Scientia Horticulturae 236 (2018) 238–243

Contents lists available at ScienceDirect

Scientia Horticulturae

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

Identification of repetitive DNA sequences in the *Chrysanthemum boreale* genome

So Youn Won[a,*], Yoon-Jung Hwang[b], Jae-A Jung[c], Jung Sun Kim[a], Sang-Ho Kang[a], Seong-Han Sohn[a]

[a] National Institute of Agricultural Sciences, Rural Development Administration, Jeonju, 54874, Republic of Korea
[b] Department of Life Science, Sahmyook University, Seoul, 01795, Republic of Korea
[c] National Institute of Horticultural and Herbal Science, Rural Development Administration, Wanju, 55365, Republic of Korea

ARTICLE INFO

Keywords:
Asteraceae
Chrysanthemum
Fluorescence in situ hybridization
Repeat
Satellite
Transposon

ABSTRACT

We previously revealed that the *Chrysanthemum boreale* genome is highly repetitive; however, the types and nucleotide sequences of repetitive DNA in this diploid wild chrysanthemum are not known. Here, we characterized repetitive DNA sequences in the *C. boreale* genome by analysing genomic sequences obtained by Illumina sequencing and confirmed their repetitive nature by conducting fluorescence in situ hybridization (FISH) analyses. Annotation of the obtained DNA sequences revealed that microsatellite-containing genomic sequences exhibited similarity with genomic sequences in *Chrysanthemum morifolium*, indicating sequence conservation of repetitive DNA sequences between the two *Chrysanthemum* species. Two superfamilies of repetitive DNA, *Copia* and *Gypsy,* belonging to the long-terminal repeat (LTR) class of retrotransposons, are abundant in the *C. boreale* genome. We propose that *Copia* and *Gypsy* retroelements contribute to the current genome architecture of *C. boreale*. Whole genome sequencing, which is currently in progress, will reveal the extent to which these repetitive DNA sequences contribute.

1. Introduction

Repetitive DNA sequences are a major component of eukaryotic genomes and play a critical role in regulating genes and genomes (Grewal and Jia, 2007; Lisch, 2013; Mehrotra and Goyal, 2014). While the types, abundance, and proportion of repeats in a genome vary among species, repetitive sequences account for over 80–90% of the genomes of some plants, such as several members of the *Fritillaria* genus in the Liliaceae family, in which genome sizes exceed 40 Gb (Kelly et al., 2015). In plants, repeats are transcriptionally silenced by repressive epigenetic marks, which maintain genome stability and regulate gene expression (Lisch, 2013). Repeats clustered in specific genomic regions, such as the centromere or telomere, can influence chromosome structure and function, for example during cell division (Mehrotra and Goyal, 2014). Repeats are also dispersed across all chromosomes, and genetic or epigenetic variation in some repeats has been shown to influence the function and expression of juxtaposed genes (Lisch, 2013; Mehrotra and Goyal, 2014).

Transposons are a type of repetitive sequence that can move from one genomic locus to another (reviewed in (Levin and Moran, 2011)). Initially, they were classified into two classes based on their mode of mobilization and structures (Finnegan, 1989). Retrotransposons (Class I) transpose in a copy-and-paste manner via reverse-transcribed RNA intermediates and therefore amplify their copy number in the genome (Finnegan, 1989). The genomic location of DNA transposons (Class II) is altered through a cut-and-paste mechanism involving their terminal inverted repeats and transposase (Finnegan, 1989). With the identification of new transposons and their transposition mechanisms, new classification systems of transposons have been proposed (Jurka et al., 2005; Piégu et al., 2015; Wicker et al., 2007). In this study, we use the system described in Wicker et al. (Wicker et al., 2007), because it is compatible with the repeat database used (Repbase and RepeatMasker) and is frequently used for plant genome annotations. Similar to the original classification system (Finnegan, 1989), transposons are first divided into retrotransposons and DNA transposons based on the involvement of RNA intermediates for transposition, and are further classified in a hierarchical manner (Wicker et al., 2007). For example, retrotransposons are classified as long-terminal repeats (LTRs), long interspersed elements (LINEs), and short interspersed elements (SINEs) etc, based on their insertion mode, arrangement of proteins, and enzymology (Wicker et al., 2007). LTR retrotransposons are grouped into several superfamilies, including *Copia* and *Gypsy,* which are the most

