Genomics 110 (2018) 310–317

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
Available online 13 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).

Multifunctionality was calculated by summing up the number of GO biological process terms assigned to each gene identifier [38]. Domain number was obtained by summing up the Pfam [39] accession against each gene identifier.

2.6. Single nucleotide polymorphism (SNPs)

Data for Single Nucleotide polymorphism (SNPs) of the genome of *A. thaliana* was also obtained from Biomart of Ensembl Plants. The coordinates of the SNPs were mapped to both the exonic and intronic positions of genes of *A. thaliana.* The mapping of coordinates was done by using in-house PERL script.

2.7. Statistical tests

Statistical analyses were performed using SPSS v.13. Mann-Whitney *U* test [40] was used to compare the average values of different variables between two classes of genes since the values were not normally distributed in our dataset. For correlation analysis, we performed the Spearman's rank correlation coefficient ρ [41], where the significant correlations were denoted by $P < 0.05$. *Z*-test was also carried out to study the proportion difference between groups.

3. Results and discussions

3.1. PMGs are more intron rich than SMGs and NMGs in A. thaliana as well as in other plant groups

A previous study showed that PMGs are more intron-rich than SMGs in *A. thaliana* [25]. Here, we have also studied the non-metabolic genes of *A. thaliana* to get a complete picture of the intronic distribution in *A. thaliana* regarding metabolic and non-metabolic genes. We have considered a total of 2094 genes as metabolic genes and 24903 genes as non-metabolic genes (NMGs). Of these metabolic genes, 1821 genes are associated with primary metabolism while 273 genes are related with secondary metabolism. It was observed that PMGs on an average have higher intron number as compared to NMGs and SMGs (Fig. 1) (Mann-Whitney *U* test, $P < 10^{-6}$).

We then, analysed whether this trend (PMGs have higher intron number than SMGs and NMGs) is present in other groups of plants too. We have studied the differences between the average intron number in NMGs, PMGs and SMGs in all the seven species (Fig. 2). For this, we have considered the PMGs, SMGs and NMGs of *A. thaliana* and their orthologous genes from the other six species. It was found that in all the investigated species, PMGs always showed more introns than SMGs and NMGs. However, in *C. reinhardtii* (an aquatic alga), there is no significant differences between the three groups while in *P. patens*, *S. moellendorffii* and in the two monocot species, significant difference was found between NMGs and PMGs. However, in the two species of *Arabidopsis*, significant differences between PMGs and NMGs as well as between PMGs and SMGs have been found with respect to intron number. From these results, it can be concluded that PMGs gathered significantly more introns than NMGs or SMGs over time. It was also clear that early land plants showed a similar pattern before the monocot-dicot divergence. After that, these two groups showed significant differences with respect to intron number in PMGs, SMGs and NMGs.

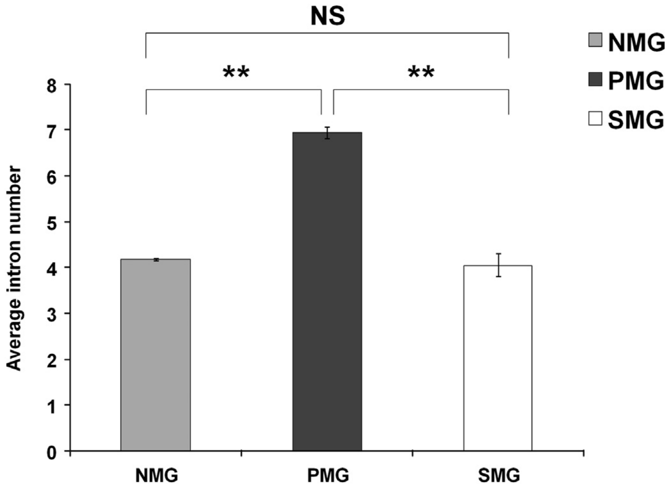

Fig. 1. Bar diagram showing the difference of intron number amongst different groups, PMGs, NMGs and SMGs. ** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

3.2. PMGs are more primitive and conserved than SMGs and NMGs in A. thaliana

We have estimated taxonomic distribution of PMGs, SMGs and NMGs using two unique indexes, i) Persistence index and ii) Age index. It was observed that in *Arabidopsis*, protein coding genes showed a marked variation in their degree of persistence across different plant species. The persistence index as well as age index of a given gene ranges from 0 (present only in *Arabidopsis* and of most recent origin) to 1 (present in all the investigated species and genes with most ancient origin). It was observed that primary metabolic genes (PMGs) show higher level of primitiveness and persistence as compared to NMGs and secondary metabolic genes (SMGs) (Fig. 3 A and B). Although, PMGs possessed significantly higher PI as well AI values compared to NMGs and SMGs ($P < 0.01$), SMGs did not show any significant difference of PIs and AIs as compared to NMGs ($P > 0.05$) (Fig. 3). It was observed that there was a significantly strong positive correlation (Spearman's $\rho = 0.993$, $P = 10^{-6}$, $N = 26997$) between PI and AI indicating that genes with most ancient origin are the ones that are more persistent across wide range of plant genomes. We also checked the phyletic age of the metabolic genes using Phylostratography (https://lighthouse.ucsf.edu/proteinhistorian/). It was observed that majority of the genes of PMGs have ancestral origin than SMGs and NMGs. However, the proportion of PMGs decrease gradually with the gene age. We also observed that the NMGs are mostly quite recent in their origin (Fig. 4). As PMGs in *A. thaliana* are more ancient in terms of their origin, as showed more PI, AI and ancient phyletic origin than the other two groups and they also retained more introns over time, there must be some selective advantage of retaining more and more introns in PMGs. Therefore, from here onwards, we would investigate the role of intron number in guiding the persistence of *Arabidopsis* genes.

