Physiol Mol Biol Plants
https://doi.org/10.1007/s12298-022-01188-2

RESEARCH ARTICLE

Interplay between gene expression and gene architecture as a consequence of gene and genome duplications: evidence from metabolic genes of *Arabidopsis thaliana*

Dola Mukherjee[1] · Deeya Saha[1] · Debarun Acharya[1] · Ashutosh Mukherjee[2] · Tapash Chandra Ghosh[1]

Received: 30 March 2021 / Revised: 16 May 2022 / Accepted: 18 May 2022

Abstract Gene and genome duplications have been widespread during the evolution of flowering plant which resulted in the increment of biological complexity as well as creation of plasticity of a genome helping the species to adapt to changing environments. Duplicated genes with higher evolutionary rates can act as a mechanism of generating novel functions in secondary metabolism. In this study, we explored duplication as a potential factor governing the expression heterogeneity and gene architecture of Primary Metabolic Genes (PMGs) and Secondary Metabolic Genes (SMGs) of *Arabidopsis thaliana*. It is remarkable that different types of duplication processes controlled gene expression and tissue specificity differently in PMGs and SMGs. A complex relationship exists between gene architecture and expression patterns of primary and secondary metabolic genes. Our study reflects, expression heterogeneity and gene structure variation of primary and secondary metabolism in *Arabidopsis thaliana* are partly results of duplication events of different origins. Our study suggests that duplication has differential effect on PMGs and SMGs regarding expression pattern by controlling gene structure, epigenetic modifications, multifunctionality and subcellular compartmentalization. This study provides an insight into the evolution of metabolism in plants in the light of gene and genome scale duplication.

Keywords Whole genome duplication · Evolutionary rate · Tissue specificity · Essentiality · Methylation

Introduction

It is well known that *A. thaliana* has experienced at least three WGD (whole genome duplication) events as well as many single-gene duplications which contributed to the origin of a considerable portion of genes in this species (Wang et al. 2013). Of these, alpha- and beta-WGD are recent events causing the divergence of *Arabidopsis* from other members of the Brassicales clade while gamma-WGD was a more ancient event (Bowers et al. 2003). The primary and secondary metabolic genes must have experienced one or more of these duplication events and their expression and evolution must have been shaped by these duplication events.

Gene duplication is one major determinant for gene expression divergence (Panchy et al. 2016). Among eukaryotic organisms, plants are exceptional as duplicate loci compose a large fraction of their genomes (Moore and Purugganan 2005). For example, in *Arabidopsis thaliana*, 60% of the total nuclear genes arose from duplication (The Arabidopsis Genome Initiative 2000). Gene duplication can arise from chromosomal or genome duplication, unequal crossing over or retroposition (Zhang 2003). Small-scale duplication (often involves a single gene) can occur at any time and may be retained in the course of evolution while large-scale duplication may involve many genes or even the entire genome, with the later being known as whole-genome duplication (WGD) (Hakes et al.

✉ Ashutosh Mukherjee
ashutoshcaluniv@gmail.com

✉ Tapash Chandra Ghosh
tapashbic@gmail.com

[1] Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, India

[2] Department of Botany, Vivekananda College, 269, Diamond Harbour Road, Thakurpukur, Kolkata, West Bengal 700063, India

Published online: 02 June 2022

2007). The outcomes of these duplication processes are quite different (Zhang 2003). For example, in plants, genes related to protein kinases, transcription factors and ribosomal proteins are preferentially retained following WGDs (Blanc and Wolfe 2004; Maere et al. 2005), genes associated with biotic and abiotic stresses are more likely to be retained following local duplications (Hanada et al. 2009; Rizzon et al. 2006) while transpositions of genes are more common in some gene families such as MADS-box, F-box, NBS-LRR, and defensins than others (Freeling 2009; Freeling et al. 2008). While gene duplication is an important mechanism for evolution of functional novelty (Crow and Wagner 2006; Flagel and Wendel 2009; Magadum et al. 2013; Wang et al. 2011), the fate of duplicated genes resulting from WGD and small-scale duplication events differs (Wang et al. 2013).

Duplication is also responsible for increased metabolic diversity in plants. For example, Chae et al. (2014) showed that the expansion of specialized metabolic genes in plants appears to be the result of local gene duplication. Moreover, it has been shown that the two recent WGD events gave rise to novel pathways for the synthesis of indole and Met-derived glucosinolates (Kliebenstein 2008). In fact, according to Pichersky and Gang (2000), gene duplication is assumed to be a major driving force for the recruitment of genes for plant secondary metabolism. In *Arabidopsis thaliana*, primary metabolic genes (PMGs) show higher expression level than secondary metabolic genes (SMGs) while SMGs are more tissue specific than PMGs (Mukherjee et al. 2016). However, we have previously reported that unlike animals where compact genes show higher expression, PMGs in *A. thaliana* are highly expressed in spite of being significantly longer than SMGs (Mukherjee et al. 2016). Moreover, in *A. thaliana*, PMGs are, in general, older than SMGs and have orthologous genes in different plant lineages ranging from unicellular alga to higher plants (Mukherjee et al. 2018). Thus, it seems that the expression heterogeneity and genic properties of metabolic genes in *A. thaliana* is connected to the history of its genome evolution.