* Corresponding author.
E-mail address: soyounwon@korea.kr (S.Y. Won).

https://doi.org/10.1016/j.scienta.2018.03.055
Received 15 November 2017; Received in revised form 19 March 2018; Accepted 26 March 2018
Available online 30 March 2018

abundant types of LTR retrotransposons in plant genomes, and are involved in shaping the genome architecture (Kumar and Bennetzen, 1999; Todorovska, 2007). DNA transposons are classified into two subclasses. Subclass I transposes in the canonical cut-and-paste manner and includes terminal inverted repeats (TIRs) and Cryptons, whereas subclass II transposons, are characterized by copy-and-paste transposition without DNA double-stranded cleavage and include Helitron and Maverick (Wicker et al., 2007).

Microsatellites (also known as simple sequence repeats, SSRs) are another overrepresented class of repeats characterized by tandem repeats of mono- to hexanucleotides (reviewed in (Ellegren, 2004; Vieira et al., 2016)). They are highly polymorphic with respect to the number of repeats within a unit, and the length of SSRs has therefore been used to develop molecular markers (Miah et al., 2013; Vieira et al., 2016). Specifically, nonrepetitive sequences flanking both ends of a microsatellite repeat can be conserved among closely related species or within a species and serve as a resource to design primers that amplify genomic regions encompassing polymorphic SSRs. Microsatellite markers have numerous benefits, including being hypervariable, abundant, neutral, co-dominant, and reproducible. Information on microsatellites and their flanking regions is generated by sequencing microsatellite-enriched genomic DNA or transcriptomes using next-generation sequencing (NGS) technologies.

Currently, we are conducting whole genome sequencing of *Chrysanthemum boreale*, a wild diploid chrysanthemum used for ornamental purposes and as a herb. Our preliminary genome sequencing and assembly results obtained using Illumina's Hiseq platform revealed many short contigs and scaffolds with average read lengths of as short as 8.8 kbp (unpublished data, SY Won). We assumed that the repetitiveness of the genome likely generated the highly fragmented genome assembly. To test this, we fractionated three levels of repetitive DNA (C_0t-1, C_0t-10, and C_0t-100) by using the reassociation kinetics of the denatured DNA in a C_0t analysis in which C_0t-1, C_0t-10, and C_0t-100 were considered as pools of highly repetitive (HR), highly + moderately repetitive (MR), and highly + moderately repetitive + low-copy (LC) DNAs, respectively (Chang et al., 2008; Peterson et al., 2002), and we determined their chromosomal location by FISH analysis (Cuyacot et al., 2016). Indeed, the *C. boreale* genome contained a large fraction of repetitive sequences (Cuyacot et al., 2016). However, the types and nucleotide sequences of these repeats were unknown.

In the present study, we used genomic, bioinformatic, and cytogenetic approaches to identify repeats in the *C. boreale* genome. Repetitive DNA sequences were identified by assembling low-coverage NGS reads and counting the number of reads mapped to each assembled contig. As the repetitiveness of a genomic region is proportional to the number of NGS reads mapped to it, the contigs with the deepest read depth were defined as repeat elements in *C. boreale*. These sequences were then annotated and their chromosomal distribution was determined by FISH. This report—the first to uncover overrepresented repeats in the *Chrysanthemum* genus— will help elucidate how the present genome architecture evolved within this genus.