3.3. Intron number is correlated with persistence index in A. thaliana

We observed a strong positive association between persistence index and intron enrichment (Spearman's $\rho_{\text{PI-intron number}} = 0.297$, $P < 10^{-6}$, $N = 26997$, Spearman's $\rho_{\text{PI-intron length}} = 0.225$, $P < 10^{-6}$, $N = 26998$). In animal genomes, previous studies suggested that persistence of genes or its conservation is highly correlated to its intron enrichment [20,42]. In agreement with these studies, our study also found a strong association between intron enrichment and gene persistence index. In addition to it, our study also showed that genes that are older and has wider pan-taxonomic distribution, have higher intron enrichment as compared with genes of more recent origin. Next, we intended to find out whether total intron lengths or intron number was the more prominent predictor of persistence. Intron number per gene was found to correlate highly with total intron length (Spearman's $\rho = 0.809$, $P < 10^{-6}$, $N = 26997$). So, we performed a partial correlation between PI and intron number after controlling for total intron length (Spearman's $\rho = 0.113$, $P < 10^{-6}$) and observed that there was a significant impact of intron number over PI. On the contrary, when intron number was controlled and correlation between

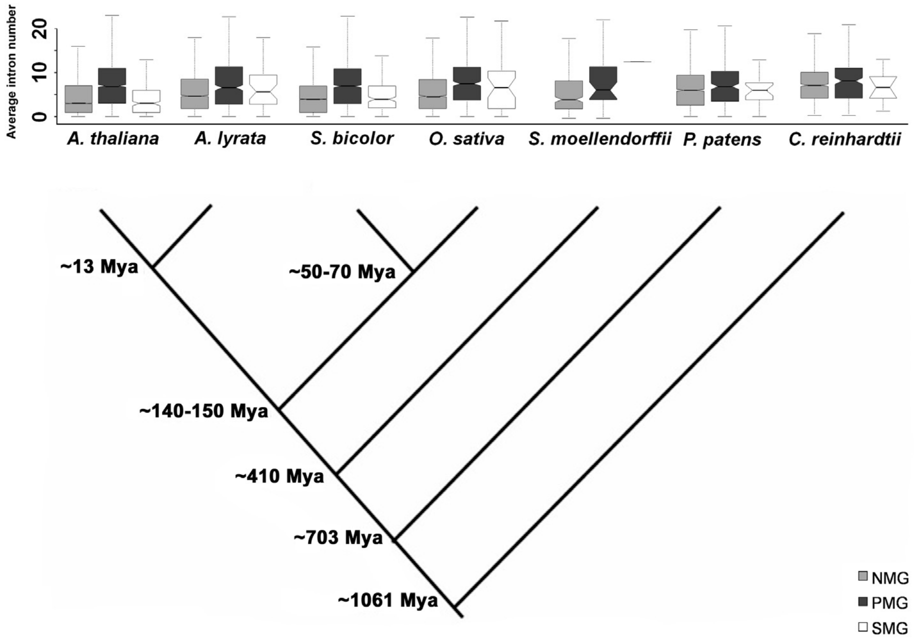

Fig. 2. Comparison of intron number between PMGs, SMGs and NMGs across diverse taxonomic plant groups.** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

total intron length and conservation was noticed, it was observed that there was a very weak significant correlation between them (Spearman's $\rho = 0.044$, $P < 10^{-6}$). Thus, the effect of intron length was negligible over conservation.

3.4. *Difference in conservation of PMGs, NMGs and SMGs in A. thaliana is independent of gene expression levels*

Gene expression level has been shown to be a major determinant of protein level conservation in plants and animals [43]. Henceforth, we were interested to study the effect of intron number over gene expression level of *A. thaliana*. It has been previously proposed that intron number negatively influences gene expression level in animals [20,42]. However, we obtained a strong positive correlation between intron number and gene expression level (Spearman's $\rho = 0.253$ $P < 10^{-6}$, $N = 21049$). Our results are in agreement with previous work [26] which also showed that highly expressed genes in plants contain more introns. Hence we were interested to study the effect of expression over conservation of PMGs. It was observed that PMGs have significantly higher expression level as compared to NMGs and SMGs. It was also observed that gene expression level positively correlates with PI (Spearman's $\rho_{PI} = 0.312$ $P < 10^{-6}$). Next, we intended to explore whether the difference of PI between PMGs, NMGs and SMGs were due to their difference in the expression level. In this context, we binned gene expression values into four bins-Bin1 (containing genes having gene expression value 2.00–5.00), Bin2 (gene expression value 5.00–8.00), Bin3 (gene expression value 8.00–11.00) and Bin4 (gene expression value > 11.00). Bin4 also showed absence of any SMGs. It was observed that in each bin, PMGs has significantly higher persistence as compared to NMGs and SMGs except Bin3 where expression level of PMGs and SMGs were insignificant (Fig. 5A).This indicates that difference of PI between PMGs and NMGs as well as PMGs and SMGs