Considering all these facts, in this study, we have tried to address the relationship of gene expression patterns with gene architecture of primary and secondary metabolic genes of *A. thaliana* in the light of gene and genome duplication.

Materials and method

Dataset preparation

Metabolic genes of *Arabidopsis thaliana* (L.) Heynh. was obtained from the supplementary data of Mukherjee et al. (2016). Initially, we had compiled a dataset of 2512 metabolic genes out of which 2030 were PMGs and 482 were SMGs. We had then filtered out 209 metabolic genes from our dataset which participated in both primary and secondary metabolic pathways. Finally, we had 1821 PMGs and 273 SMGs. Paralogs with at least 40% sequence identity were considered as duplicates and these paralogs were obtained from the Biomart interface (http://plants.ensembl.org/biomart/martview) (Kinsella et al. 2011) of Ensembl Plants (Yates et al. 2022).

Expression data

We had obtained the microarray expression data for *A. thaliana* from PLEXdb (www.plexdb.org/) (Dash et al. 2012). The accession number of expression dataset was AT40 and the microarray platform that was used was ATH1-121501. The tissue specificity index τ (tau) was measured following Yanai et al. (2005) as follows.

$$\tau = \frac{\sum_{j=1}^{n}\left[1 - \frac{\log S(i,j)}{\log S(i,\max)}\right]}{n-1}$$

where n is the number of tissues and conditions, and S (*i, max*) is the highest expression of gene *i* across the n tissues. The index τ ranges from 0 to 1, with higher value signifying higher specificity.

Dataset preparation of mechanism of origin of duplicates

We had prepared the dataset for classifying the PMGs and SMGs according to the modes of duplication from the supplementary data of Wang et al. (2013). We had considered all three broad modes of gene duplication in *Arabidopsis*: Whole Genome Duplication (WGD), Local Duplication (LD) and Transposed Duplication (TD) as the major origin of duplicates in *Arabidopsis thaliana*. Moreover, subtypes of WGD (alpha, beta and gamma), LD (tandem and proximal) and TD [duplication occurred < 16 Million years ago (Mya) and 16-107 Mya] have also been considered. We have identified 1152 such duplication events involving our dataset from the supplementary data of Wang et al. (2013). Some singleton genes as considered by this study also showed some mode of duplication from the data of Wang et al. (2013). For the study of different genic, functional or epigenetic parameters, we have not included them in WGD, TD or LD groups. The final dataset of duplicated and singleton genes of both PMGs and SMGs with mode of duplication is given in the supplementary table.

Evolutionary rate determination and gene structural parameters

We had extracted the corresponding *Arabidopsis lyrata* orthologs (with 1:1 orthology and with at least 80% sequence similarity) of the *Arabidopsis thaliana* genes using Biomart interface of Ensembl Plants database (http://plants.ensembl.org/biomart/martview) as well as obtained their pair wise non-synonymous (d_N-rate of non-synonymous substitution per non-synonymous site) and synonymous (d_S -rate of synonymous substitution per synonymous site) substitution rates to compute gene specific evolutionary rate (d_N / d_S). Protein coding sequences of these genes were also acquired from Ensembl database. For genes with more than one isoform, the longest isoform was considered. Also, we had studied the synonymous substitution rate (d_S) of the paralogs as a proxy of age of the duplicated genes. We had also obtained gene length, coding sequence length, intron number, exon length and number of transcripts per gene from Biomart.

Functional parameters and essentiality

The functionality of the metabolic genes was measured through study of GO (Gene Ontology) terms associated with the genes which were obtained from Biomart. For the study of essentiality, we have obtained lethal phenotype scores of *A. thaliana* genes from the supplementary data of Lloyd et al. (2015). This score is a value between 0 and 1 where higher values signify higher confidence whose disruption causes a gene to display a lethal phenotype (Lloyd et al. 2015). As the authors designated "lethal genes" as "essential genes", we have designated the "lethal phenotype scores" as "essentiality" in this study. It should be noted that essentiality of SMGs are hard to estimate as these are more involved in environmental responses. However, comparative study of lethal phenotype scores will give a overall view of essentiality of PMGs and SMGs.

Catalytic properties of metabolic enzymes

As the dataset of metabolic genes encodes for metabolic enzymes, we decided to analyze catalytic properties of these enzymes. To study this, we chose product diversity of the PMGs and SMGs as a measure of enzyme's functionality. As enzymes catalyse biochemical reactions, we obtained the number of products catalyzed by these metabolic enzymes. The biochemical reactions were obtained from AraCyc (https://www.arabidopsis.org/biocyc/) (Mueller et al. 2003).

