Gene xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Gene

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

Research paper

Evolutionary dynamics of triosephosphate isomerase gene intron location pattern in Metazoa: A new perspective on intron evolution in animals

Bing Chen, Jingru Shao, Huifu Zhuang, Jianfan Wen *

State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, China

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 13 November 2016
Accepted 14 November 2016
Available online xxxx

Keywords:
Intron evolution
Intron location pattern (ILP)
Exon-intron arrangement
Triosephosphate isomerase gene (*tim*)
Metazoan divergence

ABSTRACT

Intron evolution, including its dynamics in the evolutionary transitions and diversification of eukaryotes, remains elusive. Inadequate taxon sampling due to data shortage, unclear phylogenetic framework, and inappropriate outgroup application might be among the causes. Besides, the integrity of all the introns within a gene was often neglected previously. Taking advantage of the ancient conserved triosephosphate isomerase gene (*tim*), the relatively robust phylogeny of Metazoa, and choanoflagellates as outgroup, the evolutionary dynamics of *tim* intron location pattern (ILP) in Metazoa was investigated. From 133 representative species of ten phyla, 30 types of ILPs were identified. A most common one, which harbors the maximum six intron positions, is deduced to be the common ancestral *tim* ILP of Metazoa, which almost had formed in their protozoan ancestor and was surprisingly retained and passed down till to each ancestors of metazoan phyla. In the subsequent animal diversification, it underwent different evolutionary trajectories: within Deuterostomia, it was almost completely retained only with changes in a few species with relatively recently fast-evolving histories, while within the rapidly radiating Protostomia, besides few but remarkable retention, it usually displayed extensive intron losses and a few gains. Therefore, a common ancestral exon-intron arrangement pattern of an animal gene is definitely discovered; besides the 'intron-rich view' of early animal genes being confirmed, the novel insight that high exon-intron re-arrangements of genes seem to be associated with the relatively recently rapid evolution of lineages/species/genomes but have no correlation with the ancient major evolutionary transitions in animal evolution, is revealed.

1. Introduction

Spliceosomal introns (or introns for short), existing exclusively and ubiquitously in eukaryotic genomes and leading to the appearance of the split gene structure, might play crucial roles in the origin of eukaryotes and also in the subsequent evolutionary diversification of them (Koonin, 2006; Rogozin et al., 2012). The evolution of this gene structural component is still a subject of intense debate. Meanwhile, intron evolution along with the ramifying lines of life, remains elusive.

Four basic parameters are often used to characterize introns: number (or density), position (or location), size (or length), and sequence context. However, intron sequence and size are usually more variable even among closely related organisms, while intron number and position (or intron density and location) are relatively conserved even over great evolutionary time scales and seem to be altered following certain evolutionary rules (Rogozin et al., 2003; Irimia and Roy, 2014). Thus the latter ones are often investigated and compared in the study of intron evolution. In the early days, this sort of studies were performed mainly on single-gene scale or based on several genes of limited species (Hankeln et al., 1997; Cho et al., 2004), due to the shortage of orthologous gene data. The postgenomic era opens the opportunity to perform the investigation on whole-genome scale and among much wider range of organisms, and thus some more important progress has thus been made, e.g. the revelation of that introns most likely arose before the last eukaryotic common ancestor (LECA) (since they have been found ubiquitously in all fully sequenced eukaryotic genomes), and the LECA and even the common ancestors of each of the six eukaryotic supergroups were probably intron-rich (Rogozin et al., 2012; Irimia and Roy, 2014), and that intron loss had dominated the evolution of eukaryotic genes (Csuros et al., 2011). However, these genome-scale studies have still been limited by the shortage of genome

Abbreviations: bp, base-pair(s); cDNA, DNA complementary to RNA; HFP, high-frequency position; ILP, intron location pattern; LFP, low-frequency position; ORF, open reading frame; LECA, the last eukaryotic common ancestor; TIM, triosephosphate isomerase; *tim*, triosephosphate isomerase gene.

* Corresponding author at: State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 32th East Jiaochang Rd., Kunming 650223, Yunnan, China.

E-mail addresses: chenbing@mail.kiz.ac.cn (B. Chen), shaojr@mail.kiz.ac.cn (J. Shao), zhuanghuifu@mail.kib.ac.cn (H. Zhuang), wenjf@mail.kiz.ac.cn (J. Wen).

http://dx.doi.org/10.1016/j.gene.2016.11.027
0378-1119/

data from enough representative species within a certain lineage, and/or by the situation where though having genome data, the sampled species are usually distantly related in evolution – both can result in inadequate taxon sampling. It has been pointed out, however, adequate taxon sampling, especially including the deep-branching intron-rich species, is crucial for reconstructing intron evolution (Nikitin and Aleoshin, 2013). In addition, incompletely annotated genomes, inability to accurately predict genes and their introns, and inability to identify the orthologous genes (especially when there exist multiple-member gene families and alternative splicing), these factors each often make the whole-genome intron statistics or homologous intron comparison infeasible or less reliable. Actually the broad taxonomic and dense phylogenetic sampling strategy has seldom been achieved previously. The fact that quite a number of introns identified to be newly gained in earlier studies (Babenko et al., 2004; Coghlan and Wolfe, 2004), were later denied by adding samples of other closely related species (Roy and Penny, 2006) or by finding the failure of the outgroups early used (Roy and Penny, 2007), exactly illustrates this situation. Probably just for these reasons, different (and even contradictory) scenarios of intron evolution, from 'overwhelming domination of intron losses' (Roy and Gilbert, 2005) to 'dramatic excess of intron gains' (Qiu et al., 2004), have ever been proposed.