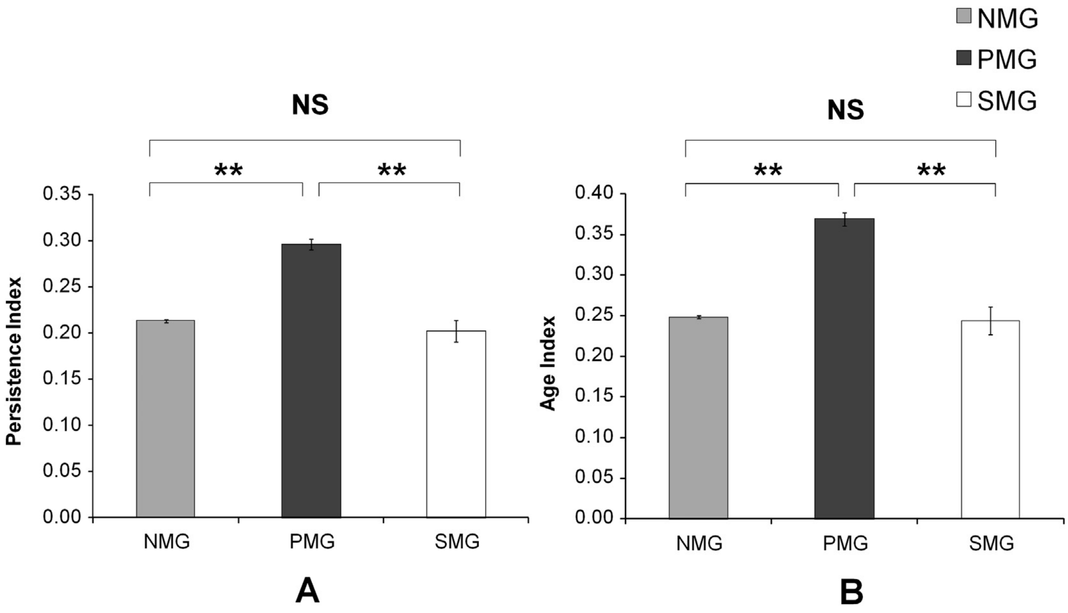

Fig. 3. Bar diagram showing the significant difference of (A) Persistence index and (B) Age Index, amongst different groups, PMGs, NMGs and SMGs.** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

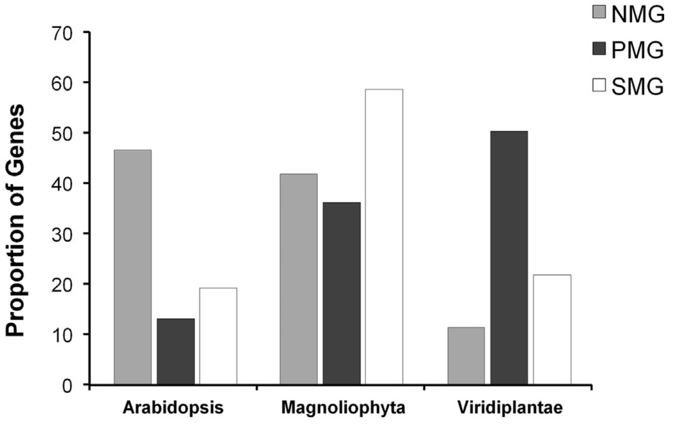

Fig. 4. Proportion of genes in three categories of phyletic age. All the proportions are significantly different (Chi-square test, $P < 0.001$).

are independent of expression level. It is also noticeable that there is a constant rise in PI up to category 3 of expression level and at category 4 (Highest expression values) there were no SMGs, so we could only compare PMGs and NMGs at this category. Interestingly, category 4 showed a lower magnitude of conservation level. This once again indicated that relationship between gene expression and conservation is non-linear. Next we analysed whether intron enrichment has an influence in governing the PI difference between PMGs, NMGs and SMGs. For this we binned intron number into four bins-Bin1 (containing genes having intron number 1–10), Bin2 (intron number 11–20), Bin3 (intron number 21–30) and Bin4 (intron number > 30). Bin 1 represented the group of genes with least number of introns, while Bin4 had genes with highest intron numbers. Bin3 and Bin4 also showed absence of any SMGs. It was observed that PI of PMGs, NMGs and SMGs did not follow any particular trend (Fig. 5B). This once again revealed that intron enrichment has a role in determining the conservation of the genes.

We propose a two-fold explanation behind such an observation. Firstly, it has been proposed earlier that older genes are under more complex regulation [44]. Introns have a definitive effect over gene expression regulation in both animals [45] and plants [46]. Hence, acquisition of large number of introns in plants could be due to the result of more complex regulation of older and highly persistent genes. Given the fact that intron number is correlated with gene conservation and intron number gradually increases with increase in degree of persistence, it is questionable that, what roles introns have in maintaining gene's degree of conservation.