Epigenetic parameters

The epigenetics of the metabolic genes were studied through promoter and body methylation. The data of both promoter and body methylation were obtained from TEA database (http://tea.iis.sinica.edu.tw) (Su et al. 2011).

Statistical analyses

Statistical analyses were performed using SPSS v.13. Mann–Whitney *U* test (Mann and Whitney 1947) with Bonferroni correction 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 (Spearman 1904) analysis, where the significant correlations were denoted by $P < 0.05$. To study the significant proportion test between two groups, *Z*-test analysis was carried out.

Results

Duplication patterns in PMGs and SMGs

It was observed that SMGs possess significantly more duplicated genes (81.68%) compared to PMGs (73.31%) ($Z = 2.95$; $P < 0.05$). Of these, we have obatined 1152 duplication events from the supplementary data of Wang et al. (2013). Of these, WGDs are significantly more widespread in PMGs (44.34%) than SMGs (23.32%) ($Z = 5.90$, $p < 10^{-5}$). However, of different kinds of WGDs, alpha and gamma-WGDs are significantly more widespread in PMGs than SMGs (Fig. 1a). On the contrary, LDs are more prevalent in SMGs (34.98%) than PMGs (11.31%) ($Z = 9.24$; $p < 10^{-5}$). Of these, tandem duplication created significantly ($P < 0.01$) more duplicates in SMGs than PMGs (Fig. 1a). Proximal duplications showed no significant difference in this regard among two groups. We considered synonymous substitution rate per synonymous site (d_S) between paralogs as a proxy measure of time of duplication among PMG and SMG duplicates. It was observed that in both PMGs and SMGs, LD duplicates are the most recent whereas TD duplicates are most primitive and WGDs are in between two (Table 1).

Regarding paralog number, PMGs shows the following pattern: LD > WGD ≈ TD ($p < 0.05$, Mann–Whitney *U* test with Bonferroni correction). SMGs, on the contrary, showed the following pattern: LD > TD > WGD (Table 1) ($p < 0.05$, Mann–Whitney *U* test with Bonferroni correction). In both the cases, mean paralog numbers of LDs are significantly higher than WGDs and TDs (Fig. 1b, c), showing that local duplications produce more gene copy

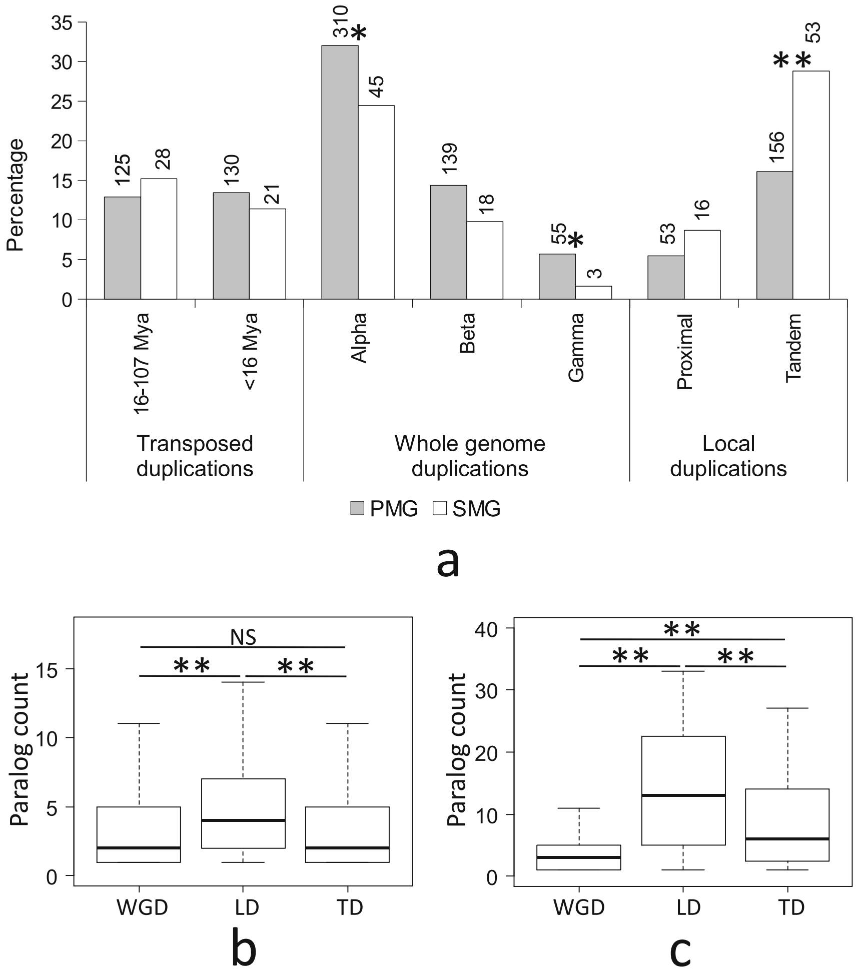

Fig. 1 **a** Percentage of PMGs and SMGs in different types of duplication events. Gene duplication numbers of each category are also shown above each bar. * and ** indicate significance at $P < 0.05$ and $P < 0.01$, respectively (Z test). **b** Difference in paralog count in different types of duplicates in PMGs; **c** Difference in paralog count in different types of duplicates in SMGs. NS and ** indicate nonsignificant and significance at $P < 0.01$, respectively (Mann–Whitney *U* test with Bonferroni correction)

