Biochimica et Biophysica Acta 1859 (2016) 366–380

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

What makes up plant genomes: The vanishing line between transposable elements and genes

Dongyan Zhao, Ann A. Ferguson, Ning Jiang *

Department of Horticulture, Michigan State University, 1066 Bogue Street, East Lansing, MI 48824, USA

ARTICLE INFO

Article history:
Received 17 August 2015
Received in revised form 9 December 2015
Accepted 11 December 2015
Available online 17 December 2015

Keywords:
Transposon
Retrotransposon
Gene expression
Gene duplication
Domestication

ABSTRACT

The ultimate source of evolution is mutation. As the largest component in plant genomes, transposable elements (TEs) create numerous types of mutations that cannot be mimicked by other genetic mechanisms. When TEs insert into genomic sequences, they influence the expression of nearby genes as well as genes unlinked to the insertion. TEs can duplicate, mobilize, and recombine normal genes or gene fragments, with the potential to generate new genes or modify the structure of existing genes. TEs also donate their transposase coding regions for cellular functions in a process called TE domestication. Despite the host defense against TE activity, a subset of TEs survived and thrived through discreet selection of transposition activity, target site, element size, and the internal sequence. Finally, TEs have established strategies to reduce the efficacy of host defense system by increasing the cost of silencing TEs. This review discusses the recent progress in the area of plant TEs with a focus on the interaction between TEs and genes.

1. Introduction

Transposable elements (TEs) were first discovered by Barbara McClintock in the 1940s using maize (*Zea mays*) as a model system. Maize, a member of the grass (Poaceae) family, is an excellent model for genetic studies. Its large flowers with separate male (the tassels) and female (the ears) reproductive organs allow easy genetic crosses. In addition, the large size of maize chromosomes makes them readily visible under a light microscope. In a normal maize plant, chromosome breakage is rarely observed. Nevertheless, McClintock noticed very frequent breakage on a particular locus of chromosome 9 in one special maize line [1]. Subsequently, she discovered that two loci were required for the breakage. One locus found at the site of breakage and was called *Ds* (*Dissociation*). The other, which was required to "activate" the breakage, was therefore called *Ac* (*Activator*). Since the location of *Ac* and *Ds* appears to be variable between generations, McClintock proposed that they were actually genetic elements capable of transposition [1]. Thus, the first TEs were discovered through a forward genetics approach.

In 1983, more than 30 years later, McClintock won a Nobel Prize for her discovery of TEs. This coincided with the cloning and sequencing of the *Ac* and *Ds* elements by multiple research groups [2,3,4,5]. It was discovered that *Ac* encodes a transposase protein responsible for the transposition of itself as well as *Ds*. Meanwhile, more TEs were identified from maize and other organisms including other plants, animals, fungi, algae, and bacteria. These TEs form distinct superfamilies, families, and subfamilies (defined below)—*Ac*/*Ds* represent only one of numerous TE families. Moreover, not all TEs at present are capable of transposition; the ones with current transposition activity are called active TEs. When an active TE creates a new copy, the two copies are often identical at the time of transposition. Over time, the two copies diverge and become less similar.

In 2002, approximately 20 years after the cloning of *Ac*/*Ds*, the entire genome of rice (*Oryza sativa*), a relative of maize, was sequenced [6,7,8]. Rice was the first crop genome sequenced and the quality of its sequence remains the best among all crops. In 2003, the first active DNA element in rice, called *mPing*, was discovered independently by three research groups through two distinct approaches [9,10,11]. One group identified *mPing* through a mutation called *slender glume*, caused by the insertion of this element [11], resembling the classical pathway for TE discovery. A reverse genetics approach was employed by the two other groups [9,10]. With this method, candidate active elements were identified through computational search of available genomic sequences. The rationale for this search is that if an element is active, or was active in the recent past, identical or highly similar copies should be identified in the genome. Among the 52 copies of *mPing* in Nipponbare, the sequenced rice cultivar, 40 are identical to each other, strongly suggesting a recent activity. Since *mPing* appeared to be a good candidate for an active TE, its transposition activity was tested in tissue culture, where excisions and new insertions of *mPing* were observed [9,10]. The advances of technology, especially the availability of genomic and transcriptomic sequences, have revolutionized our approaches to study TEs and brought about a burst of information about TE biology. In this review, we will discuss some of the latest progress in plant TEs.

