Gene 594 (2016) 151–159

Contents lists available at ScienceDirect

Gene

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

Research paper

Distribution patterns and impact of transposable elements in genes of green algae

Gisele S. Philippsen [a,b], Juliana S. Avaca-Crusca [a], Ana P.U. Araujo [a], Ricardo DeMarco [a,*]

[a] Departamento de Física e Ciência Interdisciplinar, Instituto de Física de São Carlos, Universidade de São Paulo, São Carlos, São Paulo, Brazil
[b] Universidade Federal do Paraná, Jandaia do Sul, Paraná, Brazil

ARTICLE INFO

Article history:
Received 30 March 2016
Received in revised form 1 September 2016
Accepted 6 September 2016
Available online 07 September 2016

Keywords:
Transposable element
Transposable element domestication
Gene evolution

ABSTRACT

Transposable elements (TEs) are DNA sequences able to transpose in the host genome, a remarkable feature that enables them to influence evolutive trajectories of species. An investigation about the TE distribution and TE impact in different gene regions of the green algae species *Chlamydomonas reinhardtii* and *Volvox carteri* was performed. Our results indicate that TEs are very scarce near introns boundaries, suggesting that insertions in this region are negatively selected. This contrasts with previous results showing enrichment of tandem repeats in introns boundaries and suggests that different evolutionary forces are acting in these different classes of repeats. Despite the relatively low abundance of TEs in the genome of green algae when compared to mammals, the proportion of poly(A) sites derived from TEs found in *C. reinhardtii* was similar to that described in human and mice. This fact, associated with the enrichment of TEs in gene 5′ and 3′ flanks of *C. reinhardtii*, opens up the possibility that TEs may have considerably contributed for gene regulatory sequences evolution in this species. Moreover, it was possible identify several instances of TE exonization for *C. reinhardtii*, with a particularly interesting case from a gene coding for Condensin II, a protein involved in the maintenance of chromosomal structure, where the addition of a transposomal PHD finger may contribute to binding specificity of this protein. Taken together, our results suggest that the low abundance of TEs in green algae genomes is correlated with a strict negative selection process, combined with the retention of copies that contribute positively with gene structures.

1. Introduction

Transposable elements (TEs) were discovered in the 1940s by Barbara McClintock, who was the first to propose that mobile DNA elements could regulate gene expression (Biémont, 2010; McClintock, 1984). Regarded skeptically by scientific community until 1969, when their recognition emerged from bacterial studies, TEs were placed to junk DNA status due their apparent lack of function (Janicki et al., 2011; Rebollo et al., 2012). This view was prevalent up to 1990s, when the advances in sequencing techniques revealed that the majority of eukaryotes harbor mobile elements into their genomes, suggesting their relevance in evolution (Biémont, 2010; Levin and Moran, 2011). Currently, albeit their selfish nature, TEs are considered as a source of genetic variability able to influence evolutionary trajectories (Biémont, 2010; Feschotte, 2008; Janicki et al., 2011; Krom et al., 2007; van de Lagemaat et al., 2003; Lopes et al., 2008; Nekrutenko and Li, 2001; Venancio et al., 2010).

TEs are commonly classified into two main groups based on their transposition mechanism: retrotransposons, which transposes by an intermediate RNA copy, and DNA transposon, which transposes by moving their DNA from one genomic location to another without RNA intermediate (Finnegan, 1992; Jurka et al., 2007). The preeminent feature of mobile elements in the evolutionary question is related with their transposition capability. In fact, as a result of a TE insertion, genes might have their structure or regulation modified, as well can be induced to silence or function disruption (Feschotte, 2008; Jurka et al., 2007; Kazazian, 2004; Levin and Moran, 2011; McDonald, 1995).