numbers than other means of duplications in *A. thaliana* metabolic genes. Moreover, paralog numbers of LDs and TDs of SMGs are higher compared to that of PMGs while no such difference was observed regarding WGDs (Table 1). It indicates that gene level duplication has a profound role in increasing the gene number in SMGs irrespective of their age while genome duplication has similar effects in both primary and secondary metabolism. As α WGD predates the divergence of *Arabidopsis* and *Brassica* (Yu et al. 2017), these results suggest that several PMGs evolved in *Arabidopsis*-*Brassica* clade, while expansion of many SMGs occurred in the genus *Arabidopsis* by local duplications.

Evolutionary rate heterogeneity of PMGs and SMGs in the light of duplication

It is known that in general, PMGs have been more conserved than SMGs during their evolution (Mukherjee et al. 2016). To analyse evolutionary rate heterogeneity between singletons and duplicates and between different types of duplicates, we have performed Mann–Whitney *U* test with Bonferroni correction. No significant difference has been found regarding d_N/d_S between singletons of PMGs (0.149 ± 0.006) and SMGs (0.155 ± 0.020) ($p > 0.05$). However, duplicates of PMGs have been observed as more conserved than duplicates of SMGs during their evolution (mean d_N/d_S of PMG duplicates = 0.137 ± 0.004, mean d_N/d_S of SMG duplicates = 0.193 ± 0.011; $p = 1.83 \times 10^{-8}$, Mann–Whitney *U* test with Bonferroni correction). Again, regarding origin of duplicates, PMGs and SMGs both showed no significant differences between LDs and TDs, but in PMGs, WGDs are more conserved than TDs during their evolution ($p < 0.05$) and LDs ($p < 0.01$). In SMGs, however, although WGDs are significantly more conserved than LDs ($p < 0.01$), it has not been significantly more conserved than TDs ($p > 0.05$). Interestingly, only LDs of PMGs are significantly more conserved than LDs of SMGs ($p < 0.05$). The TDs and WGDs are similar in PMGs and SMGs ($p > 0.05$). Thus, as local duplications of SMGs are predominant, they actually contribute to the overall higher evolutionary rate of SMGs compared to PMGs. It was also observed that only in SMG duplicates, paralog number is significantly correlated to $d_N/$

Table 1 Difference between different parameters of different types of duplicated genes in PMGs and SMGs

Parameters	PMGs (Mean ± SE)	SMGs (Mean ± SE)	*P*-value
Paralog count			
WGD	3.85 ± 0.16	4.24 ± 0.54	0.92^{NS}
TD	3.46 ± 0.20	8.45 ± 1.03	9.54×10^{-8}***
LD	4.85 ± 0.27	14.45 ± 1.21	6.68×10^{-13}***
d_S			
WGD	11.87 ± 0.62	9.55 ± 1.78	0.788^{NS}
TD	17.70 ± 1.26	12.70 ± 2.70	1.0^{NS}
LD	9.12 ± 1.01	5.18 ± 0.81	0.95^{NS}
Expression level			
WGD	7.74 ± 0.10	7.06 ± 0.29	0.068^{NS}
TD	7.68 ± 0.15	6.43 ± 0.34	0.004**
LD	6.70 ± 0.22	5.53 ± 0.22	0.01*
Tissue specificity (tau)			
WGD	0.24 ± 0.01	0.29 ± 0.02	0.006**
TD	0.22 ± 0.01	0.30 ± 0.03	0.002**
LD	0.31 ± 0.01	0.33 ± 0.02	0.662^{NS}
Gene length (bp)			
WGD	2815.00 ± 50.45	2375.71 ± 104.26	0.046*
TD	2992.80 ± 88.08	2098.25 ± 94.33	2.8×10^{-6}***
LD	2367.71 ± 88.81	2321.54 ± 122.92	0.738^{NS}
Intron number			
WGD	6.33 ± 0.22	2.92 ± 0.38	4.88×10^{-6}***
TD	6.63 ± 0.31	2.73 ± 0.43	1.25×10^{-8}***
LD	4.80 ± 0.33	3.01 ± 0.42	1.32×10^{-4}***
Average exon length (bp)			
WGD	421.42 ± 18.53	703.49 ± 91.39	1.43×10^{-5}***
TD	395.09 ± 23.69	689.39 ± 67.73	3.5×10^{-8}***
LD	431.02 ± 35.91	639.93 ± 50.42	9.28×10^{-6}***
Transcript count			
WGD	1.80 ± 0.05	1.69 ± 0.14	1.0^{NS}
TD	1.90 ± 0.08	1.33 ± 0.10	0.006**
LD	1.74 ± 0.10	1.51 ± 0.12	0.14^{NS}
Molecular function			
WGD	4.72 ± 0.09	5.34 ± 0.23	0.02*
TD	4.97 ± 0.13	5.90 ± 0.28	3.6×10^{-4}***
LD	4.58 ± 0.15	5.42 ± 0.30	0.02*
Biological process			
WGD	4.38 ± 0.11	5.04 ± 0.46	0.14^{NS}
TD	4.82 ± 0.17	3.65 ± 0.30	0.008**
LD	4.29 ± 0.21	2.51 ± 0.17	1.64×10^{-8}***
Cellular compartments			
WGD	4.02 ± 0.12	2.95 ± 0.21	0.08^{NS}
TD	3.73 ± 0.14	2.52 ± 0.18	0.006**
LD	3.35 ± 0.21	2.39 ± 0.16	0.14^{NS}
Essentiality			
WGD	0.102 ± 0.003	0.063 ± 0.007	6×10^{-4}***
TD	0.181 ± 0.006	0.111 ± 0.014	2.16×10^{-6}***
LD	0.098 ± 0.005	0.086 ± 0.004	0.52^{NS}

