Gene 568 (2015) 89–99

Contents lists available at ScienceDirect

Gene

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

Research paper

The Ty1-*copia* LTR retroelement family PARTC is highly conserved in conifers over 200 MY of evolution

Andrea Zuccolo [a,b,*], Douglas G. Scofield [c], Emanuele De Paoli [d], Michele Morgante [b,d]

[a] Institute of Life Sciences, Scuola Superiore Sant'Anna, 56127 Pisa, Italy
[b] Istituto di Genomica Applicata, Via J. Linussio 51, 33100 Udine, Italy
[c] Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, SE-75236 Uppsala, Sweden
[d] Università degli Studi di Udine, Via delle Scienze 208, 33100 Udine, Italy

ARTICLE INFO

Article history:
Received 8 January 2015
Received in revised form 6 April 2015
Accepted 11 May 2015
Available online 14 May 2015

Keywords:
LTR retrotransposon
Ty1-*copia*
PARTC
Gymnosperms
Conifers
Picea
Pinus
Abies
Taxus
Juniperus
Ginkgo

ABSTRACT

Long Terminal Repeat retroelements (LTR-RTs) are a major component of many plant genomes. Although well studied and described in angiosperms, their features and dynamics are poorly understood in gymnosperms. Representative complete copies of a Ty1-*copia* element isolate in *Picea abies* and named PARTC were identified in six other conifer species (*Picea glauca*, *Pinus sylvestris*, *Pinus taeda*, *Abies sibirica*, *Taxus baccata* and *Juniperus communis*) covering more than 200 million years of evolution. Here we characterized the structure of this element, assessed its abundance across conifers, studied the modes and timing of its amplification, and evaluated the degree of conservation of its extant copies at nucleotide level over distant species. We demonstrated that the element is ancient, abundant, widespread and its paralogous copies are present in the genera *Picea*, *Pinus* and *Abies* as an LTR-RT family. The amplification leading to the extant copies of PARTC occurred over long evolutionary times spanning 10 s of MY and mostly took place after the speciation of the conifers analyzed. The level of conservation of PARTC is striking and may be explained by low substitution rates and limited removal mechanisms for LTR-RTs. These PARTC features and dynamics are representative of a more general scenario for LTR-RTs in gymnosperms quite different from that characterizing the vast majority of LTR-RT elements in angiosperms.

1. Introduction

Transposable elements (TEs) are major components of eukaryotic genomes (Wessler, 2006). They are particularly abundant in plants where their content ranges from at least 10% in *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), to 35% in rice (International Rice Genome Sequencing Project, 2005) and up to 85% in maize (Schnable et al., 2009). TEs are classified in two major classes according to the molecule acting as intermediate during their transposition process. Class I TEs (or retrotransposons) use RNA whereas Class 2 TEs (or DNA-TEs) use DNA (Finnegan, 1989). The most abundant subclass of TEs in plants is that of Long Terminal Repeat retrotransposons (LTR-RTs) (Feschotte et al., 2002). Two LTR-RT superfamilies are present in plants: Ty1-*copia* and Ty3-*gypsy*, classified according to the order of the gene coding for proteins involved in retrotransposition (Kumar and Bennetzen, 1999).

TEs were long dismissed as parasitic or "junk" DNA (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Because of their high copy number and peculiar mechanism of transposition, TEs have since been recognized as important players in shaping eukaryote genome structure and evolution. TEs can mediate rearrangements including insertions, deletions, inversions and duplications (Gray, 2000); capture, transfer and rearrange coding segments (Jiang et al., 2004; Morgante et al., 2005); activate or inactivate genes, contributing to transcriptional control (Kobayashi et al., 2004; Fernandez et al., 2010; Butelli et al., 2012); and can also provide sequences that are co-opted (or exapted) by host genomes (Hoen and Bureau, 2012).

LTR-RTs often accumulate by waves of sustained retrotranspositional activity that can take place over very short evolutionary times. In maize, LTR-RT amplification mostly occurred during the last 6 million years (SanMiguel et al., 1998), and the *Oryza australiensis* genome doubled its size in less than 3 million years because of sustained amplification of three families of LTR-RTs (Piegu et al., 2006).

