Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

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

The role of introns in the conservation of the metabolic genes of *Arabidopsis thaliana*

Dola Mukherjee[a], Deeya Saha[a], Debarun Acharya[a], Ashutosh Mukherjee[b], Sandip Chakraborty[a], Tapash Chandra Ghosh[a,*]

[a] Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata, 700 054, West Bengal, India
[b] Department of Botany, Vivekananda College, 269, Diamond Harbour Road, Thakurpukur, Kolkata, 700063, West Bengal, India

ARTICLE INFO

Keywords:
Conservation patterns
Metabolic genes
Intron enrichment
Protein versatility
Mutational load

ABSTRACT

In *Arabidopsis thaliana*, primary metabolic genes (PMGs) are more evolutionarily conserved and intron-rich than secondary metabolic genes. We observed that PMGs are more primitive and pan-taxonomically persistent as compared to secondary (SMGs) and non-metabolic genes (NMGs). This difference in primitiveness and persistence is primarily correlated with intron number and is independent of gene expression level. We propose a twofold explanation behind higher intron enrichment in PMGs. Firstly, introns might increase protein versatility amongst PMGs through alternative splicing, providing selective advantage of PMGs and making them more persistent across diverse plant taxa. Also, multifunctional PMGs may acquire functional domains by increasing the intronic burden. Additionally, single nucleotide polymorphisms (SNPs) accumulate at a higher rate in introns as compared to exons. Moreover, a strong negative correlation between cumulative exonic SNPs density and intron number indicates that introns may protect the exonic regions against the deleterious effect of these mutations, making them more conserved.

1. Introduction

Introns are non-coding sequences that interrupt the coding regions of eukaryotic genes. They act as a hallmark of eukaryotic protein coding genes [1–3] and are important components of genome adaptation [4]. Although, spliceosomal introns are common amongst eukaryotic genomes, their density varies greatly across genomes as well as genes within the same genome [5] and deciphering the uneven phylogenetic distribution of introns is a major challenge for evolutionary genomics [4]. Understanding the function and evolution of introns have gained much attention since its discovery in the late 1970s [6]. The rapidly accumulating fully sequenced eukaryotic genomes are also allowing high-resolution reconstruction of evolutionary history of introns [7].

However, introns are thought to impose a considerable burden to the host [7], and there could be at least three possible deleterious effects on gene expression [4]: First, spliceosomal introns requires a spliceosome [7] and thus, splicing multiple introns is biologically expensive [8,9]. Second, intron transcription is costly in terms of time and energy [10–12]. Third, malfunction of any of the snRNPs will have a general detrimental effect on the cell [7]. Moreover, some studies showed that highly expressed genes are compact, especially, concerning intron size [13,14]. Finally, the mutational hazard hypothesis says that non-coding sequences have slightly deleterious effects on fitness because of the hazard of accumulating deleterious mutations [15–17]. Thus, to minimize the mutational hazard, selection would preferentially remove the excess DNA from genomes [5].

On the other hand, some recent studies highlight various advantages of having introns [7,18]. It has been reported that introns increase the protein diversity by exon shuffling and alternative splicing [5,19,20]. Some introns also regulate gene expression [5]. Moreover, introns play a pivotal role in mRNA export, transcription coupling, splicing, etc. [21] and also give rise to non-coding RNAs that participate in regulatory processes [22]. Introns can also boost the gene expression, and this positive effect is called intron-mediated enhancement (IME) [23].

Indeed, the relationships between gene expression and intron numbers have been a matter of debate. For example, Vinogradov showed that in humans, housekeeping genes and tissue specific genes differed in their genomic complexities and regulation [24]. While the former category harbored compact, broadly and highly expressed genes, the later was tissue specific. Such observations on the properties of housekeeping genes were assessed using an older dataset. However, a different trend was observed in the model plant *Arabidopsis thaliana*, where primary metabolic genes, being mostly housekeeping in nature exhibited not only elevated expression but also higher intron number

* Corresponding author.
E-mail address: tapash@jcbose.ac.in (T.C. Ghosh).

https://doi.org/10.1016/j.ygeno.2017.12.003
Received 26 September 2017; Received in revised form 6 December 2017; Accepted 8 December 2017

[25].

Some recent studies have indicated that the relation between gene expression and introns are much more complex than previously thought. While in animals like *Caenorhabditis elegans* and *Homo sapiens*, highly expressed genes contain less and compact introns [14], in plants like *Oryza sativa* and *Arabidopsis thaliana*, it was found that highly expressed genes contained more and longer introns than genes expressed at a low level [26]. However, when the intron length between model plant and animal were compared, the introns were found to be relatively shorter in the model plant *Arabidopsis thaliana* than the mammalian mouse model, indicating the cost of transcription is negligible [4], which may favor intron retention. Previous studies indicated that variation of intron size is influenced by various factors [27]. The metabolic requirements and spatiotemporal economy might also act as selective forces to resume surplus DNA [27]. For example, housekeeping genes that are required to express at a certain level in every cell comprise shorter introns than other genes in humans [28]. On the contrary, Gorlova et al. [20] showed that evolutionary conserved and primitive genes are more functionally important and have a more intron enrichment in human, which opens up the opportunity for novel functions. Genes expressed in pollens of *A. thaliana* have smaller introns than genes expressed in sporophytes [29]. However, it is unclear to what extent the genomic configuration of plant has been shaped by functional requirement and natural selection [13].