3.5. Introns increase protein versatility

Previously it has been proposed that gradual segmentation of a given gene into smaller exonic regions interrupted by introns may facilitate alternative splicing and thus might increase protein versatility of the concerned gene [47]. It is quite obvious that genes with high protein diversity would tend to be more conserved than genes that yield fewer number of protein isoforms [20]. In other words, as genes grew older with time, it acquired many different number of spliced variants which increases its diversity in both transcript and protein level [48].

Intron enrichment is considerably higher in PMGs and they also acquire higher number of spliced variants as compared to NMGs and SMGs. As PMGs are older and are more persistent across diverse plant taxa, it is more likely for them to gain different molecular functions with time. It has been previously proposed that PMGs are more multifunctional in nature as compared to SMGs [25]. This high multifunctionality may be attributed to higher number of spliced variants and increased protein versatility. Previous studies have suggested intron number might increase protein versatility through alternative splicing [49]. Here we investigated that whether introns in *A. thaliana* increases protein diversity by the mechanism of alternative splicing. In this context, we estimated the number of unique proteins and unique transcripts per gene of PMGs, SMGs and NMGs respectively. It was observed that PMGs possessed significantly higher splice variant and protein diversity as compared to SMGs and NMGs (Table 1). We found that transcript and protein diversity (measured as number of unique transcript ids/protein ids) has a significant strong positive correlation with intron number (Spearman's $\rho_{\text{intron no-transcript count}} = 0.214$ $P < 10^{-6}$; Spearman's $\rho_{\text{intron no-protein count}} = 0.210$ $P < 10^{-6}$). The increased diversity in transcript and protein level could be due to alternative splicing of these genes. It was also revealed that PI is also correlated with multifunctionality (Spearman's $\rho_{\text{PI-Multifunctionality}}$ using GO biological process terms $= 0.219$, $P < 10^{-6}$; Spearman's $\rho_{\text{PI-Multi-functionality}}$ using Pfam domain number $= 0.017$, $P = 1.2 \times 10^{-2}$) and transcript count (Spearman's $\rho_{\text{PI-transcript count}} = 0.117$, $P < 10^{-6}$). Overall our data suggests that acquisition of large number of introns could eventually increase protein versatility through exon shuffling mechanisms which may ultimately cause conservation of genes in *A. thaliana.*

In order to gain multiple functions, primary genes might harbour elevated number of functional domains within them. It was observed that PMGs were indeed enriched in functional domains as compared to the SMGs (Mann-Whitney *U* Test, $P < 10^{-6}$). Thus, we hypothesize that intron enrichment in primary genes could be correlated with functional domain acquisition. In agreement to our hypothesis, we observed a significant correlation between functional domain count and

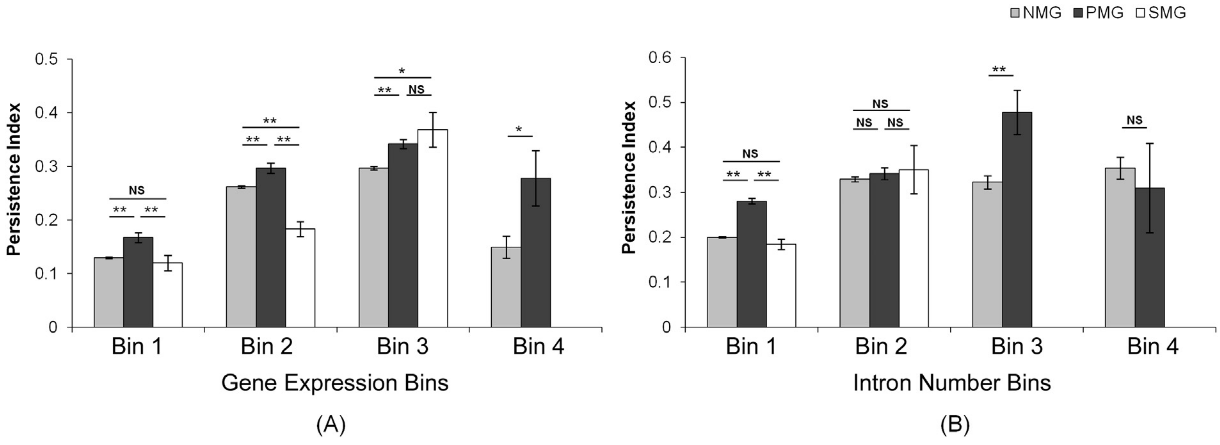

Fig. 5. Bar diagram showing the difference of Persistence index between PMGs and NMGs in each bin of (A) gene expression and (B) intron number. The expression values were divided into 4 bins: $2.0 < \text{Bin1} < 5.0$, $5.0 < \text{Bin2} < 8.0$, $8.0 < \text{Bin 3} < 11.0$, $\text{Bin 4} > 11$. The intron numbers were divided into 4 bins such that: $1 \leq \text{Bin1} \leq 10$, $11 \leq \text{Bin2} \leq 20$, $21 \leq \text{Bin3} \leq 30$, $\text{Bin 4} \geq 30$. ** denotes $P < 0.01$, * denotes $P < 0.05$ and NS denotes Not significantly different values.

