Genomics 100 (2012) 222–230

Contents lists available at SciVerse ScienceDirect

Genomics

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

Characterization and functional annotation of nested transposable elements in eukaryotic genomes

Caihua Gao [a], Meili Xiao [b], Xiaodong Ren [a], Alice Hayward [c], Jiaming Yin [a], Likun Wu [a], Donghui Fu [b,*], Jiana Li [a,**]

[a] Engineering Research Center of South Upland Agriculture, Ministry of Education, College of Agronomy and Biotechnology, Southwest University, Chongqing, China
[b] Key Laboratory of Crop Physiology, Ecology and Genetic Breeding, Ministry of Education, Jiangxi Agricultural University, Nanchang 330045, China
[c] School of Agriculture and Food Sciences, Centre for Integrative Legume Research, The University of Queensland, St Lucia, 4072, Australia

ARTICLE INFO

Article history:
Received 22 November 2011
Accepted 3 July 2012
Available online 16 July 2012

Keywords:
Nested transposable element
Insertion bias
Functional annotation
Eukaryotic genome

ABSTRACT

The movement of transposable elements (TE) in eukaryotic genomes can often result in the occurrence of nested TEs (the insertion of TEs into pre-existing TEs). We performed a general TE assessment using available databases to detect nested TEs and analyze their characteristics and putative functions in eukaryote genomes. A total of 802 TEs were found to be inserted into 690 host TEs from a total number of 11,329 TEs. We reveal that repetitive sequences are associated with an increased occurrence of nested TEs and sequence biased of TE insertion. A high proportion of the genes which were associated with nested TEs are predicted to localize to organelles and participate in nucleic acid and protein binding. Many of these function in metabolic processes, and encode important enzymes for transposition and integration. Therefore, nested TEs in eukaryotic genomes may negatively influence genome expansion, and enrich the diversity of gene expression or regulation.

1. Introduction

Transposable elements (TE) are randomly mobilized repetitive DNA sequences that occupy a major part of nearly all eukaryotic genomes. Approximately 85% of the maize genome and almost half of the human genome is composed of TEs [1,2]. Movement and accumulation of TEs has significantly contributed to genome restructuring [3], resulting in profound changes in genome size [4], as well as gene function and expression [5]. In humans, mobilization of TEs has led to 25% of known disease-associated mutations [1,2].

TEs exhibit broad diversity both in their structure and transposition mechanisms. TEs are defined by their transposition mode, namely, retrotransposons and DNA transposons. Retrotransposons comprise an important part of plant genomes, and proliferate through a "copy-and-paste" mode by reverse transcription of RNA intermediates, using self-encoded or exogenous reverse transposase. Retrotransposons that have their own reverse transposase genes can be subdivided into two major classes: (1) LTR (Long Terminal Repeat), including the superfamilies *Copia* and *Gypsy*; and (2) non-LTR TEs, including the long interspersed repeat elements (LINE) and short interspersed repeat element (SINE) superfamilies [6]. DNA tranposons or terminal inverted repeat (TIR) tranposons proliferate through a "cut-and-paste" mode by DNA intermediates [7], and are classified into the following subfamilies: *En*/*Spm*, Ac/Ds, and hAT.

The process of species or genome hybridization results in the activation and amplification of replicated TEs, especially long terminal repeat (LTR) retrotransposons, whose proliferation can contribute to genome size expansion [8]. TEs can be inserted randomly into gene loci or intergenic regions as well as other repetitive sequences including pre-existing TEs. Depending on the site of insertion this can result in the creation of novel genes [9], new nested genes [10,11], or different types of TEs [12]. DNA transposons are preferentially associated with the euchromatic, or genic, component of genomes and show an insertion bias in the upstream sequences of genes [13]. This can immediately affect gene regulation and function and may have significant biological and evolutionary consequences in eukaryotic genomes [14]. The host genome can control the transposition and activity of TEs through epigenetic regulation and gene imprinting [15,16].

