Genomics 107 (2016) 199–207

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Evolution of DUF1313 family members across plant species and their association with maize photoperiod sensitivity

Jia Li [a,b,1], Erliang Hu [a,b,1], Xueying Chen [a,b], Jie Xu [a,b], Hai Lan [a,b], Chuan Li [a,b], Yaodong Hu [c], Yanli Lu [a,b,*]

[a] Maize Research Institute, Sichuan Agricultural University, Chengdu 611130, Sichuan, China
[b] Key Laboratory of Biology and Genetic Improvement of Maize in Southwest Region, Ministry of Agriculture, China
[c] Institute of Animal Genetics and Breeding, College of Animal Science and Technology, Sichuan Agricultural University, Chengdu 611130, Sichuan, China

ARTICLE INFO

Article history:
Received 6 October 2015
Received in revised form 16 December 2015
Accepted 4 January 2016
Available online 7 January 2016

Keywords:
Association analysis
DUF1313 family
Maize
Phylogenetic analysis
Selection pressure

ABSTRACT

Proteins of the DUF1313 family contain a highly conserved domain and are only found in plants; they play important roles in most plant functions. In this study, 269 DUF1313 genes from 81 photoautotrophic species were identified; they were classified into three major types based on the amino acid substitutions in the conserved region: IARV, I(S/T/F)(K/R)V, and IRRV. Phylogenic tree constructed from 51 DUF1313 genes from graminoids revealed three clades: A, B1, and B2. Clade B1 was found to have undergone episodic positive selection after a gene duplication event and included four amino acid sites under positive selection. The association between DUF1313 family members and traits investigated in maize indicated that three of four genes (GRMZM2G025646, GRMZM5G877647, GRMZM2G359322, and GRMZM2G382774) were associated with the target traits such as days to silking, days to tasselling, and plant height. The nucleotide diversity of the most primitive and highly conserved DUF1313 gene, *ELF4-like4*, was the highest in *Tripsacum* and the lowest in maize. Tajima's *D* and Fu and Li's *D* tests revealed that significant purifying selection had occurred in the coding sequence region of this DUF1313 gene in teosinte and maize. No significant signal was detected in the 5′-untranslated region of this gene in each of the three species (maize, teosinte, and *Tripsacum*) or in any gene regions of *Tripsacum*. Phylogenetic analyses revealed that the 103 accessions of maize, teosinte, and *Tripsacum* can be grouped into four clades based on the *ELF4-like4* gene sequence similarity. Thus, this gene can be used to determine the relationships between maize and its relatives, and the DUF1313 family members and alleles identified in this study might be valuable genetic resources for molecular marker-assisted breeding in maize.

1. Introduction

Maize is a facultative short-day plant: it flowers earlier when grown under short-day conditions [1]. Although most temperate maize cultivars do not respond to changes in day length, tropical maize cultivars are usually sensitive to such variation. In addition, like the wild relatives, tropical maize shows greater genetic variation than temperate maize. Thus, when grown in temperate regions, tropical maize cultivars require improved adaptability to long-day conditions [2]. Plant breeders are currently investigating strategies to improve the adaptability of tropical maize by exploring its high level of genetic diversity.

The domain of unknown function (DUF) is a protein domain with unknown function. Protein families with such domains have been deposited together in the Pfam (http://pfam.xfam.org/family) database and labeled using the prefix "DUF" followed by a number [3]. More than 20% of all protein domains are currently annotated as DUF, many of which are highly conserved, indicating that they play important roles in the biology of the plants in which they are found [4,5]. The DUF1313 family consists of several hypothetical plant proteins having length of around 100 residues. DUF1313 genes are known to be only present in plants and contain a highly conserved domain [6]. *Early Flowering 4* (*ELF4*) belongs to the DUF1313 family of plant proteins [7]. In *Arabidopsis*, *ELF4* encodes a protein of 111 amino acids without a known protein signature. The *Arabidopsis elf4* mutants flower early under short-day conditions because of the reduced ability to sense variations in day length. The tendency of *elf4* mutants to flower early is thought to be attributed to the elevated expression of *CONSTANS* (*CO*) [6]. Furthermore, under red-light conditions, the normal expression level of *ELF4* is induced by *phytochrome B* (*PHYB*), whereas *phyb* decreases the expression of *ELF4*. *ELF4* also plays a role in the *PHYB*-mediated induction of de-etiolation of seedlings under red-light conditions; *elf4* mutant seedlings exhibit relatively weak sensitivity to red light, leading to delayed de-etiolation [8]. The *ELF4*-homologous gene *Die Neutralis* (*DNE*) was found to affect the rhythmic expression of clock genes in peas under continuous light and dark conditions [9].