Table 1
Details of Mann Whitney *U* test of transcript and protein diversity between PMGs, NMGs and SMGs of *A. thaliana.*

	PMGs	NMGs	SMGs	P-values
Transcript diversity				
Mean	1.44	1.28	1.25	$P_{PMG\text{-}NMG} = 10^{-6}$
Standard deviation	0.78	0.64	0.57	$P_{PMG\text{-}SMG} = 10^{-6}$
				$P_{NMG\text{-}SMG} = 0.53$
Protein diversity				
Mean	1.43	1.30	1.25	$P_{PMG\text{-}NMG} = 10^{-6}$
Standard deviation	0.79	0.69	0.57	$P_{PMG\text{-}SMG} = 10^{-6}$
				$P_{NMG\text{-}SMG} = 0.46$

intron number in PMGs (Spearman's $\rho_{\text{domain no-intron no}} = 0.176$, $P < 10^{-6}$). Our results thus, indicate that introns increase protein function by acquisition of functional domains, and thus plays important role in protein multifunctionality.

3.6. Introns may serve as buffer for mutations in coding regions

Another probable explanation behind acquisition of high number of introns could be the fact that introns being themselves non-coding might retain mutational disturbances and thus buffers the coding exons from mutations, as explained by Jo and Choi [18]. We, thus, analysed the single nucleotide polymorphisms (SNPs) as the mutational force. A strong negative correlation between intron number and exonic SNP density (Spearman's $\rho = -0.312$, $P < 10^{-6}$) accompanied by a significantly higher enrichment of SNPs in the intronic regions as compared to exonic ones suggest that along with alternative splicing, intron enrichment is helpful for persistence of old genes to protect themselves from mutations. On the other way round, SNPs in intronic region could also guide splicing as observed in many different previous studies [50]. In this study, we have also found that intronic SNP density is significantly correlated with transcript count (Spearman's $\rho = 0.155$, $P < 10^{-6}$) in *A. thaliana*. These results show that SNPs do have a role in alternative splicing mechanisms. Moreover, exonic SNP density was found to have a slight yet significant negative correlation with transcript count (Spearman's $\rho = -0.074$, $P < 10^{-6}$). Thus, proteins with fewer number of splice variants have a slightly more chance of gathering SNPs in the exonic regions. Thus, we have searched the intronic regions for the presence of SNPs and tried to understand their role in conservation.

Previous studies [51] suggested that mutation through single nucleotide polymorphisms (SNPs) are more in the intronic regions of the genes as compared to the exonic counterparts. It has also been suggested that introns could possibly buffer mutations and protect the exons [18]. In this study, we hypothesized that introns may absorb more mutational shocks which allow the genes to retain normal protein function and hence be conserved. To elaborate this, we studied the distribution pattern of SNPs. We have observed a significantly higher count of SNPs in the introns than exons in all the groups (Mann-Whitney *U* test, $P < 0.01$). Here, we observed introns of PMGs have highest SNP density followed by NMGs and least in SMGs and NMGs have significantly more exonic SNPs than PMGs and SMGs (Fig. 6 A and B). However, exonic SNP density of PMGs and SMGs did not vary significantly. In addition to it, as shown above, we obtained very strong negative correlation between total exonic SNPs density and intron number, indicating introns could possibly absorb the mutational load of the genes, which is also supported by the previous notion of Jo and Choi [18]. Finally, intronic SNP density showed a slight yet significant correlation with PI (Spearman's $\rho = 0.079$, $P = 2.35 \times 10^{-9}$). However, there was no correlation of total exonic SNP density with PI (Spearman's $\rho_{PI} = 0.011$, $P = 0.362$). This shows that intronic SNPs indeed have a role in evolutionary conservation of genes. To further authenticate the study, we generated the SIFT score of the SNPs of the coding exons of PMGs, NMGs and SMGs using the webserver (SIFT 4.0) [52]. Density of deleterious mutations (no. of deleterious mutations/cds length) was highest in SMG (0.01), followed by NMG (0.008) and PMG (0.006) (Sig at $P < 0.01$, Mann-Whitney *U* test). Moreover, intron number is significantly negatively correlated with this density of deleterious mutations (Spearman's rho of -0.036, $P < 0.001$). We have also showed that intron number is highest in PMG, followed by NMG and SMG. Surprisingly, this is also true for density of tolerated mutations (no. of tolerated mutations/cds length). It was highest in SMG (0.05), followed by NMG (0.048) and PMG (0.041) (Sig at $P < 0.01$, Mann-Whitney *U* test). Moreover, intron number is significantly negatively correlated with this density of deleterious mutations (Spearman's rho of -0.167, $P < 0.001$). Thus, it is clear that more number of introns somehow preventing the gene from accumulating more mutations (be it deleterious or tolerated) in the coding regions. However, it may be the fact that as introns rich genes are highly expressed, mutation accumulation is less [26]. We also conducted the cause and effect estimation of intron number and deleterious/tolerated mutations to understand the influence of the factors based on van der Lee et al. [53]. The result in both cases reveals the number of introns to be the cause of mutations be it deleterious or tolerated (Table 2) The presence of introns within the coding regions brings down the overall mutation of the exons, leading to the functional conservation of vital primary metabolic genes.

