAG	62 retrotransposons	OS11G0219000
GCGCGCGT	ORSiTETNOOT00104 (transposons)	OS02G0264700

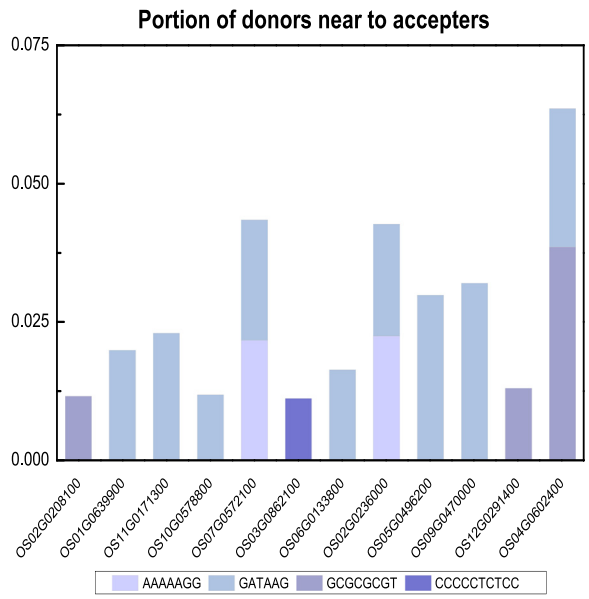

Fig. 3 Proportion of donors that were proximal to the acceptor where experimentally validated *cis*-regulatory element (motif) resides. For each acceptor (X axis) and each potentially recruited motif (colored block), the proportion of donors (colored segments along Y axis) that were near to acceptor where the acceptor gene resides is indicated

MAST [32] (http://meme.nbcr.net/meme/tools/mast) to check its distribution in the promoters of genes in the genome.

We then examined whether the motifs differentially existed between C_4 genes and their orthologous C_3 genes. For those motifs differentially distributed between C_4 and C_3 orthologs, we examined its distribution in the rice gene regulatory network to determine whether there is a possibility that motifs in C_4 orthologs were recruited from the neighboring genes.

Distribution of the candidate acceptors and donors in the chromosome

To assess whether the donor genes were proximal to the acceptor gene, we identified the donors residing within a region around 1/10 of the length of chromosome surrounding the acceptor gene (i.e., $d(accepter, donor) < 0.1$). The length and number of genes in all 12 rice chromosomes were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/assembly/GCF_000005425.2). The genes locus and description were downloaded from RAP-DB (http://rapdb.dna.affrc.go.jp/). The distance between acceptor and donor was calculated as follows:

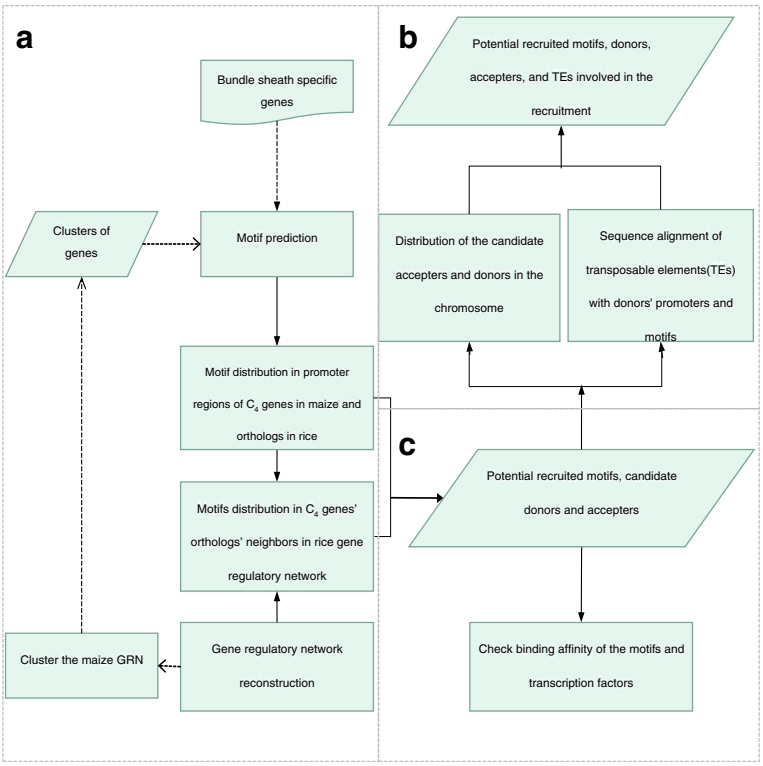

Fig. 4 Pipeline for the analysis of motif recruitment. **a** analysis of the possibility of motif recruitment; **b** analysis to determine whether a transposable element is involved in the recruitment; **c** analysis to determine whether a recruited motif is a potential binding target of transcription factors

$$d(accepter,\ donor) = \frac{donor\ start\ site - accepter\ start\ site}{chromosome\ length}$$