Despite the likely deleterious effect, a TE insertion into coding region might change or adapt the original function of the gene in an advantageous manner (Krom et al., 2007; Lopes et al., 2008; Nekrutenko and Li, 2001). There are two mechanisms through which TEs may be integrated into coding regions: inserting directly inside a coding region or inserting in the non-coding region and posteriorly being conducted to exaptation as a new exon, possibly because mobile elements contain potential splice sites (Nekrutenko and Li, 2001). The fact that approximately 4% of human genes display TE fragments in their coding sequences reinforces the relevance of these elements in the genomes

Abbreviations: DNA, deoxyribonucleic acid; HMM, hidden Markov model; LTR, long terminal repeat; PCR, polymerase chain reaction; PHD, plant homeodomain; poly(A), polyadenilation; RNA, ribonucleic acid; RT, reverse transcription; SMC, structural maintenance of chromosome; TE, transposable elements; TLS, translation start site; TSS, transcription start site; UTR, untranslated region.

[*] Corresponding author.
E-mail address: rdemarco@ifsc.usp.br (R. DeMarco).

http://dx.doi.org/10.1016/j.gene.2016.09.012
0378-1119/© 2016 Elsevier B.V. All rights reserved.

evolution (Nekrutenko and Li, 2001). It is already recognized that TEs are also able to introduce new regulatory elements by inserting upstream from a gene (Faulkner et al., 2009; Rebollo et al., 2012; Thornburg et al., 2006) and it is estimated that 24% of the human promoters harbor a TE derived sequence (Jordan et al., 2003).

The green algae species *C. reinhardtii* and *V. carteri* are considered important biological models to address relevant issues as asymmetric cell division, cell differentiation and multicellularity evolution (Hallmann, 2010; Merchant et al., 2007; Prochnik et al., 2010). Both species are model organisms of volvocine algae and represent the extremes of size and complexity within the clade. While the unicellular *C. reinhardtii* has a life cycle performed sequentially by the somatic stage (biflagellate cell state) followed by the reproductive stage (gonidia cell state), *V. carteri* is a multicellular organism with two specialized cell types with division of labor: approximately 2000 somatic cells located at the surface of a transparent glycoprotein-rich sphere and approximately 16 reproductive cells located below the surface of the sphere (Hallmann, 2010; Merchant et al., 2007; Miller and Kirk, 1999; Prochnik et al., 2010).

Several TEs were described for the *C. reinhardtti* and *V. carteri* species. The first TE identified for *C. reinhardtii* was the TOC1 (Day et al., 1988; Lefebvre and Silflow, 1999), a 5.7-kbp LTR element that was initially identified in the second intron of the OEE1 (oxygen-evolving enhancer 1) gene. A 12-kb DNA transposon, named Gulliver, was identified in the *C. reinhardtii* and was the first TE employed as transposon-tagging in this species (Ferris, 1989; Lefebvre and Silflow, 1999). The first complete *copia*-like TE described in a green algae was the element Osser (4875 bp), which was identified in the *V. carteri* species (Lindauer et al., 1993). The nonautonomous DNA transposon Jordan (1595-bp), identified in *V. carteri* (Miller et al., 1993), was also employed as transposon-tagging for the study of important genes, like the regA and glsA genes (Kirk et al., 1999; Miller and Kirk, 1999) associated to the genetic program of germ-soma differentiation in *V. carteri* (Hallmann, 2010). Currently, libraries of TE sequences derived from the *C. reinhardtii* and *V. carteri* organisms are available in the Repbase public database (Jurka et al., 2005), establishing a suitable dataset to perform analyses regarding TE issues in these green algae species.

In this context, the availability of the genomes and the annotation data of *C. reinhardtii* and *V. carteri* green algae species (Merchant et al., 2007; Prochnik et al., 2010) provided an appropriated scenario to further study the contribution of TEs to gene function. Despite the fact that a preliminary study of TEs representation in these genomes was performed (Merchant et al., 2007; Prochnik et al., 2010), their distribution relative to genic structures have not been addressed. The aim of this work is to investigate the TE distribution and contribution to the architecture of genes from these two related species, which are microalgae model organisms (Hallmann, 2010; Prochnik et al., 2010).