^{NS}indicates nonsignificant variation while *, ** and *** indicate significant variation at $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively (Mann–Whitney *U* test with Bonferroni correction)

d_S (Spearman's ρ = 0.330, $p = 0.001$). Thus, gene duplication appears to favour SMGs to increase the protein diversity. On the other hand, paralog count is not correlated with d_N/d_S in PMGs, indicating duplication is not favouring these genes towards novelty, rather they maintain their conserved nature.

Effect of duplication on gene expression patterns of PMGs and SMGs

It was observed that in both PMGs and SMGs, the expression level was significantly lower in duplicates as compared to singletons (Table 2) (Fig. 2a) while tau *i.e.* tissue specificity was significantly higher in duplicates as compared to singletons (Table 2) (Fig. 2b). For both duplicates and singletons, PMGs showed significantly higher expression level than SMGs ($p = 0.022$ for singletons; $p = 5.02 \times 10^{-51}$ for duplicates, Mann–Whitney *U* test with Bonferroni correction). On the contrary, SMGs showed higher tissue specificity than PMGs in both duplicates and singletons ($p = 0.028$ for singletons; $p = 6.86 \times 10^{-12}$ for duplicates, Mann–Whitney *U* test with Bonferroni correction). However, looking at the *p*-value, it is evident that regarding gene expression and tissue specificity, although significant differences exist between PMG and SMG singletons, the difference is much higher and more significant in duplicates. Further, in PMGs, significant reduction in expression level in LDs compared to TDs and WGD has been observed (Fig. 3a). However, genes duplicated by LD showed significantly

Table 2 Difference between different parameters of duplicated and singletone genes in PMGs and SMGs (*p*-values obtained with Mann–Whitney *U* test with Bonferroni correction)

Parameter	PMG (Mean ± SE)			SMG (Mean ± SE)		
	Duplicates	Singletons	*P*-value	Duplicates	Singletons	*P*-value
Expression pattern						
Expression level	7.64 ± 0.07	8.45 ± 0.08	4.02×10^{-10}***	6.21 ± 0.15	7.42 ± 0.37	0.006**
Tissue specificity (tau)	0.24 ± 0.004	0.15 ± 0.004	2.6×10^{-29}***	0.32 ± 0.01	0.23 ± 0.03	6.6×10^{-4}***
Gene architecture						
Gene length (bp)	2847.43 ± 36.64	3095.32 ± 71.37	0.004**	2386.72 ± 68.40	2900.08 ± 181.19	0.004**
Intron number	6.31 ± 0.14	7.81 ± 0.26	1.63×10^{-7}***	3.27 ± 0.24	6.94 ± 0.64	2.46×10^{-8}***
Average exon length	408.84 ± 11.72	306.71 ± 13.97	1.53×10^{-7}***	640.91 ± 34.29	313.46 ± 45.61	9.96×10^{-9}***
Transcript count	1.84 ± 0.04	1.92 ± 0.06	0.06^{NS}	1.55 ± 0.07	1.94 ± 0.14	0.002**
Multifunctionality						
Molecular function	4.78 ± 0.6	4.68 ± 0.11	0.236^{NS}	5.53 ± 0.15	5.20 ± 0.24	0.404^{NS}
Biological process	4.51 ± 0.07	4.29 ± 0.13	0.122^{NS}	3.71 ± 0.18	4.47 ± 0.34	0.022*
Cellular compartments	3.69 ± 0.07	3.99 ± 0.12	0.008**	2.56 ± 0.10	3.27 ± 0.23	0.004**
Product diversity	8.02 ± 0.45	6.71 ± 0.53	1.0^{NS}	37.20 ± 4.11	9.8 ± 3.58	0.006**
Essentiality	0.124 ± 0.002	0.316 ± 0.006	3.6×10^{-110}***	0.083 ± 0.003	0.273 ± 0.033	1.11×10^{-9}***
Promoter methylation						
CG methylation	0.104 ± 0.005	0.114 ± 0.008	1.68×10^{-4}***	0.077 ± 0.010	0.102 ± 0.022	0.998^{NS}
CHG methylation	0.054 ± 0.003	0.058 ± 0.005	0.02*	0.036 ± 0.006	0.032 ± 0.007	0.002**
CHH methylation	*0.036 ± 0.002*	*0.037 ± 0.003*	0.08^{NS}	0.024 ± 0.003	0.025 ± 0.005	0.03*
Gene body methylation						
CG methylation	0.140 ± 0.004	0.195 ± 0.007	1.18×10^{-16}***	0.073 ± 0.006	0.150 ± 0.018	0.002**
CHG methylation	0.027 ± 0.001	0.020 ± 0.0004	0.02*	0.026 ± 0.002	0.019 ± 0.001	0.228^{NS}
CHH methylation	0.023 ± 0.0005	0.020 ± 0.0004	4×10^{-4}***	0.024 ± 0.001	0.019 ± 0.001	0.058^{NS}