The insertion of TEs into pre-existing or other TEs produces nested TEs. Compared with non-nested TEs, nested TEs are relatively less abundant, but are present in many species. Organisms with high TE density exhibit a certain amount of nested TEs. Chimeric retrogenes are bipartite nested elements that originate from the fusion of two retrotransposons [17]. For instance, LINEs in mammalian and fungal genomes can become chimeric TEs through the fusion of DNA replicates of the cellular transcripts to each other or to the 3′ part of another LINE [17,18]. In rice, 42% of primary retrogenes generate functional chimeric genes [19]; in mammals, LINEs may prevent

* Corresponding author. Fax: +86 791 83813185.
** Correspondence to: J. Li, Engineering Research Center of South Upland Agriculture, Ministry of Education, China. Fax: +86 791 83813185.
E-mail addresses: fudhui@163.com (D. Fu), ljn1950@swu.edu.cn (J. Li).

0888-7543/$ – see front matter.
doi:10.1016/j.ygeno.2012.07.004

pre-existing genes from generating exon–intronic structures [20]. The Alu, hAT, ERV1, and MaLR type of TEs preferentially insert into host and young TEs nested within certain types of older ones as opposed to within non-TE intergenic regions [18]. These nested TEs are significant and informative in understanding genome evolution, gene organization, and the regulation of gene expression. However, few studies have focused on the biased distribution of nested TEs in different species, the characterization of nested TEs, the mechanism of nested TE occurrence, and the biological and evolutionary relevance functions of these chimeric TEs.

In the present study we evaluated available eukaryotic genome sequences for intact and nested TE sequences [21]. We detected the proportion of nested TEs, as well as the insertion bias of young nested TEs, and analyzed the relationship between nested TE and host TEs. Functional annotations were performed to gain better insight into the biological and evolutionary significance of nested elements. Overall, we aimed to investigate the distribution, nature, evolution and putative roles of nested elements in genomic regions.

2. Results

2.1. Establishment of the TE database and identification of the nested TEs

We obtained 16,137 intact, non-redundant TE sequences from almost all publicly-available 50,935 repetitive sequences (120 Mb) collected through RJPrimers [21]. These sequences were classified into 26 groups across 200 species according to their annotation using Repeatmasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). A total of 8995 TEs from unknown species were excluded from analysis, while 11,329 TEs with defined annotations were retained. These retained TE sequences were artificially screened by discarding TEs of the same type. A total of 690 sequences (6.1% of all definite 11,329 TEs) from 36 different species were predicted to be nested TEs, which were host to 802 different inserted TEs. In addition, 12.8% (88/690) of the nested TEs contained a total of 140 inverted DNA transposons.

2.2. Statistics of nested TEs in different species

A total of 36 species contained nested TEs from the TE database (Fig. 1). An average of 19 (690/36) nested TEs per species was detected across the Eukaryota, including an average of 4.3 nested TEs per species in Coelomata, and an average of 46.9 nested TEs per species in Embryophyta. Within this grouping, the nested TEs were mainly distributed in Gnathostomata (81 nested TEs) and Poaceae (511 nested TEs). In Gnathostomata, the majority of the nested TEs were from the species *Branchiostoma floridae* (16 nested TEs) and *Equus caballus* (15 nested TEs). In plants, ten times more nested TEs were identified in monocotyledons than within dicotyledons. In the monocot rice, 482 nested TEs were found, comprising 14.3% of the 3366 non-redundant TE sequences in the genome. This makes the rice genome the most highly populated with nested TEs, followed by *Zea mays* (with only 22 nested TEs). Interestingly, 440 *En/spm*-type DNA TEs from all species were preferentially inserted into DNA/hAT TEs, comprising 99.1% of the nested patterns in DNA/hAT. In addition, 63 LTR/*Gypsy*-type TEs showed bias toward insertion into *Ty1/copia*, contributing to 68.5% of all the nested patterns in rice (Table A.1).

