Agri Gene 3 (2017) 87–98

Contents lists available at ScienceDirect

Agri Gene

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

Genomic organization and structural diversity of germin-like protein coding genes in foxtail millet (*Setaria italica* L.)

Dipnarayan Saha *[,1], Rajeev Singh Rana [1], Lalit Arya, Tapan Kumar Mondal [2]

Genomics Resources Division, ICAR-National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi 110 012, India

ARTICLE INFO

Article history:
Received 16 November 2016
Received in revised form 15 December 2016
Accepted 16 December 2016
Available online 20 December 2016

Keywords:
Foxtail millet
Genome-wide analysis
Gene expression dynamics
Germin-like proteins
Protein homology modeling

ABSTRACT

Germin-like protein (GLP) genes in plants constitute a multigene family that functions in a variety of biological processes, such as plant growth and development and respond to abiotic and biotic stresses. Available genome sequences of foxtail millet (*Setaria italica*) have facilitated the identification and delineation of the GLP gene family, which is perhaps the first report in any millet crop. A total of 20 SiGLP genes were mapped in six of the nine foxtail millet chromosomes. The majority of these SiGLP genes, except six, clustered into five known germin (GER) groups and revealed group-specific variations in their gene structure and conserved protein motifs. The SiGLP genes shared homologous relationship within and across plant genomes as per their syntenic chromosomal regions, indicating gene duplication and expansion in course of plant evolution. Gene expression dynamics revealed four out of the 20 SiGLP genes having a putative functional role in different tissues. The presence of diverse cis-acting regulatory elements in the putative SiGLP promoter sequences and homology-based 3-D protein structure predictions revealed their possible functional diversity. The present work encompasses a molecular insight into the organization and functions of SiGLP gene family members. This may serve as a platform for functional analysis and their efficient utilizations in genetic improvement of foxtail millet and other related crops.

1. Introduction

The Germins are highly conserved proteins (Homo-hexamer glycoprotein) with more than 90% sequence identity. They were first identified as a marker protein in germinating wheat and barley embryos (Lane et al., 1993). The germins are from the cupin superfamily and typically features a motif to form a jellyroll beta-barrel structure that possesses metal binding properties and oxalate oxidase (OXO) activity (Dunwell et al., 2008). Germins or OXOs are highly specific to cereals whereas, germin-like proteins (GLPs) are ubiquitous in all land plants, including gymnosperms and moss (Breen and Bellgard, 2010). Unlike germins, GLPs are the diverse group of protein family consisting of heterogenous members with 30–70% sequence identities to the true OXOs; thus phylogenetically placed in distinct sub-groups (Breen and Bellgard, 2010). GLPs are thus actually germin motif-containing proteins devoid of any OXO activity, but with conserved sequence features and function (Dunwell et al., 2008; Zimmermann et al., 2006).

Structurally, GLPs consists of 140–150 amino acid-long conserved cupin domain with small conserved motifs of approximately 20-amino acid residues known as 'germin box' (Lane et al., 1993). The average length and molecular weight of GLPs are approximately 220 amino acids and 23.0 kDa, respectively. At the molecular level, GLPs form a multimeric structure that provides stability to various denaturing agents like proteases, temperature, and pH (Dunwell et al., 2008). They are involved in a variety of functions, such as plant growth and development, and in response to various biotic and abiotic stresses at the physiological level (Breen and Bellgard, 2010; Dunwell et al., 2008; Manosalva et al., 2009). Phylogenetically, GLPs are grouped into six germin (GER) subfamilies, one to six, with concordant enzymatic properties exhibited by the members of subfamilies (Manosalva et al., 2009). True germin or GER-1 subfamily, possessing OXO activity, is primarily reported to be involved in plant organ development, various physiological processes, such as lignification and senescence, and somatic and zygotic embryogenesis in plants (Dunwell et al., 2008; Federico et al., 2006). The other traits governed by the GLPs are tolerance to abiotic stresses, such as drought stress (Ke et al., 2009), salinity stress (Cramer et al., 2007) and metal stress (Houde and Diallo, 2008). But, the most important role attributed to GLPs is the broad-spectrum defence response in grasses (Breen and Bellgard, 2010; Dunwell et al., 2008; Manosalva et al., 2009). It was suggested that in response to pathogen infection, the OXO, SOD and H_2O_2 properties encoded by GLPs play

Abbreviations: bp, base pairs; CARES, cis-acting regulatory elements; CDS, coding sequence; GER, germins; GLP, germin-like proteins; GO, gene ontology; HMM, hidden Markov Model; MYA, million years ago; OXO, oxalate oxidases; pI, iso-electric point; RT-PCR, Reverse transcriptase-polymerase chain reaction.