2. Materials and methods

2.1. TE annotation in the genomes

Annotation of TEs was performed in the publicly available *C. reinhardtii* genome (version v5.3) (Merchant et al., 2007), in the version 1 (v1) and version 2 (v2) of *V. carteri* genome (Prochnik et al., 2010) and in the *A. thaliana* genome assembly TAIR10 (Lamesch et al., 2012) (available at Phytozome (Goodstein et al., 2012) – http://phytozome.net) independently. *A. thaliana* was considered in some analyses only as a comparative outgroup species of green algae, doesn't consisting the focus of this work. The TE sequence libraries derived from these organisms were obtained from Repbase (Jurka et al., 2005) (http://www.girinst.org/Repbase/). In cases where retrotransposons LTRs were represented as a separated Repbase entry in relation to the internal portion of the element, we merged the two portions of element to reconstitute an integral copy. TE copies were mapped by a BLASTn (Korf et al., 2003) search (hits cutoff e-value $<10^{-10}$) of TE sequences against the genome. Hits that belong to the same TE family, are positioned at the same chromosome, in the same orientation, with a distance <100 bp and collinearity with respect to TE sequence were considered as a single insertion. This approach was adopted to avoid an incorrect overestimation in the copy number of TEs into the genomes. Overlapping insertions larger than 50 bp were removed by selecting only the copy with higher score. Copies were considered as full-length when displaying 99% identity and coverage in relation to the full-length TE sequence obtained from public databases.

2.2. Analysis of orthologous genes and their association with TEs

Gene annotation and peptide sequences data derived from *C. reinhardtii*, *V. carteri* (v1 and v2) and *A. thaliana* genomes were obtained from Phytozome. All analyses were carried out considering only a subset of those genes representing assigned pairs of orthologous genes from those genomes. Approaches focused on those orthologous genes minimize the risk that genes considered in the analysis may constitute gene prediction artifacts, since it is not expected that such artifacts are conserved between species. To establish pairs of orthologous genes, we employed the InParanoid software (version 4.1) (Sonnhammer and Östlund, 2015). This approach resulted in the establishment of 8170 ortholog pairs from *C. reinhardtii* and *V. carteri* v1 genomes and 8701 ortholog pairs from *C. reinhardtii* and *V. carteri* v2 genomes. The subsequent analyses carried out in this study for the *C. reinhardtii* species considered the set of orthologous genes respective to *V. carteri* v1. In the case of the outgroup species *A. thaliana*, genes orthologous to either *C. reinhardtii* or *V. carteri* v1 were considered to perform the subsequent analyses, leading to a set of 3678 genes. With the purpose of investigating the association of orthologous genes with TEs, we considered annotated insertions that occurred up to 2 kbp upstream of the translation start site (TLS), within the gene (defined as a sequence from the start codon to the stop codon) or up to 2 kbp downstream of the stop codon as been relative to gene. Insertions that partially overlap those regions were also considered. In this way, we could assign the mobile elements copies relative to a gene.

2.3. Random TE insertions

To simulate random TE insertions in the genome we used a methodology similar to that utilized in a previous work (Zhang et al., 2011). In a round of simulation, for each TE family we generated a number of random genomic *loci* equal to observed frequency in the genome to mimic insertions. Then, in each gene context of interest, it was possible to estimate the expected TE frequency (f_{exp}) by the average number of random copies, obtained in 20,000 simulation rounds.

2.4. Statistical significance

The relative frequencies distribution of random copy number, for each gene context considered in this study, could be approximated by a Normal distribution, as showed in the Supplementary Fig. 1. This random variable can be denoted by $X{\sim}N(f_{exp},\sigma^2)$, with f_{exp} and σ^2 derived from simulation, and can be transformed in the Standard Normal distribution $Z{\sim}N(0,1)$. The z-value statistic associated with the observed frequency (f_{obs}) can be assigned by

$$z = \frac{f_{obs}-f_{exp}}{\sigma}$$

From the z-value, it is possible to estimate the p-value. p-Values estimates considered are: $|z| > 3.0902$: p-value <0.001; $|z| > 2.3263$: p-value <0.01; $|z| > 1.6449$: p-value <0.05.