Sequence alignment of TEs with the promoters and motifs of donor genes

We further examined whether a motif resides within the overlapping region between TEs and the promoter region of the donor gene. To do this, we first checked the distribution of motifs in different categories of TEs, i.e., retrotransposons, class II transposons, including the miniature inverted-repeat transposable elements (MITEs), which have earlier been shown to be important in determining species diversity in *Oryza sativa* [33–35]. The sequences of these TEs were downloaded from Plant Repeat Databases [36] (http://plantrepeats.plantbiology.msu.edu/index.html). We aligned the sequences of TEs containing a particular motif with the promoter regions of candidate donors of this motif by using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to identify the overlapping region, and checked the distribution of motifs across the overlapping region.

Enrichment of motifs in TF binding sites

To explore whether a recruited motif can potentially function as a binding site for TFs, we aligned motifs with the position weight matrix information for TFs (downloaded from TRANSFAC) (http://www.gene-regulation.com/pub/databases.html). To test whether the motif enrichment was statistically significant, for each motif, we aligned to the TF binding sites PWM, we randomly constructed 1,000 short sequences with the same length of the motif and aligned these to TF binding sites PWM. For both the potential recruited motifs and the randomly constructed sequences, we calculated the match ratio (MR), which is defined as $MR(elements,\ TF) = \frac{|elements\ matched\ to\ the\ TF\ bingding\ site|}{|elements|}$; where $|S|$ is the size of a set S, and *elements* can either be the potentially recruited motifs set or random sequences set of the same length as the potentially recruited motifs. The binding affinity of the motifs and the TFs were then assessed by the relative match ratio (relative MR) $relative\ MR(motifs,\ TF) = \frac{MR(motifs,\ TF)}{MR(random\ elements,\ TF)}$; where random elements are random sequences of the same length as potentially recruited motifs. Only elements with lengths > 5 were considered in this study.

Availability of supporting data

The data sets used to construct GRN for rice and maize in the present study along with their NCBI accession numbers are listed in Additional file 1: Tables S1 and S2 (http://www.ncbi.nlm.nih.gov/sra/). The sequences of genes for both rice and maize were downloaded from Ensembl Plant (http://plants.ensembl.org/index.html). Location and description of rice genes were downloaded from (RAP-DB, http://rapdb.dna.affrc.go.jp/). Information on genomes of *Oryza sativa* was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/assembly/GCF_000005425.2). Sequences of transposable elements in *Oryza sativa* were downloaded from a plant repeat database (http://plantrepeats.plantbiology.msu.edu/index.html). Sequences of TF binding sites were downloaded from TRANSFAC (Additional file 1: Table S5). The list of C_4 genes is presented in Additional file 1: Table S8.

Additional files

Additional file 1: Table S1. Basic information on the collected RNA-SEQ data on rice mature leaves that was used in the construction of the rice gene regulatory network. **Table S2.** Basic information of collected RNA-SEQ data on maize mature leaves that was used in the construction of the maize gene regulatory network. **Table S3.** Candidate acceptors and donors of potentially recruited BS cell-specific motifs. Those donors that were not farther than 10 genes along the same chromosome where the acceptor gene resides are listed. A brief description of the acceptor gene is also provided. **Table S4.** Brief description of TFs binding sites obtained from the TRANSFAC database. **Table S5.** Experimentally validated cell-specific motifs. Those potentially recruited motifs during C_4 evolution are labeled as red. **Table S6.** Candidate acceptors and donors of potentially recruited C_4 related motifs were experimentally validated. Donors not farther than 10 genes along the same chromosome where the acceptor gene resides are listed. A brief description of the acceptor gene is also provided. **Table S7.** The potentially recruited experimentally validated motifs show higher binding affinities with TFs compared to randomly generated sequences of the same length as the motif. **Table S8.** List of C_4 genes used in the analysis of the present study. These genes were obtained from Li et al. 2010. (PDF 360 kb)

Additional file 2: Potentially recruited network-derived motifs related to C4 photosynthesis. (XLSX 15 kb)

Additional file 3: Potentially recruited network-derived motifs located within the overlapping region of transposable elements and the promoter region of the donor gene. (XLSX 1074 kb)

Additional file 4: Candidate acceptors and donors of the recruited network-derived motifs, along with a brief description of each acceptors. Only those donor genes that were not farther than 10 genes from the acceptor gene are listed. (XLSX 110 kb)

Additional file 5: Potentially recruited network-derived motifs detected within transcription factor binding sites (TFBS). (XLSX 27 kb)

Additional file 6: Network-derived motifs with the probability of 'A,' 'T,' 'C,' 'G' in each site. (PDF 708 kb)

Competing interests

The authors declare no competing interests.