In addition, the integrity or totality of all the introns within a gene was often neglected in previous intron evolution studies due to the prevailing use of individual intron statistics approach. But there might exist a certain extent of correlation structurally and functionally among all the introns within a single gene, since they are usually arranged with a proper distance from each other (probably to avoid splicing errors), and are transcribed into a common primary transcript and then removed all together when RNA splicing; in particular, in those genes with alternative splicing, introns in different combinations are removed from the same pre-mRNA to produce different functional proteins. Therefore, besides the previous individual-intron investigation method, we think that viewing all the introns within a gene in their totality might be a new and reasonable perspective for the study of intron evolution. We call the location of all the individual intron positions within a gene 'intron location pattern (ILP)', which can not only reflect the exon-intron arrangement of a gene overall, but also be much more expediently identified and compared among organisms, and we try to investigate the evolutionary dynamics of ILP rather than individual introns.

There already exist a few investigations of exon-intron arrangement of a certain gene or gene family (e.g. alpha-amylase gene family) within a certain lineage (e.g. Bilateria and Basidiomycetes) (Da Lage et al., 2011, 2013). But, besides introns being considered individually as mentioned above, the gene used in these studies is relatively recently and horizontally transferred from bacteria, and therefore is somewhat incompetent to represent the general evolution of exon-intron arrangement of the majority of eukaryotic genes. In addition, the lineages they investigated are relatively special and small clades within eukaryotes.

Triosephosphate isomerase (TIM) is a ubiquitous, extremely ancient conserved energy metabolic enzyme. There usually exists a single copy TIM gene (*tim*) in animal genomes, and no alternative splicing has ever been observed in this gene. Thus, it is fairly easy to identify and occasionally needed but not difficult to distinguish its orthology from paralogy, and most importantly, the moderate size and high conservation of its coded protein sequences make it possible to precisely determine and unambiguously align its intron positions among diverse organisms. Largely for these reasons, this gene had once become an emblematic gene in the research into the origin of spliceosomal introns, especially used by the proponents of the "intron-early" theory (Straus and Gilbert, 1985; Tittiger et al., 1993), and later, this gene was also used as a favorite model for intron evolution study in some lineages (Kwiatowski et al., 1995; Logsdon et al., 1995), though only a handful of *tim* gene data were available at that time. Now the vast accumulated genome data make it feasible to study this gene in a much wider range of organisms and by much denser phylogenetic sampling.

To reconstruct the evolutionary history of introns, relatively clear phylogenetic framework of the investigated lineage (and at best also clear between the investigated lineages and outgroup relatives) is important. Up to now, the animal inter- and intra-lineage phylogenetic relationships have become relatively clear and increasingly robust (Edgecombe et al., 2011). Moreover, that choanoflagellates are the closest unicellular relatives of Metazoa (King et al., 2008), is widely accepted now, and thus they can be served as effective outgroup for the evolutionary analysis of Metazoa. The importance of the proper outgroup in elucidating intron evolutionary history has been emphasized particularly by Roy and Penny (2007).

Here, taking advantage of the explosively increasing genome data, 146 *tim* genes, together with their ILPs, were obtained from 133 metazoan genomes in ten phyla, and as many as 30 types of distinct *tim* ILPs were identified and compared. Interestingly, some novel observations on the evolutionary dynamics of *tim* ILP and new insights into intron evolution in animals were obtained.

2. Materials and methods

2.1. Sequence retrieving and database searching

Metazoan *tim* gene sequence data were firstly retrieved from public database such as GenBank and Ensembl. Then, taking advantage of the vast accumulated genome data, lots of *tim* genes, especially from an increasing number of metazoan taxa, are available. Genomic sequences released from ongoing metazoan genome projects were downloaded from public sequencing centers to local computer. Both vertebrate (human) and invertebrates (*Drosophila* and *Caenorhabditis*) TIM sequences were used as queries to carry out local BLAST searches against these genomic sequences (e-value $< e^{-10}$) according to Altschul et al. (1997). The homologous TIM protein sequences were obtained and their corresponding genomic sequences were retrieved for further identification. In addition, the same manipulation was performed on protozoa choanoflagellates, *Salpingoeca* sp. and *Monosiga brevicollis*.

2.2. Identification of tim genes and their exon-intron structures

From the public databases, the exactly known exon-intron structures were directly collected into our dataset. As for the *tim* gene candidates that lack the EST data or come from genomic sequences for further confirmation, their introns within the coding region (namely open reading frame, ORF) and corresponding exon-intron structures were determined manually according to the following criteria: i) the deduced protein sequence should contain ~250 amino acids and include all known catalytic residues (e.g., Asn12, Lys14, His96 and Glu166. The numbering of amino acids is according to human TIM) and conserved motifs (e.g., extremely conserved 169-WAIGTG-174 and 209-YGGS-212, and a little less conserved 181-QAQEV-185), and those candidates were discarded when only partial coding sequences could be determined; ii) the sequence identity with known metazoan counterparts should be >50% and notable residue deletion or insertion (>3 residues) in the following alignment is not allowed – this is an indispensable qualification; iii) each exon sequence should be longer than 30 base-pairs; and iv) the typical consensus sequence 'GT⋯AG' and occasionally other sequences such as 'GC⋯AG' are used to assist the identification of introns.

2.3. Multiple sequence alignment and ILP definition

Multiple sequence alignments of both coding sequences and protein sequences were constructed using the online program ClustalW2 (EMBL), and the DNA sequence alignment results were modified according to the conservative property of protein sequence. For those closely related insect species from a common genus, only one TIM from one species was selected as representative. Then, all the identified

Please cite this article as: Chen, B., et al., Evolutionary dynamics of triosephosphate isomerase gene intron location pattern in Metazoa: A new perspective on intron evolution ..., Gene (2016), http://dx.doi.org/10.1016/j.gene.2016.11.027