* Corresponding author.
E-mail address: jiangn@msu.edu (N. Jiang).

http://dx.doi.org/10.1016/j.bbagrm.2015.12.005
1874-9399/

2. Classification of plant transposable elements

Based on the transposition mechanism, TEs fall into two classes. Class I, the retrotransposons, use a "copy and paste" mechanism and utilize an RNA intermediate for transposition [12]. Class II, the DNA transposons, transpose via a DNA intermediate through a "cut and paste" mechanism [13]. Based on their coding capacity, both classes of TEs can be divided into autonomous and non-autonomous elements. Autonomous elements (such as *Ac*) encode the protein products required for their transposition. Non-autonomous elements (such as *Ds*) do not encode the relevant products and rely on their cognate autonomous elements for transposition. When an element inserts into a genomic locus, a small piece of flanking sequence in the insertion site is duplicated and this sequence is called target site duplication (TSD).

For both RNA and DNA TEs, they are often further categorized into superfamilies. A superfamily is classified by elements sharing transposases with significant similarity at the protein level and having similar length of TSDs. A family of elements typically shares a sequence that is critical for transposition, such as the element ends. A subfamily is composed of members wherein the entire element sequence or at least the major portion of the elements is conserved (Fig. 1). However, during genome-wide annotation, a family is defined by a practical method using sequence similarity among different elements due to the difficulty in defining elements sharing transposition machinery or *cis*-sequences. A common practice is that if two elements share 80% or higher identity at the nucleotide level for over 80% of the length of the shorter element, the two elements are considered to belong to the same family [14]. It is helpful to follow this community standard whenever possible, but certainly, these parameters may be modified to suit the objectives of each individual study.

2.1. Class I elements (retrotransposons)

Class I elements can be further divided into several groups, including the long terminal repeat (LTR) elements, long interspersed nuclear elements (LINEs), and short interspersed nuclear elements (SINEs) (Table 1). Due to the lack of LTRs, LINEs and SINEs are called non-LTR retrotransposons. Non-LTR retrotransposons are often associated with a poly-A tail at the 3′ end of the elements as a consequence of transcription. LTR elements are further classified into *Gypsy* superfamily and *Copia* superfamily, depending on the gene order in the internal region of the elements [12]. The internal regions of LTR elements encode *gag*, *pol*, and *int* genes that are synthesized as a polyprotein. The *gag* gene encodes structure proteins of virus-like particles which are responsible for packaging of retrotransposon RNA and proteins. The *pol* gene encodes reverse transcriptase and RNase H activities that are required for replication/transposition of the retrotransposon, and *int* encodes the integrase that allows the DNA form of the retrotransposon to insert at a new chromosomal location [15]. During the transposition of class I elements, the element mRNAs are converted into cDNA through the action of reverse transcriptase (from the *pol* gene) which is encoded within the element, and the TE cDNAs are then inserted into a target site in the genome. LTR elements are associated with a TSD of 5 bp whereas non-LTR retrotransposons generate TSDs with variable length and in some cases even deletion of the flanking sequence [16].