2.3. Characterization of nested TE sequences

A total of 69 nested TEs, accounting for 10.0% of all 690 nested TEs, were predicted to contain 111 microsatellites with a range of 1 to 22 microsatellites per nested TE sequence. There were also 103 nested TEs (14.9%) containing a total of 415 tandem repeats (ranging from 1 to 37 tandem repeats per nested TE sequence), as well as 28 nested

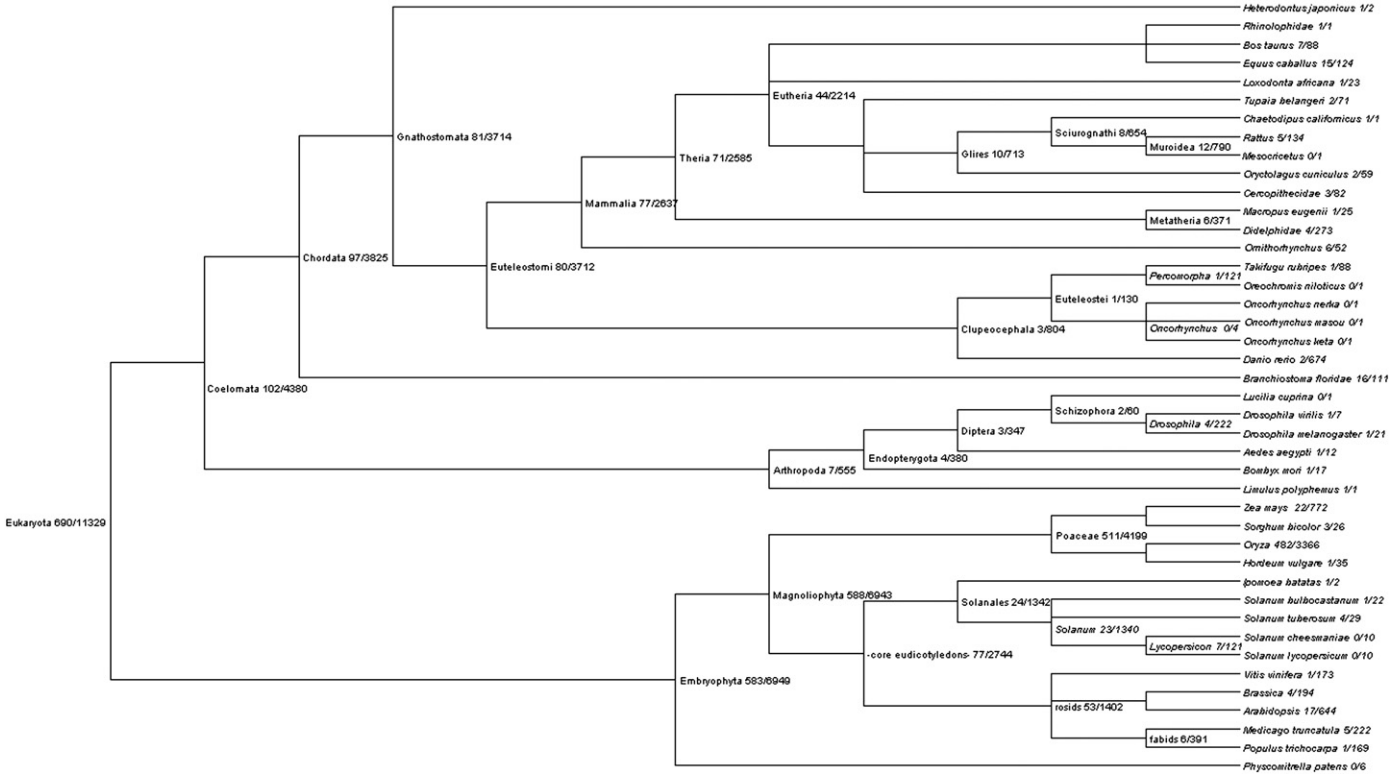

Fig. 1. Numbers of nested transposable elements in different taxonomic classes. Ratios of the number of nested TEs (numerator) to the total number of TEs per clade (denominator) are shown.