Authors' contribution

XGZ conceived the study and wrote the manuscript. CC conducted all analyses, wrote the manuscript, and generated the figures. JX predicted the bundle sheath cells specific motifs. GZ constructed the rice and maize gene regulatory network. XGZ and GZ read and approved the final manuscript. All authors read and approved the final manuscript.

Funding

The Bill and Melinda Gates Foundation (#OPP1014417) and Ministry of Science and Technology of China (#2015CB150104) supported this study.

Received: 23 August 2015 Accepted: 24 February 2016
Published online: 08 March 2016

References

1. Zhu X-G, Long SP, Ort DR. What is the maximum efficiency with which photosynthesis can convert solar energy into biomass? Curr Opin Biotechnol. 2008;19:153–9.
2. Raghavendra AS, Sage RF. C_4 Photosynthesis and Related CO_2 Concentrating Mechanisms. Volume 32. Springer Science & Business Media; Springer Netherlands, 2010.
3. Covshoff S, Hibberd JM. Integrating C_4 photosynthesis into C_3 crops to increase yield potential. Curr Opin Biotechnol. 2012; 23:209–14.
4. Hibberd JM, Covshoff S. The regulation of gene expression required for C_4 photosynthesis. Annu Rev Plant Biol. 2010;61:181–207.
5. Sheen J. C_4 gene expression. Annu Rev Plant Biol. 1999;50:187–217.
6. Aubry S, Brown NJ, Hibberd JM. The role of proteins in C_3 plants prior to their recruitment into the C_4 pathway. J Exp Bot. 2011;62:3049–59.
7. Langdale JA. C_4 cycles: past, present, and future research on C_4 photosynthesis. Plant Cell. 2011;23:3879–92.
8. Von Caemmerer S, Quick WP, Furbank RT. The development of C_4 rice: current progress and future challenges. Science. 2012;336:1671–2.
9. Roider HG, Kanhere A, Manke T, Vingron M. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics. 2007;23:134–41.
10. Palumbo MJ, Newberg LA. Phyloscan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. Nucleic Acids Res. 2010;38: W268-274.
11. Jia H, Li J. Finding transcription factor binding motifs for co-regulated genes by combining sequence overrepresentation with cross-species conservation. J Probab Stat. 2012;2012. http://dx.doi.org/10.1155/2012/830575
12. Zhang X, Zhao X-M, He K, Lu L, Cao Y, Liu J, Hao J-K, Liu Z-P, Chen L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics. 2012;28:98–104.
13. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
14. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, Snyder MP, Wang T. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 2014;24:1963-76.
15. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet. 2013;45:836–41.
16. Terashima I, Hanba YT, Tazoe Y, Vyas P, Yano S. Irradiance and phenotype: comparative eco-development of sun and shade leaves in relation to photosynthetic CO_2 diffusion. J Exp Bot. 2006;57:343–54.
17. Wang T, Zeng J, Lowe CB, Sellers RG, Salama SR, Yang M, Burgess SM, Brachmann RK, Haussler D. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci. 2007;104:18613–8.
18. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nat Genet. 2011;43:1154–9.
19. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 2012;46:21–42.
20. Jacques P-E, Jeyakani J, Bourque G. The majority of primate-specific regulatory sequences are derived from transposable elements. PLoS Genet. 2013;9, e1003504.
21. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M, Feschotte C. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013;9, e1003470.
22. Gowik U, Bräutigam A, Weber KL, Weber APM, Westhoff P. Evolution of C_4 photosynthesis in the genus *Flaveria*: how many and which genes does it take to make C_4? Plant Cell. 2011;23:2087–105.
23. Sage RF, Christin PA, Edwards EJ. The C_4 plant lineages of planet earth. J Exp Bot. 2011;62:3155–69.
24. Christin PA, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V, Salamin N. Oligocene CO_2 decline promoted C_4 photosynthesis in grasses. Curr Biol. 2008;18:37–43.
25. Monson RK. The origins of C_4 genes and evolutionary pattern in the C_4 metabolic phenotype. In: Sage RF, Monson RK, editors. C_4 plant biology. San Diego, CA, USA: Academic Press; 1989. p. 377–410.