Because of their replicative transposition mechanism, class I elements can amplify very rapidly and constitute the largest portion of DNA content in most plant genomes. Particularly, LTR elements are often responsible for the expansion of plant genomes, which is different from that in mammals where non-LTR retrotransposons are very abundant [17]. For instance, 75% of the maize genome is occupied by LTR retrotransposons, including both *Gypsy* and *Copia* like elements [18]. In contrast, non-LTR retrotransposons are relatively scarce in plants, usually accounting for less than 5% of the genome with some exceptions, *e.g.*, apple (*Malus domestica*, 7.95%), sacred lotus (*Nelumbo nucifera*, 6.4%), sugar beet (*Beta vulgaris*, 5.67%), and banana (*Musa acuminata*, 5.41%) [19,20,21,22].

2.2. Class II elements (DNA transposons)

DNA transposons are usually associated with terminal inverted repeats (TIRs) and transpose via a DNA intermediate [13]. In general, DNA transposons excise from one site (donor site) and reinsert elsewhere (target site) in the genome. The departure of element from donor site is called "excision" whereas the integration of element into target site is called "insertion/reinsertion" or "forward transposition".

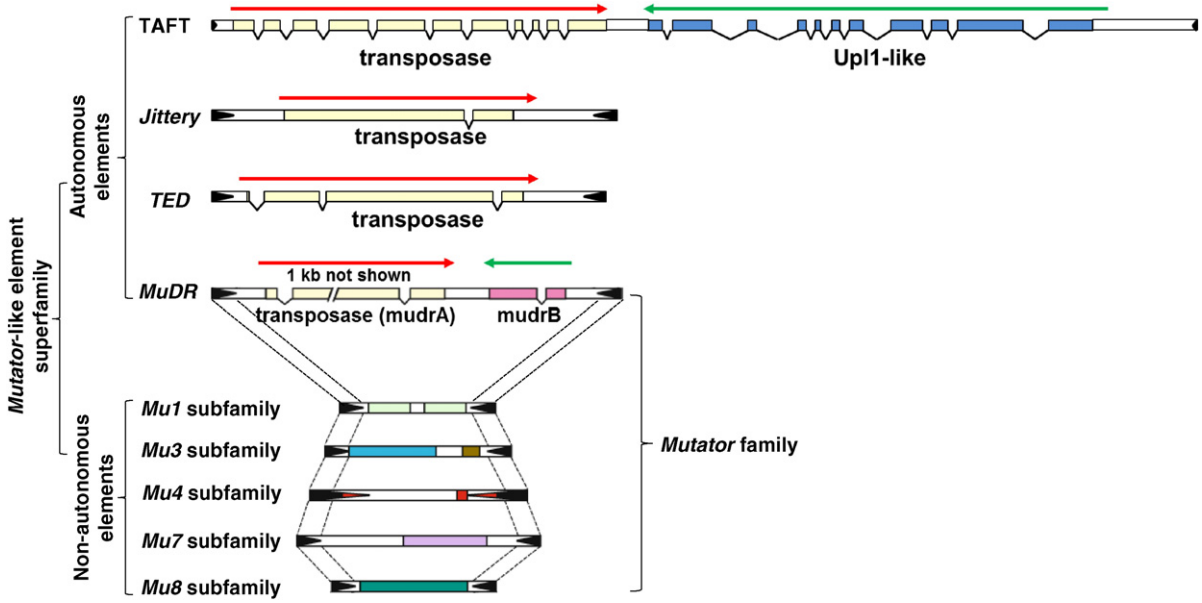

Fig. 1. Structural organization of a superfamily of transposable elements. *Mutator*-like superfamily (DNA transposons) in maize is used as an example. Terminal inverted repeats (TIRs) are depicted as black triangles. Colored boxes represent open reading frames in autonomous elements or acquired gene fragments in non-autonomous elements, and other sequences are depicted as white boxes. The homologous sequences between *MuDR* and its non-autonomous counterparts are connected with dashed lines. Introns are shown as lines connecting exons. In *Mu4*, the red triangles indicate that a portion of the acquired gene fragment was duplicated, inverted, and became part of the TIR. Colored arrows above the elements indicate the transcribed regions as well as the orientation of transcription.