It was earlier reported that in *Arabidopsis thaliana*, primary metabolic pathway genes contain significantly more introns than secondary metabolic pathway genes [25]. Additionally, the primary metabolic pathway genes are evolutionary more conserved than secondary metabolic pathway genes on the basis of the ratio of synonymous and non-synonymous substitution rates (d_N/d_S). However, no correlation has been found between d_N/d_S and intron number. This may not be surprising as d_N/d_S, by definition, addresses the evolutionary rate of the coding regions. Thus, the difference of intron number of these two categories of genes in *A. thaliana* is still enigmatic. So, to address this issue, we have taken a different approach here. Encouraged by the work of Gorlova et al. [20], we have introduced, in this study, two new indices named *Persistence Index* (PI) and *Age Index* (AI) to see whether this intron number variation is correlated with the evolutionary history as well as the taxonomic distribution of the concerned gene within the plant kingdom. We have taken this approach as in plants, primary metabolic pathways are almost omnipresent while secondary metabolic pathways are restricted to specific plant groups [30]. Moreover, a gene's level of evolutionary conservation reflects its functional significance [31,32].

Thus, the objective of our study is to find whether higher intron enrichment of primary metabolic pathway genes (PMGs) over secondary metabolic pathway genes (SMGs) confer any selective advantages to them which can answer the primitiveness and pan-taxonomic distribution of PMGs. For analysis of PI and AI, we have selected six other plant species along with *A. thaliana* whole genome sequences are available. These include one dicot and two monocot species, one species each from pteridophyta, bryophyta and algae. Our analysis showed that in *A. thaliana*, these two indices differ in PMGs, SMGs and NMGs (Non Metabolic pathway Genes) and both PI and AI are significantly correlated with intron number. Moreover, introns accumulate more single nucleotide polymorphisms in PMGs than SMGs as well as NMGs and may act as buffer to protect the coding region of the genes to accumulate mutations. Our study shows that introns confer some advantages for evolutionary conservation of primary metabolic pathway genes in *A. thaliana*.

2. Materials and methods

2.1. Dataset preparation

We collected the whole genome data of *Arabidopsis thaliana* from Biomart interface [33] of Ensembl Plants [34] (http://plants.ensembl.org/). The metabolic gene dataset was prepared from KEGG Database (http://www.genome.jp/kegg) [35]. Initially, we obtained a dataset of 2512 metabolic genes out of which 2030 were PMGs and 482 were SMGs. We filtered out 209 metabolic genes from our dataset which participated in both primary and secondary metabolism. Finally, we had 1821 PMGs and 273 SMGs. The rest of the protein coding genes that did not participated in metabolism were categorised as non-metabolic genes or NMGs. We compiled a dataset of 24,903 NMGs. The complete gene list of PMG, NMG and SMG are provided in the Supplementary file 1.

2.2. Estimation of conservation of genes

We used the pan-taxonomic distribution of metabolic genes as a measure of conservation of the metabolic genes of *A. thaliana* rather than the protein level conservation. Previously, Gorlova et al. have formulated the conservation index as a measure of genes' degree of preservation [20]. The concept of Conservation index as perceived by Gorlova et al. [20] was further redefined by us as persistence index (PI) and age index (AI) to study the pan-taxonomic distribution of *A. thaliana* genes amongst the various plant taxa. PI reflects the distribution of orthologous genes of *A. thaliana* between the other plant taxa while AI denotes the primitiveness of the orthologous genes. We have detected orthologous set of genes in six of the below mentioned plant species: *A. lyrata* (dicot), *Sorghum bicolor* (monocot), *Oryza sativa* var. *japonica* (monocot), *Selaginella moellendorffii* (lycophyte), *Physcomitrella patens* (moss) and *Chlamydomonas reinhardtii* (alga) [36]. These species were ranked on the basis of their evolutionary distance from *A. thaliana*. We assigned rank 0 to those genes which are unique to *A. thaliana* while rank 6 was assigned to those genes which have orthologs on *C. reinhardtii*. Persistence index (PI) $= \Sigma x_i / (N - 1)$, where x_i, represents the count of orthologous gene across the selected plant taxa and N is the total of plant species selected apart from *A. thaliana*. Age index (AI) $= x_i / (N - 1)$, here x_i represents the rank where the primitive most ortholog of *A. thaliana* genes could be traced. The indices value ranges from 0 to 1. '0' depicting the genes confined only to *A. thaliana* and recent origin while '1' representing the most persistent and orthologs that could be traced to all other groups and hence more primitive. To explain the indices better, we put a hypothetical example where a gene of *A. thaliana* is present in 3 other groups so its PI is 0.5, now if the most primitive ortholog could be traced to Chlamydomonas, and then the AI is 1. We also checked the primitiveness of the *A. thaliana* genes by using Phylostratography (https://lighthouse.ucsf.edu/proteinhistorian/). Here, we have categorised the *A. thaliana* genes according to their phylogenetic origin into three groups-Arabidopsis (recent), Magnoliophyta (medium) and Viridiplantae (ancient).

2.3. Gene expression

Microarray expression data for *A. thaliana* was obtained from PLEXdb (www.plexdb.org/) [37]. The accession number of expression dataset is AT40 and the microarray platform used was ATH1-121501.

2.4. Intron enrichment

Both intron counts within each gene along with the intron length considered separately for studying the intron enrichment of the respective genes. The intronic coordinates were obtained from Biomart of Ensembl Plants (http://plants.ensembl.org/biomart).

2.5. Other genomic parameters

Intron count, intron length, transcript length, GO terms accessions and Pfam accessions were downloaded from Biomart of Ensembl Plant (http://plants.ensembl.org/biomart).