4. Conclusions

Primitiveness and conservation of metabolic genes is largely correlated with intron number and is expression independent. Unlike that of animal genomes, where housekeeping genes possesses shorter introns and have compact genetic architecture; Primary metabolic genes of *A. thaliana* (which has a basic housekeeping functionality) represent quite a different and unique set of characters. As a matter of fact PMGs share a combination of features that partially resembles both housekeeping and tissue specific genes. PMGs are pan-taxonomically conserved like that of housekeeping genes, but unlike animal housekeeping genes entails higher intron enrichment. Plants being autotrophic can harness their own energy. Hence, energy cost for processing large number of introns might not be a limitation amongst PMGs. At the same time primary genes, in course of evolution could give birth to secondary metabolic genes, which are again tissue specific. Thus, primary genes represent a complex trade-off between housekeeping and tissue specific genetic architectures in *A. thaliana.*

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2017.12.003.

Abbreviations

PMG	primary metabolic gene
SMG	secondary metabolic gene
NMG	non metabolic gene
PI	persistence index
AI	age index
SNP	single nucleotide polymorphism

Declarations

Availability of data and materials

All data were obtained from publicly available databases (mentioned in the Materials and methods section) and are freely available online. The dataset used in the study can be found in the Additional file 1 (Microsoft Excel Worksheet). Furthermore, the datasets used and/or analysed during the current study are also available from the corresponding author on reasonable request.

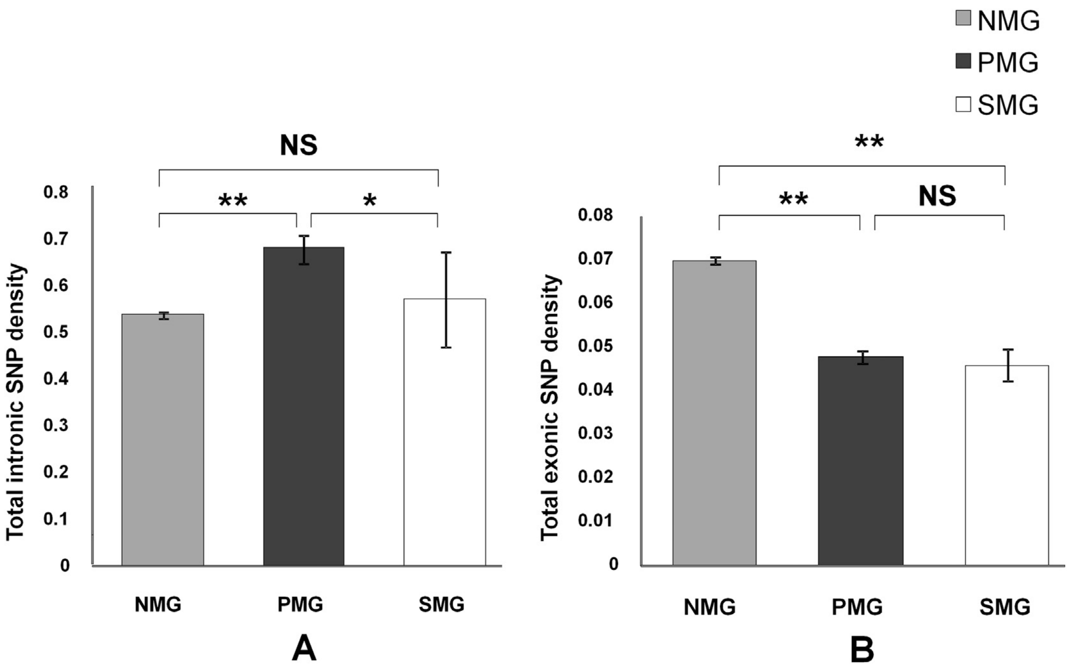

Fig. 6. Bar diagram showing the average values of (A) total intronic SNP density and (B) total exonic SNP density amongst different groups, PMGs, NMGs and SMGs. ** denotes P < 0.01, * denotes P < 0.05 and NS denotes Not significantly different values.

Table 2
Conditional probability study between intron number and Tolerated/Deleterious mutations.

		Tolerant mutation		
		High (T_H)	Low (T_L)	
Intron number	High (I_H)	4077	7432	11509
	Low (I_L)	9138	5788	14926
		13,215	13220	
		Deleterious mutation		
		High (D_H)	Low (D_L)	
Intron number	High (I_H)	5983	5559	11542
	Low (I_L)	7337	8057	15394
		13320	13616	

	Event(E)	Condition (C)	Probability (Event|Condition) = P(E∩C)/P(C)
A	Deleterious mutation low	High intron number	$P(D_L\|I_H) = \frac{5559}{11542} = 0.482$
	High intron number	Low deleterious mutation	$P(I_H\|D_L) = \frac{5559}{13616} = 0.408$
B	Tolerant mutation low	High intron number	$P(T_L\|I_H) = \frac{7432}{11542} = 0.644$
	High intron number	Low tolerant mutation	$P(I_H\|T_L) = \frac{7432}{13220} = 0.562$

Competing interests

The authors declare that no financial and/or non-financial competing interests exist.