Since TEs may have potentially dramatic effects on genome function, host genomes attempt to restrict their expression and proliferation

Abbreviations: LTR-RT, Long Terminal Repeat retroelements; nt, nucleotides; AA, amino acids; TE, transposable element; MY, million years; PBS, primer binding site; PPT, poly purinic tract; BAC, Bacterial Artificial Chromosome.

* Corresponding author at: Scuola Superiore Sant'Anna, Piazza Martiri della liberta' 33, 56127 Pisa, Italy.
E-mail address: a.zuccolo@sssup.it (A. Zuccolo).

http://dx.doi.org/10.1016/j.gene.2015.05.028
0378-1119/

through epigenetic modifications (Slotkin and Martienssen, 2007; Lisch and Slotkin, 2011; Lisch and Bennetzen, 2011). Epigenetic mechanisms may help prevent proliferation, but they do not lead to removal of TEs from host genomes. A long-debated question has been whether plant genomes have effective mechanisms to counteract the inflation of their genome sizes mediated by TE proliferation (Bennetzen and Kellogg, 1997). The analysis of an ever increasing amount of genome sequence data has provided some clues leading to the emergence of an increase/decrease model for plant genome sizes (Vitte and Panaud, 2005). The evidence gathered indicates that known mechanisms of DNA loss can effectively counterbalance genomic "obesity" by TE proliferation (Hawkins et al., 2009). The most important of such mechanisms are intra-strand homologous recombination and illegitimate recombination (Devos et al., 2002). The turnover of TEs in angiosperms' genomes is extremely fast: in rice the half-life of an LTR-RT insertion was estimated to be less than 6 million years, and during the last 8 million years at least 190 Mb of LTR-RT-related sequences have been removed from this genome (Ma et al., 2004).

So far most of the information related to plant TEs has been collected studying angiosperms, and the data available for TEs in gymnosperms is quite limited. The few studies of TEs and the repetitive DNA landscape in gymnosperms that are available are focused on a handful of species in the genera Taxus (Hao et al., 2011), Cupressus (Liu et al., 2011), Pinus (Kovach et al., 2010; Magbanua et al., 2011; Wegrzyn et al., 2013) and Picea (Hamberger et al., 2009; Morgante and De Paoli, 2011). In particular, characterizations of LTR-RTs in gymnosperms are lacking in comparison to angiosperm; just a few papers targeting conifers are available (Kamm et al., 1996; Elsik and Williams, 2000; L'Homme et al., 2000; Friesen et al., 2001; Stuart-Rogers and Flavell, 2001), and in most cases analyses are limited to small sequence data sets and/or partial short conserved sequences. Furthermore, only a few complete LTR-RTs have been identified and characterized in gymnosperms: IFG (Kossack and Kinlaw, 1999), PrRT1 (Rocheta et al., 2007) and Gymny (Morse et al., 2009). All three are Ty3-gypsy elements and were identified in Pinus species. Conifer genome sequencing projects (Picea abies, Nystedt et al., 2013; Picea glauca, Birol et al., 2013; Pinus taeda, Neale et al., 2014) have already provided a robust data framework to start filling this gap in knowledge. For example, analysis of the recently sequenced Norway spruce genome together with the genomes of several other gymnosperms suggested that the dynamics of TEs amplification and removal are strikingly different in angiosperms and gymnosperms. In particular, in gymnosperms LTR-RTs seem to have accumulated slowly over tens of million years rather than in bursts, and seem not to have undergone efficient removal (Nystedt et al., 2013).

Gymnosperms are organized into 85 genera which include more than 1000 species. The largest subclade, the conifers (Pinophyta), contains about 68 of these genera with a total of about 600 species (Christenhusz et al., 2011). The oldest gymnosperm fossil records date back to the Triassic period (Klavins et al., 2003; Pott and Krings, 2010). Extant gymnosperms and angiosperms are estimated to have separated around 300 MYA (Goremykin et al., 1997; Bowe et al., 2000).