26. Kajala K, Brown NJ, Williams BP, Borrill P, Taylor LE, Hibberd JM. Multiple Arabidopsis genes primed for recruitment into C_4 photosynthesis. Plant J. 2012;69:47–56.
27. Gowik U, Burscheidt J, Akyildiz M, Schlue U, Koczor M, Streubel M, Westhoff P. *cis*-regulatory elements for mesophyll-specific gene expression in the C_4 plant *Flaveria trinervia*, the promoter of the C_4 phosphoenolpyruvate carboxylase gene. Plant Cell. 2004;16:1077–90.
28. Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, Kajala K, Hibberd JM. Independent and parallel recruitment of preexisting mechanisms underlying C_4 photosynthesis. Science. 2011;331:1436–9.
29. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84.
30. Pavesi G, Mereghetti P, Zambelli F, Stefani M, Mauri G, Pesole G. MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. Nucleic Acids Res. 2006;34:W566–70.
31. Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR. The developmental dynamics of the maize leaf transcriptome. Nat Genet. 2010;42:1060–7.
32. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics. 1998;14:48–54.
33. Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, Okumoto Y, Tanisaka T, Wessler SR. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature. 2009;461:1130–4.
34. Kuang H, Padmanabhan C, Li F, Kamei A, Bhaskar PB, Ouyang S, Jiang J, Buell CR, Baker B. Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: new functional implications for MITEs. Genome Res. 2009;19:42–56.
35. Yang G, Lee Y-H, Jiang Y, Shi X, Kertbundit S, Hall TC. A two-edged role for the transposable element Kiddo in the rice ubiquitin2 promoter. Plant Cell. 2005;17:1559–68.
36. Ouyang S, Buell CR. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 2004;32 suppl 1:D360–3.
37. Yu X, Zheng G, Shan L, Meng G, Vingron M, Liu Q, Zhu XG. Reconstruction of gene regulatory network related to photosynthesis in Arabidopsis thaliana. Frontiers in Plant Sci. 5:273.doi: 10.3389/fpls.2014.00273.
38. Mukherjee K, Choudhury AR, Gupta B, Gupta S, Sengupta DN. An ABRE-binding factor, OSBZ8, is highly expressed in salt tolerant cultivars than in salt sensitive cultivars of indica rice. BMC Plant Biol. 2006;6:18.
39. Singh K, Dennis ES, Ellis JG, Llewellyn DJ, Tokuhisa JG, Wahleithner JA, Peacock WJ. OCSBF-1, a maize ocs enhancer binding factor: isolation and expression during development. Plant Cell. 1990;2:891–903.
40. Maddaloni M, Donini G, Balconi C, Rizzi E, Gallusci P, Forlani F, Lohmer S, Thompson R, Salamini F, Motto M. The transcriptional activator Opaque-2 controls the expression of a cytosolic form of pyruvate orthophosphate dikinase-1 in maize endosperms. Mol Gen Genet MGG. 1996;250:647–54.
41. Song YH, Yoo CM, Hong AP, Kim SH, Jeong HJ, Shin SY, Kim HJ, Yun D-J, Lim CO, Bahk JD. DNA-binding study identifies C-box and hybrid C/G-box or C/A-box motifs as high-affinity binding sites for STF1 and LONG HYPOCOTYL5 proteins. Plant Physiol. 2008;146:1862–77.
42. Verdier J, Thompson RD. Transcriptional regulation of storage protein synthesis during dicotyledon seed filling. Plant Cell Physiol. 2008;49:1263–71.
43. Tabata T, Nakayama T, Mikami K, Iwabuchi M. HBP-1a and HBP-1b: leucine zipper-type transcription factors of wheat. EMBO J. 1991;10:1459–67.
44. Sell S, Hehl R. Functional dissection of a small anaerobically induced bZIP transcription factor from tomato. Eur J Biochem. 2014;271:4534–44.
45. Yoshida T, Fujita Y, Maruyama K, Mogami J, Todaka D, Shinozaki K, Yamaguchi-Shinozaki K. Four Arabidopsis AREB/ABF transcription factors function predominantly in gene expression downstream of SnRK2 kinases in abscisic acid signalling in response to osmotic stress. Plant Cell Environ. 2015;38:35–49.
46. Azakhsh M, Kirienko AN, Zhukov VA, Lebedeva MA, Dolgikh EA, Lutova LA. KNOTTED1-LIKE HOMEOBOX 3: a new regulator of symbiotic nodule development. J Exp Bot. 2015;66:7181–95.
47. Corral JM, Vogel H, Aliyu OM, Hensel G, Thiel T, Kumlehn J, Sharbel TF. A conserved apomixis-specific polymorphism is correlated with exclusive exonuclease expression in premeiotic ovules of apomictic Boechera species. Plant Physiol. 2013;163:1660–72.