2. Materials and methods

2.1. Plant material

C. boreale (accession number IT121002 in the RDA GeneBank, Republic of Korea) was collected from the Republic of Korea as previously described (Hwang et al., 2013). Shoot cuttings were transferred into pots and grown under natural daylight conditions in a greenhouse at the National Institute of Horticultural & Herbal Sciences, Rural Development Administration, Republic of Korea.

2.2. Whole genome sequencing

Young leaf tissue was harvested, frozen in liquid nitrogen, and stored at −80 °C until use. Genomic DNA was extracted as previously described (Kim et al., 2006) and its quantity and quality were assessed using an ND-1000 NanoDrop Spectrophotometer (NanoDrop Technologies, USA) and an Agilent 2100 Bioanalyzer (Agilent Technologies, USA). A paired-end library with an insert size of 300 bp was constructed using the Illumina Paired-End DNA Sample Prep Kit (Illumina, USA) according to the manufacturer's instructions and sequenced on an Illumina HiSeq1000 following standard protocols for 2 × 100 bp by the Genomics Division, National Institute of Agricultural Sciences, Rural Development Administration, Republic of Korea. Raw reads were submitted to the Sequence Read Archive (accession number: SRA618418) at the National Center for Biotechnology Information (NCBI).

2.3. Identification of repeat sequences

Raw reads of 1.3 Gbp were randomly extracted and used for the *de novo* assembly of low coverage whole-genome shotgun (WGS) sequences (dnaLCW) as previously described (Kim et al., 2015). Briefly, reads were preprocessed using the CLC-quality trim tool in CLC ASSEMBLY CELL package ver. 4.06 beta. 67189, (http://www.clcbio.com/products/clc-assembly-cell/) with Phred scores of 20 as a cutoff value. Preprocessed reads were assembled with the CLC *de novo* assembler using default parameters. After preprocessed reads were mapped to contigs using the CLC mapping tool, contigs with high read depth were extracted for further analysis. These contigs were subjected to blastn searches against the NCBI non-redundant (nr) database to remove sequences encoding chloroplast, mitochondrial, and ribosomal DNA. The 20 contigs with the highest read coverage were annotated with CENSOR, software that detects repetitive elements by WU-BLAST against Repbase, a reference database of repeats in eukaryotes (www.girinst.org/censor/index.php) (Kohany et al., 2006). They were further annotated using the RepeatMasker program in the RepeatExplorer pipeline (Smit et al., 2013-2015; Smit et al., 2013). Additionally, the presence of SSRs was examined using MISA (MIcroSAtellite identification tool) (Yuan et al., 2016).

2.4. Probe preparation

PCR primers were designed using Primer3 (Koressaar and Remm, 2007) and are listed in Supplemental File 1. PCR was conducted using a PCR DIG Probe Synthesis Kit (Roche, Germany) following the manufacturer's instructions. PCR products were electrophoresed in an agarose gel to examine PCR amplification and used as probes for FISH analysis.

2.5. FISH

FISH was performed as previously described with minor modifications (Hwang et al., 2013). To prepare chromosomes, root tips were pre-treated with 2 mM 8-hydroxyquinoline at 25 °C for 5 h, fixed in freshly prepared ethanol:acetic acid (3:1, v/v) at room temperature (RT) for 2 h and stored in 70% ethanol at −20 °C prior to use. After rinsing with distilled water, root tips were treated with an enzyme mixture (0.3% cellulose, 0.3% pectolyase, and 0.3% cytohelicase) at 37 °C for 1 h, squashed in a drop of 60% acetic acid on a clean slide, and air-dried. Chromosomes on slides were treated with 100 µg/mL RNaseA in 2X saline sodium citrate (SSC) at 37 °C for 1 h, washed with 2X SSC three times, post-fixed in 4% paraformaldehyde for 10 min, and dehydrated in an ethanol dilution series. The hybridization mixture was incubated at 90 °C for 10 min and immediately placed on ice for 5 min