* Corresponding author.
E-mail addresses: dipsaha72@yahoo.com, dipnarayan.saha@icar.gov.in (D. Saha).
[1] Present address: Division of Crop Improvement, ICAR-Central Research Institute for Jute and Allied Fibres, Kolkata 700 121, India.
[2] Present address: ICAR-National Research Centre on Plant Biotechnology, Pusa Campus, New Delhi 110 012, India.

http://dx.doi.org/10.1016/j.aggene.2016.12.001
2352-2151/

an active role in mediating oxidative burst to counter the stress response. Thus, GLPs can act as a candidate of choice to impart defence response to biotic and abiotic stresses in plants (Dunwell et al., 2008).

GLPs were reported from a variety of plant systems besides cereals, such as *Arabidopsis* (Membré et al., 2000), soybean (Lu et al., 2010), *Brassica* (Rietz et al., 2012), and peanut (Wang et al., 2013). The availability of genome sequence of various crop plants in the recent years has facilitated rapid identification, phylogenetic studies and functional analysis of GLPs in grain crops, such as rice (Manosalva et al., 2009) and barley (Zimmermann et al., 2006). Foxtail millet (*Setaria italica*), is an important food and feed crop of the arid and semi-arid tropics. Its genome sequence is available for comparative genomics utilization in bio-fuel research and C_4 photosynthesis system owing to its close genome synteny to bioenergy crop, switchgrass (Bennetzen et al., 2012; Zhang et al., 2012). Blast disease caused by the fungal pathogen, *M. grisea* is one of the major biotic stress of foxtail millet and other related millet crops, such as finger millet (*Eleusine coracana*) (Saha et al., 2016a). Candidate genes with a broad-spectrum defence response may not only address this problem in foxtail millet, but will also enrich the genomic resources in millets (Saha et al., 2016b). Thus, to identify and characterize candidates of durable defence-related genes, such as GLP, in the foxtail millet genome is of paramount interest. We discuss here a comprehensive genome-wide identification, phylogeny, evolutionary relationships and gene expression analysis to offer a strong foundation for functional studies of GLPs in foxtail millet and their utilizations in genetic improvement of millets.

2. Materials and methods

2.1. Identification of GLP genes

Gene, CDS and amino acid sequences of putative GLP proteins with cupin domain (PFAM IDs - PF00190 and PF07883) were derived from the foxtail millet genome at Phytozome v11.0.7 (https://phytozome.jgi.doe.gov/pz/portal.html#) database and cross-checked against the annotations at SiGDB of plantGDB (http://www.plantgdb.org/SiGDB/). The protein sequences were further scanned with an HMM profile at Pfam 29.0 (Finn et al., 2016). A software program Simple Modular Architecture Research Tool (SMART) (Letunic et al., 2014), was used to confirm annotation of the GLP proteins based on the presence of the conserved cupin domain. Information regarding chromosomal localization and sequence lengths was retrieved from the Phytozome database. Features of GLP proteins, such as molecular weight and isoelectric point (pI) were deduced from the compute pI/Mw tool of ExPASy (http://web.expasy.org/compute_pi/). Transmembrane domain detection of the GLP proteins was carried out using the TMHMM Server v 2.0 (Krogh et al., 2001).

2.2. Mapping of GLP genes on foxtail millet chromosomes

Putative GLP genes were mapped on foxtail millet chromosomes based on their genomic coordinates retrieved from the Phytozome database. Locations of GLP genes on the foxtail millet chromosomes were plotted in megabase scale using the software Graphical Geno Typing v2.0 (GGT 2.0) (van Berloo, 2008). The GLP genes were named from SiGLP01 to SiGLP20 according to their occurrence on the foxtail millet chromosomes in ascending order. Gene clusters were identified when two or more genes were found within the 200 kb chromosomal region (Holub, 2001). The SiGLP gene duplication types (tandem or block) were identified from the Monocots PLAZA v.3.0 (http://bioinformatics.psb.ugent.be/plaza/), which is a database of plant comparative genomics (Proost et al., 2015) and local genome synteny was checked using GEvo tool of the CoGe database (https://genomevolution.org/CoGe/GEvo.pl).