In this work, we present the results of our structural characterization and phylogenetic analysis of a Ty1-*copia* element isolated in the conifer *Picea abies*. We show a striking degree of conservation of this element among six other conifer species (*Picea glauca*, *Pinus sylvestris*, *Pinus taeda*, *Abies sibirica*, *Taxus baccata* and *Juniperus communis*) covering together at least 200 MY of evolution (Crisp and Cook, 2011). The element is abundant, ancient and widespread across conifers. Its copies in *Picea*, *Pinus* and *Abies* are sufficiently conserved so as to be assignable to the same LTR-RT family, here named PARTC. Notably, the amplification of PARTC in conifers mostly took place after the different species arose from a common ancestor. These PARTC features and dynamics describe a scenario for LTR-RTs in gymnosperms quite different from that characterizing the large majority of LTR-RTs in angiosperms.

2. Materials and methods

2.1. Identification of PARTC elements

The 85-AA tract of Reverse Transcriptase of the complete *P. abies* PARTC element isolated in this work was used as a query in tBlastN (Altschul et al., 1997) searches to identify homologous copies in the available large genome sequences of *P. glauca* and *P. taeda* (Supplemental Table S5). The sequence coordinates of hits having at least 80% identities over the whole length of the query were used to extract large genomic tracts possibly containing the complete element. The tracts were self-compared in dot plot analysis carried out using the software Dotter (Sonnhammer and Durbin, 1995) and visually inspected to identify complete copies of the elements (Supplemental Fig. S1). The same strategy was adopted to identify PARTC-related elements in the complete genomes of the angiosperms *Amborella trichopoda*, *Vitis vinifera*, *Populus trichocarpa*, *Arabidopsis thaliana*, *Prunus persica*, *Citrus clementina* and *Oryza sativa* (Supplemental Table S2). To obtain complete PARTC sequences in conifer species for which only shallow-coverage whole-genome 454 reads were available (Table 1), the following strategy was adopted: (1) All 454 reads sharing similarity with complete PARTC elements identified in *Picea* spp. and *P. taeda* were identified using RepeatMasker (http://www.repeatmasker.org); (2) The reads so identified were collected and assembled using CAP3 (Huang and Madan, 1999) run under default settings; (3) The contigs obtained were then compared using dotplot analysis against complete PARTC elements identified in *Picea* spp. and *P. taeda*. Dotplot results were used to assist the manual assembly by allowing for the ordering and editing of contigs (Supplemental Fig. S3) to produce a complete representative element for study species lacking a complete genome sequence.

The element assemblies obtained from this strategy represent a consensus sequence for the whole PARTC population in that species and are likely to be chimeras. For *A. sibirica*, *J. communis* and *T. baccata* PARTC elements, only the 5′ LTRs could be reliably identified. To complete the element sequences, the 5′ LTR was used as the 3′ LTR, thus the complete identity of the LTR pairs in each of these PARTC sequences (see Supplemental Data) is artifactual.

2.2. Estimates of within- and between-species nucleotide distances

A 300 nt long region from the core domain of Integrase was extracted from positions 1482–1781 of the PARTC element isolated in *P. abies*. The tract was used as query in similarity searches against all available datasets. The search was carried out using RepeatMasker, and all elements having at least 70% similarity with at least 150 nt of the query were retrieved. A divergence cutoff of ≤30% was used to avoid collecting regions from LTR-RT families other than PARTC. The 707 paralogous sequences identified were aligned using the software MUSCLE (Edgar,

Table 1
Sequence resources used and locations of identified PARTC elements in conifers. The number of nonredundant 454 reads and their total length (Mbp). The location of a complete PARTC element is shown with a download location or GenBank accession number, together with sequence coordinates where required. Data, including 4 sequenced *P. abies* BACs and 1025 *Ginkgo biloba* genomic sequences, can be retrieved from: ftp://congenie.org/congenie/Nystedt_2013/Repeats/RawData/.

Species	454 reads (Mbp)	Complete PARTC element
Picea abies	95,685 (71.4)	ftp://congenie.org/congenie/, Scaffold MA_19102:93827-98972
Picea glauca	98,700 (51.3)	GenBank, ALWZ02, Scaffold 168007737:80234-85497
Pinus taeda	–	GenBank, AC241292.1:95194-100463
Pinus sylvestris	98,269 (57.6)	GenBank, KJ000234
Abies sibirica	99,003 (60.8)	GenBank, KJ000231
Taxus bacata	98,605 (61.2)	GenBank, KJ000232
Juniperus communis	98,181 (61.3)	GenBank, KJ000233