^{NS}indicates nonsignificant variation while *, ** and *** indicate significant variation at $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively

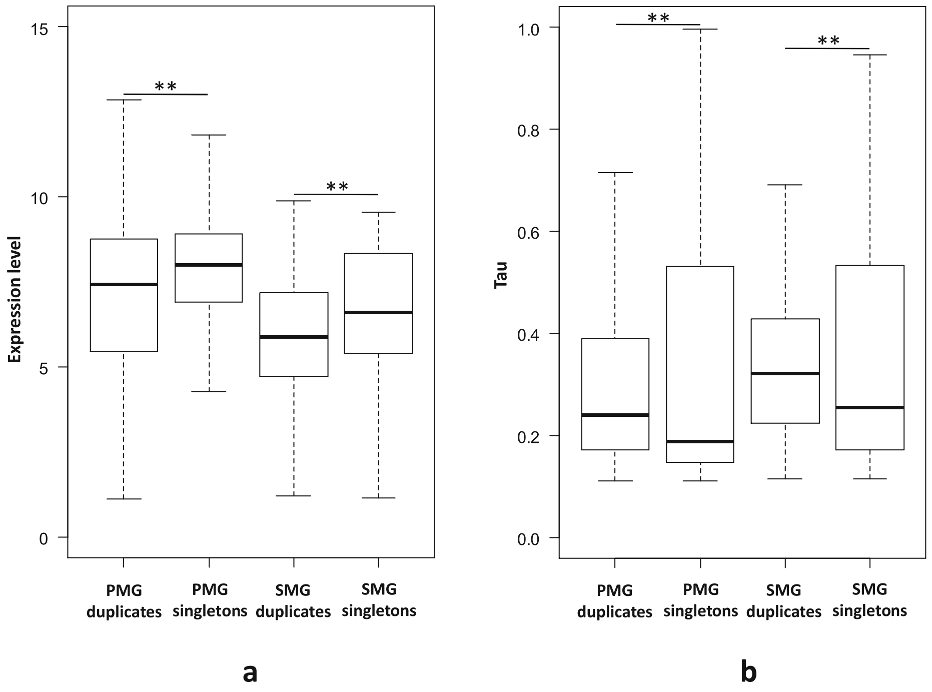

Fig. 2 **a** Boxplot showing the variation in expression level (Log_2 RMA signal intensity) among the duplicates of singletons of PMGs and SMGs. ** indicates significance at $P < 0.01$ (Mann–Whitney *U* test). **b** Boxplot showing the variation in tissue specificity (tau) among the duplicates of singletons of PMGs and SMGs. ** indicates significance at $P < 0.01$ (Mann–Whitney *U* test with Bonferroni correction)