Funding

No funding information is applicable for this manuscript.

Authors' contributions

Conceived and designed the experiments: DM, DS, DA, AM, TCG. Performed the experiments: DM, DS, SC. Analysed the data: DM, DS, AM. Wrote the paper: DM, DS, DA, AM.

Acknowledgements

The authors are thankful to Department of Biotechnology, Government of India, for funding the Bioinformatice Centre at Bose Institute Kolkata, where the work was done.

Conflict of interest: The authors declared that they had no conflict of interests.

References

[1] A.I. Lamond, RNA splicing — running rings around RNA, Nature 397 (6721) (1999) 655–656.
[2] F. Rodriguez-Trelles, R. Tarrio, F.J. Ayala, Origins and evolution of spliceosomal introns, Annu. Rev. Genet. 40 (2006) 47–76.
[3] S.W. Roy, W. Gilbert, Rates of intron loss and gain: implications for early eukaryotic evolution, Proc. Natl. Acad. Sci. U. S. A. 102 (16) (2005) 5773–5778.
[4] D. Jeffares, Rapidly regulated genes are intron poor (vol 24, pg 375, 2008), Trends Genet. 24 (10) (2008) (488–488).
[5] Y.-F. Yang, T. Zhu, D.-K. Niu, Association of intron loss with high mutation rate in *Arabidopsis*: implications for genome size evolution, Genome Biol. Evol. 5 (4) (2013) 723–733.
[6] G. Parra, K. Bradnam, A.B. Rose, I. Korf, Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants, Nucleic Acids Res. 39 (13) (2011) 5328–5337.
[7] M. Chorev, L. Carmel, The function of introns, Front. Genet. 3 (2012) 55.
[8] J.F. Wendel, R.C. Cronn, I. Alvarez, B. Liu, R.L. Small, D.S. Senchina, Intron size and genome size in plants, Mol. Biol. Evol. 19 (12) (2002) 2346–2352.
[9] K. Jiang, L.R. Goertzen, Spliceosomal intron size expansion in domesticated grapevine (*Vitis vinifera*), BMC. Res. Notes 4 (2011) (52–52).
[10] D.S. Ucker, K.R. Yamamoto, Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates, J. Biol. Chem. 259 (12) (1984) 7416–7420.
[11] M.G. Izban, D.S. Luse, Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates, J. Biol. Chem. 267 (19) (1992) 13647–13655.
[12] E. Eisenberg, E.Y. Levanon, Human housekeeping genes are compact, Trends Genet. 19 (7) (2003) 362–365.
[13] H. Yang, In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure, Biol. Direct 4 (2009) 45 (discussion 45).
[14] C.I. Castillo-Davis, S.L. Mekhedov, D.L. Hartl, E.V. Koonin, F.A. Kondrashov, Selection for short introns in highly expressed genes, Nat. Genet. 31 (4) (2002) 415–418.
[15] M. Lynch, The origins of eukaryotic gene structure, Mol. Biol. Evol. 23 (2) (2006)