Traits associated with target genes have been previously identified; this approach can be used to determine the relationship between the

* Corresponding author at: Maize Research Institute, Sichuan Agricultural University, Chengdu 611130, Sichuan, China.
E-mail address: yanli.lu82@hotmail.com (Y. Lu).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ygeno.2016.01.003
0888-7543/

DUF1313 gene family members and photoperiod sensitivity in maize. Association mapping has been successfully used to explore new functional genes in maize. Wilson [10] used an association approach to evaluate six candidate maize genes associated with kernel starch biosynthesis. Liu et al. [11] used a diverse group of 368 maize inbreds to evaluate the association between nucleic acid variation of each *ZmDREB* gene and drought tolerance. They found a significant association between the genetic variation of *ZmDREB2.7* and drought tolerance during the seedling stage. Buckler et al. [12] used a large nested association mapping (NAM) population containing 25 recombinant inbred line (RIL) populations to investigate variations in flowering time. They found numerous small-effect quantitative trait loci (QTLs) that were common among families, as well as allelic effects that differed across founder lines. These findings suggest that identifying favorable allelic variations within either gene families or a single gene is possible by using a diverse population or genetic linkage population. This study aimed to (1) identify DUF1313 family genes in plants and analyze the phylogenetic relationships among them; (2) use association mapping to identify the gene members associated with photoperiod sensitivity in maize; (3) evaluate the associations between nucleic acid variations of each candidate gene and photoperiod sensitivity by using a NAM population; and (4) clone *ELF4-like4* homologous genes and analyze the evolutionary history of maize and its wild relatives. These results might allow better understanding of the evolutionary origins of the DUF1313 gene family and its association with photoperiod sensitivity. In addition, the gene members and alleles associated with photoperiod sensitivity identified in this study might become valuable genetic resources for molecular marker-assisted breeding in maize.

2. Materials and methods

2.1. Data sets

The NCBI (http://www.ncbi.nlm.nih.gov/), JGI (http://phytozome.jgi.doe.gov/pz/portal.html), Pfam (http://pfam.xfam.org/), and UniprotKB (http://www.uniprot.org/uniprot/) databases were used to search for *Arabidopsis ELF4* homologues in plant species. An association panel of 513 maize inbred lines was used to identify single nucleotide polymorphisms (SNPs) in genes of the DUF1313 family that are significantly associated with photoperiod sensitivity. This panel included 556,809 high-quality SNPs [13,14]; it was downloaded from http://www.maizego.org/Resources.html. The traits related to photoperiod sensitivity, including days to silking (DS), days to tasselling (DT), pollen shed (PS), plant and ear height (PH and EH), tassel length (TL), tassel branch number (TBN), and the number of leaves above the ear (LN) [15], were investigated in five different environments, i.e., Ya'an (30°N, 103°E), Sanya (18°N, 109°E), and Yunnan (25°N, 102°E) in 2009, and Guangxi (23°N, 110°E) and Yunnan in 2010. The SNPs localized to four gene regions of the DUF1313 family in maize were extracted for further candidate gene based-association mapping of the photoperiod sensitivity-related traits.

Genotype-by-sequencing (GBS) analyses were conducted on a NAM population containing 25 RIL populations [12]; eight photoperiod sensitivity traits, including DS, DT, days to silking at growing degree day (GDD) (GDD-DS), days to tasselling at GDD (GDD-DT), anthesis-silking interval at GDD (GDD-ASI), tassel length (TL), and PH and EH, were measured in multiple environments during 2006–2007. The GBS and phenotypical data were downloaded from http://www.panzea.org. The SNPs localized to the four maize DUF1313 genes were extracted to verify the associations between favorable alleles and traits related to photoperiod sensitivity in the 25 RIL populations.