2.3. Multiple sequence alignment and phylogenetic analysis

Amino acid sequences of the cupin domain of SiGLP proteins were extracted using the SMART program (Letunic et al., 2014). The 140–150 amino acid-long cupin domain sequence of the SiGLP proteins was used for multiple sequence alignment in Clustal Omega tool (Sievers et al., 2011) and represented using BOXShade v3.21 (http://www.ch.embnet.org/software/BOX_form.html). Full-length protein sequences of the SiGLP proteins were used to reconstruct a phylogenetic tree using the Molecular Evolutionary Genetics Analysis (MEGA 5.10) program (Tamura et al., 2011). Representative GLP proteins from rice and barley were derived from Manosalva et al. (2009) to compare the groups. The amino acid sequences were aligned through ClustalW. Bootstrap replications of 1000, amino acid *p*-distance substitution model and pairwise gap deletions were the parameters chosen for constructing a phylogenetic tree using Neighbour-joining (NJ) method.

2.4. Gene structure and protein motif prediction

For gene structure and organization, including determination of exon-intron distribution pattern and intron splicing phase, the coding sequences and genomic sequences of the SiGLP genes were compared in the Gene Structure Display Server v.2.0 (GSDS2.0) program (Hu et al., 2015). The Multiple Expectation Maximisation for Motif Elicitation (MEME version 4.9.0) (Bailey et al., 2009) was used to find conserved motifs in the SiGLP protein sequences using the following parameters: maximum 10 motifs, any repetition of motifs, and 6–50 motif width.

2.5. Homolog identification and evolutionary analysis

A homologous sequence identification tool on the comparative genomics database, GreenPhylDB v4, was used to identify SiGLP gene orthologs through phylogeny and best blast mutual hits (BBMH) (http://www.greenphyl.org/cgi-bin/blast.cgi) from rice, maize, sorghum and *Brachypodium*. The sequences were further checked from the Monocots PLAZA v.3.0 database for collinear gene search. Genomic locations and sequences of the corresponding orthologs were individually retrieved from the Phytozome v11. Chromosomal regions of the foxtail millet genome sharing a syntenic relationship to rice, maize, sorghum and *Brachypodium* were derived from Zhang et al. (2012). All ortholog relationship data were visualized using the CIRCOS v0.64 (Krzywinski et al., 2009). Estimation of synonymous (d_S) and non-synonymous (d_N) substitution rates for each orthologous GLP gene pairs as well the tandemly duplicated SiGLPs were carried out using the program PAL2NAL (http://www.bork.embl.de/pal2nal/) of the codeml program in PAML (Suyama et al., 2006). The evolutionary time (T) of gene duplication or divergence (in million years ago, MYA) of the SiGLP homologs were calculated using a synonymous mutation rate of λ substitutions per synonymous site per year for the grass lineage ($T = d_S/2\lambda/10^6$, where $\lambda = 6.5 \times 10^{-9}$) (Gaut et al., 1996).

2.6. Gene expression analysis

Transcriptome data from four different tissues (root, leaf, stem, and spica) of the 'Zhang gu' cultivar available in the Foxtail millet Database (http://foxtailmillet.genomics.org.cn/page/species/index.jsp) were employed for gene expression analysis. The reads per kilobase transcript per million mapped reads (RPKM) values corresponding to the SiGLPs were extracted. Similarly, FPKM values of SiGLP genes in RNA-seq data from different plant tissues of varying stages and treatment conditions were also extracted from the Phytozome database. The RPKM/FPKM values >10, at least in one of the conditions per SiGLP gene, were $\log_2$ transformed to get relative expression ratios. All these values were then plotted as the heatmap using the program Cluster 3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm#ctv). The hierarchical clustering method was chosen for the heatmap construction