450–468.
[16] M. Lynch, The Origins of Genome Architecture, Vol. 98 Sinauer Associates Sunderland, 2007.
[17] M. Lynch, B. Koskella, S. Schaack, Mutation pressure and the evolution of organelle genomic architecture, Science 311 (5768) (2006) 1727–1730.
[18] B.S. Jo, S.S. Choi, Introns: the functional benefits of introns in genomes, Genomics Inform. 13 (4) (2015) 112–118.
[19] A. Kalsotra, T.A. Cooper, Functional consequences of developmentally regulated alternative splicing, Nat. Rev. Genet. 12 (10) (2011) 715–729.
[20] O. Gorlova, A. Fedorov, C. Logothetis, C. Amos, I. Gorlov, Genes with a large intronic burden show greater evolutionary conservation on the protein level, BMC Evol. Biol. 14 (1) (2014) 50.
[21] T. Maniatis, R. Reed, An extensive network of coupling among gene expression machines, Nature 416 (6880) (2002) 499–506.
[22] S.Y. Ying, S.L. Lin, Intronic microRNAs, Biochem. Biophys. Res. Commun. 326 (3) (2005) 515–520.
[23] D. Mascarenhas, I.J. Mettler, D.A. Pierce, H.W. Lowe, Intron-mediated enhancement of heterologous gene expression in maize, Plant Mol. Biol. 15 (6) (1990) 913–920.
[24] A.E. Vinogradov, Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet. 20 (5) (2004) 248–253.
[25] D. Mukherjee, A. Mukherjee, T.C. Ghosh, Evolutionary rate heterogeneity of primary and secondary metabolic pathway genes in *Arabidopsis thaliana*, Genome Biol. Evol. 8 (1) (2016) 17–28.
[26] X.-Y. Ren, O. Vorst, M.W.E.J. Fiers, W.J. Stiekema, J.-P. Nap, In plants, highly expressed genes are the least compact, Trends Genet. 22 (10) (2006) 528–532.
[27] Y.S. Rao, Z.F. Wang, X.W. Chai, Wu GZ, M. Zhou, Q.H. Nie, X.Q. Zhang, Selection for the compactness of highly expressed genes in *Gallus gallus*, Biol. Direct 5 (2010).
[28] E. Eisenberg, E.Y. Levanon, Human housekeeping genes are compact, Trends Genet. 19 (7) (2003) 362–365.
[29] C. Seoighe, C. Gehring, L.D. Hurst, Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction, PLoS Genet. 1 (2) (2005) e13.
[30] E. Pichersky, D.R. Gang, Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective, Trends Plant Sci. 5 (10) (2000) 439–445.
[31] M. Schena, The evolutionary conservation of eukaryotic gene-transcription, Experientia 45 (10) (1989) 972–983.
[32] J.Y. Yuan, Evolutionary conservation of a genetic pathway of programmed cell death, J. Cell. Biochem. 60 (1) (1996) 4–11.
[33] R.J. Kinsella, A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, et al., Ensembl BioMarts: a hub for data retrieval across taxonomic space, Database: The Journal of Biological Databases and Curation 2011 (2011) bar030.
[34] P.J. Kersey, J.E. Allen, M. Christensen, P. Davis, L.J. Falin, C. Grabmueller, D.S.T. Hughes, J. Humphrey, A. Kerhornou, J. Khobova, et al., Ensembl genomes 2013: scaling up access to genome-wide data, Nucleic Acids Res. 42 (2014) D546–D552.
[35] M. Kanehisa, S. Goto, KEGG: Kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[36] Y.-L. Guo, Gene family evolution in green plants with emphasis on the origination and evolution of *Arabidopsis thaliana* genes, Plant J. 73 (6) (2013) 941–951.
[37] S. Dash, J. Van Hemert, L. Hong, R.P. Wise, J.A. Dickerson, PLEXdb: gene expression resources for plants and plant pathogens, Nucleic Acids Res. 40 (D1) (2012) D1194–D1201.
[38] The Gene Ontology C, M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25–29.
[39] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, et al., The Pfam protein families database, Nucleic Acids Res. 40 (Database issue) (2012) D290–D301.
[40] H.B. Mann, D.R. Whitney, On a test of whether one of 2 random variables is stochastically larger than the other, Ann. Math. Stat. 18 (1) (1947) 50–60.
[41] C. Spearman, The proof and measurement of association between two things, Am. J. Psychol. 15 (1904) 72–101.
[42] L. Carmel, I.B. Rogozin, Y.I. Wolf, E.V. Koonin, Evolutionarily conserved genes preferentially accumulate introns, Genome Res. 17 (7) (2007) 1045–1050.
[43] S. Subramanian, S. Kumar, Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome, Genetics 168 (1) (2004) 373–381.
[44] M. Warnefors, A. Eyre-Walker, The accumulation of gene regulation through time, Genome Biol. Evol. 3 (2011) 667–673.
[45] Y. Imamichi, T. Mizutani, Y. Ju, T. Matsumura, S. Kawabe, M. Kanno, T. Yazawa, K. Miyamoto, Transcriptional regulation of human ferredoxin reductase through an intronic enhancer in steroidogenic cells, Biochim. Biophys. Acta, Gene Regul. Mech. 1839 (1) (2014) 33–42.
[46] A.B. Rose, J.A. Beliakoff, Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing, Plant Physiol. 122 (2) (2000) 535–542.
[47] J. Morata, S. Bejar, D. Talavera, C. Riera, S. Lois, G. Mas de Xaxars, X. de la Cruz, The relationship between gene isoform multiplicity, number of exons and protein divergence, PLoS One 8 (8) (2013).
[48] J. Roux, M. Robinson-Rechavi, Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication, Genome Res. 21 (3) (2011) 357–363.
[49] B.R. Graveley, Alternative splicing: increasing diversity in the proteomic world, Trends Genet. 17 (2) (2001) 100–107.
[50] R.A. Moyer, D. Wang, A.C. Papp, R.M. Smith, L. Duque, D.C. Mash, W. Sadee, Intronic polymorphisms affecting alternative splicing of human dopamine D2 receptor are associated with cocaine abuse, Neuropsychopharmacology 36 (4) (2011) 753–762.
[51] J. Evans, J. Kim, K.L. Childs, B. Vaillancourt, E. Crisovan, A. Nandety, D.J. Gerhardt, T.A. Richmond, J.A. Jeddeloh, S.M. Kaeppler, et al., Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass *Panicum virgatum*, Plant J. 79 (6) (2014) 993–1008.
[52] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, P.C. Ng, SIFT web server: predicting effects of amino acid substitutions on proteins, Nucleic Acids Res. 40 (W1) (2012) W452–W457.
[53] R. van der Lee, B. Lang, K. Kruse, J. Gsponer, N.S. de Groot, M.A. Huynen, A. Matouschek, M. Fuxreiter, M.M. Babu, Intrinsically disordered segments affect protein half-life in the cell and during evolution, Cell Rep. 8 (6) (2014) 1832–1844.