2.2. Identification and reconstruction of phylogeny of the DUF1313 gene family in plants

In this study, *ELF4* homologues in plant species were searched for in the Pfam, UniprotKB, NCBI, and JGI databases by using BLASTP and the amino acid sequence of *AtELF4* (NC_003071). To ensure that the complete sequence of each species was retrieved, the results of the search were repeatedly resubmitted to the database search as query terms until no new genes were found. Since the contents of different databases partially overlapped, the sequences of DUF1313 genes were integrated and redundancy was removed by manual screening. Subsequently, all gene sequences obtained were analyzed using the online software programs SMART and Pfam to predict the conserved domains and remove redundant and low-homology sequences. DUF1313 family members were then individually screened and identified in the different plant species.

MAFFT software has greater speed and higher accuracy than CLUSTALW for alignment of amino acid or nucleotide sequences [16]. Therefore, multiple sequence alignment analysis was performed using MAFFT online software version 7 (http://mafft.cbrc.jp/alignment/server/). The conserved amino acid motifs were analyzed using MEME–MAST program (http://meme-suite.org/tools/meme). Neighbor-Joining (NJ) trees were constructed using calculated amino acid substitution models in MEGA v5.0 [17]. A Jones–Taylor–Thornton + Gamma-distributed model was estimated using ProtTest v2.4 [18], and a bootstrap test with 1000 replications was performed. Maximum likelihood (ML) trees were constructed using PhyML v3.1 [19], with default parameters and bootstrap test replication number set to 1000. Further, Bayesian inference (BI) trees were constructed using MrBayes v3.2 [20]. Four Markov Chain Monte Carlo (MCMC) chains were run for 2,000,000 generations. The trees were then viewed in Figtree v1.4.2. (http://tree.bio.ed.ac.uk/software/).

2.3. dN:dS ratio estimates and selection pressure of DUF1313 genes in graminoids

How selection pressure operates on the major graminoid phylogenetic clades was determined by extracting 51 DUF1313 genes from different graminoid species. The sequence alignment of the 51 DUF1313 proteins and the corresponding coding sequences (CDSs) were converted to a codon alignment by using PAL2NAL (http://www.bork.embl.de/pal2nal). Phylogenetic trees were reconstructed using the relaxed clock model in BEAST v1.6.1 [21]. The general time reversible + I + G substitution model was selected using the jModelTest, and a Yule tree prior was used for the analyses. The analysis was run for at least 200,000,000 generations, while sampling for every 1000 states. Next, a branch model (model = 2, NS sites = 0) was used to estimate the non-synonymous to synonymous site ratios (*dN*:*dS*) for each major clade by using the CODEML program in the PAML 4.7 [22] package. The branches of each major clade were set to the foreground, and the remaining branches of the full gene tree were assigned to the background. The *dN*:*dS* values were estimated using the following equation: fix_omega = 0; omega = 1. Duplication events of DUF1313 genes were inferred using Notung 2.6 [23]. Notung offers a unified framework for incorporating duplication-loss parsimony into phylogenetic tasks. First, a rooted phylogenetic tree (graminoid species tree) and a gene tree (DUF1313 gene tree) were constructed using NCBI Taxonomy Browser (http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) and gene nucleotide sequences, respectively. Subsequently, both the trees were used to infer duplication and loss of gene events during species evolution by using Notung 2.6 [23].

Whether the major ancestral gene duplications were followed by a strong shift in selective constraints was determined by setting these nodes as the foreground branches in a branch model analysis (model = 2, NS sites = 0), with all remaining branches of the full gene tree set as background. The amino acid sites that had experienced a shift in selective pressure throughout the evolution of the gene family were identified by conducting branch-site tests. Subsequently, branch-site model analyses were performed for the branches with the ancestral gene duplications set as foreground and the remaining subtree branches set as background. The branch-site model (model = 2, NS sites = 2)