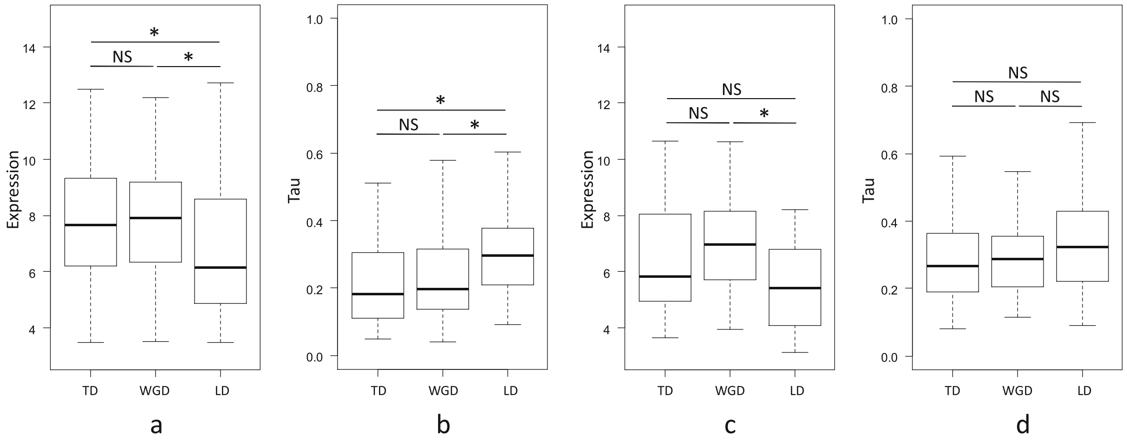

Fig. 3 Boxplots showing variation in expression level (Log_2 RMA signal intensity) and tau in different types of duplicates (duplication by TD, WGD and LD) of PMGs and SMGs. **a** Expression level variation of PMGs, **b** Tau of PMGs, **c** Expression level variation of SMGs and d) Tau of SMGs. NS and * indicate nonsignificant and significant variation (Mann–Whitney *U* test with Bonferroni correction, $P < 0.01$), respectively

more tissue specificity than the other two categories (Fig. 3b). However, in SMGs, expression level of genes originated through WGD was significantly higher than genes originated through LD (Fig. 3c). Here, TDs showed no significant difference compared to genes duplicated through LD. Regarding tissue specificity, however, no significant difference has been observed in the three categories in SMGs (Fig. 3d). It was also observed that strong negative correlations exist between paralog number and expression level in both PMG and SMG duplicates (Table 3). Regarding tissue specificity, strong positive correlation exist between paralog number and tau in both PMG and SMG duplicates (Table 3). However, in different types of duplication events, magnitude and significance level of correlation between paralog count and expression level as well as tau differed although it is evident that WGDs of both PMGs and SMGs showed negative correlation with expression level and positive correlation with tau (Table 3). All these observations showed that different types of duplication processes controlled gene expression and tissue specificity differently in PMGs and SMGs. However, in general, our results showed that duplication decreases expression level and increases tissue specificity of metabolic genes in *A. thaliana*.

Table 3 Correlation between different parameters with paralog count (*P*-values obtained with Spearman's Rank correlation test)

Parameters	PMG duplicates	SMG duplicates
Expression level	− 0.276***	− 0.372***
WGD	− 0.335***	− 0.363*
TD	− 0.067^{NS}	− 0.331^{NS}
LD	− 0.454***	− 0.174^{NS}
Tissue specificity (tau)	0.324***	0.193*
WGD	0.383***	0.381*
TD	0.133^{NS}	0.443**
LD	0.457***	− 0.119^{NS}
Gene length (bp)	− 0.080**	− 0.047^{NS}
WGD	− 0.036^{NS}	− 0.307*
TD	0.022^{NS}	0.124^{NS}
LD	− 0.373***	0.211^{NS}
Intron number	− 0.178***	− 0.040^{NS}
WGD	− 0.150***	− 0.409**
TD	− 0.106^{NS}	0.097^{NS}
LD	− 0.403***	− 0.004^{NS}
Average exon length (bp)	0.218***	0.063^{NS}
WGD	0.347***	0.405**
TD	0.255***	− 0.112^{NS}
LD	0.186***	0.107^{NS}
Transcript count	− 0.089***	− 0.109^{NS}
WGD	− 0.082*	− 0.234^{NS}
TD	− 0.071^{NS}	− 0.054^{NS}
LD	− 0.182*	− 0.034^{NS}
Essentiality	− 0.370***	0.004^{NS}
WGD	− 0.412***	− 0.346*
TD	− 0.305***	− 0.065^{NS}
LD	− 0.311***	− 0.065^{NS}

^{NS}indicates nonsignificant correlation while *, ** and *** indicate significant correlation at $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively

Effect of duplication on gene architecture of PMGs and SMGs

Genes of singletons are significantly longer than duplicates in both PMGs and SMGs (Table 2). Gene length of WGDs and TDs of PMGs are significantly longer than SMGs while LDs of the both groups are similar in length (Table 1). Paralog count was significantly negatively correlated with gene length in PMGs (although only LDs contribute to this correlation), while SMGs showed no significant correlation between paralog count and gene length (although WGDs showed some correlation). Although, coding sequence length was similar between duplicates and singletons of PMGs and SMGs ($p > 0.05$), intron number was significantly higher in singletons than duplicates in both PMGs and SMGs (Table 2). This shows that the difference between gene length of singletons and duplicates are contributed by introns. It has been noteworthy that all types of duplicates of SMGs possess lesser introns than PMGs (Table 1). Although, coding sequence of duplicates and singletons do not differ, average exon length of duplicates is higher than singletons in both PMGs and SMGs (Table 2). Thus, singletons have more introns with shorter exons while duplicates have fewer introns with longer exons in both PMGs and SMGs. Intron number showed a strong negative correlation with paralog count in PMG duplicates and a positive correlation with average exon length while this trend is absent in SMG duplicates (Table 3). As increase in intron number facilitates alternative splicing, thus, it seems that transcript count per gene should be more in intron rich genes. Here, singletons showed significantly more transcripts than duplicates in SMGs, but not in PMGs. PMG and SMG duplicates as well as PMG singletons showed significant correlation between intron number and transcript count ($p < 0.001$). Thus, increased gene length is due to increased intron number which, especially in SMGs, helps in alternative splicing. Moreover, transcript count was negatively correlated with paralog count in PMG duplicates (also in WGDs and LDs, but not in TDs) but not in SMG duplicates (also in WGDs, TDs and LDs) (Table 3). This shows that generally in PMGs, diversity of proteins is mainly achieved by alternative splicing while in SMGs, diversity is generally achieved by gene duplication.

Relationship of gene expression patterns with gene architecture in duplicates and singletons

Although PMGs are longer and showed more expression compared to SMGs, weak correlation was found between gene length and expression level in PMG (Spearman's $\rho = 0.056$, $p < 0.05$) while SMG showed no significant correlation. However, tissue specificity showed no correlation with gene length in both these groups. However, this picture changed completely when correlations were performed in duplicates and singletons separately. PMG duplicates showed positive correlation between gene length and expression level (Spearman's $\rho = 0.125$, $p = 1.6 \times 10^{-4}$) while PMG singletons showed negative correlation (Spearman's $\rho = - 0.162$, $p = 1.8 \times 10^{-3}$). SMG duplicates also showed positive correlation (Spearman's $\rho = 0.177$, $p < 0.05$) while SMG singletons showed no significant correlation. Thus, PMG singletons behave like metazoan housekeeping genes where gene length is inversely proportional to expression level. However, duplicates in both PMGs and SMGs showed opposite trend. Regarding tissue specificity, PMG duplicates showed negative correlation between gene length and tissue specificity (Spearman's $\rho = -\ 0.096$, $p < 0.01$) while PMG singletons showed positive correlation (Spearman's $\rho = 0.134$, $p < 0.05$). No correlation was found in SMGs. However, in both PMG and SMG duplicates and singletons, strong negative correlation was found between expression level and tissue specificity ($p < 0.001$).

Effect of duplication on multifunctionality of metabolic genes

As expression is related with the functional aspects of genes, we have also looked into the relationship of duplication and multifunctionality of the metabolic genes. Multifunctionality of PMGs and SMGs has been expressed as number of GO terms associated with a particular gene. We have considered both 'molecular function' as well as 'biological process' to assess multifunctionality. There was no significant difference between the duplicates and singletons of PMGs regarding the number of GO terms in 'molecular function', although it was found that WGDs, TDs and LDs of SMG duplicates take part in significantly more molecular functions than WGDs, TDs and LDs of PMGs, respectively (Table 1). SMG singletons were also found to be involved in more molecular functions than PMG singletons ($p < 0.05$). Thus, in general, it was observed that SMGs perform more molecular functions than PMGs and that reflects in all types of duplicates as well as singletons. However, paralog count was not correlated with 'molecular functions' showing duplication has no effect on alterations in molecular functions in PMGs as well as SMGs. Regarding 'biological process', PMG showed no significant difference between singletons and duplicates while in SMGs, significant differences were found (Table 2). However, it was observed that in both PMG and SMG duplicates, paralog number have a strong negative correlation with the number of biological process they are associated with (Spearman's $\rho_{\text{PMG duplicates}} = -\ 0.157$; $p = 10^{-6}$, Spearman's $\rho_{\text{SMG duplicates}} = -\ 0.289$; $p = 9.52 \times 10^{-5}$). However, transcript count showed no significant correlation with the number of biological process showing gene architecture played no role here. Moreover, it was also observed that regarding biological process, there is a significant difference between the duplicates of PMGs as well as SMGs ($p = 9.42 \times 10^{-8}$, Mann-Whitney *U* test with Bonferroni correction) while in singletons, no significant difference was found. This showed that PMG duplicates take part in more processes than SMG duplicates, a trend, absent in singletons.

It was also observed that expression level was significantly correlated with number of GO terms in 'biological process' in PMG duplicates (Spearman's $\rho = 0.455$; $p = 10^{-6}$), PMG singletons (Spearman's $\rho = 0.442$; $p = 10^{-6}$) and SMG singletons (Spearman's $\rho = 0.526$; $p = 0.02$), but not in SMG duplicates ($p > 0.05$). Tau (tissue specificity) was significantly negatively correlated with number of GO terms in 'biological process' in PMG duplicates (Spearman's $\rho = -\ 0.334$; $p = 10^{-6}$), PMG singletons (Spearman's $\rho = -\ 0.280$; $p = 10^{-6}$) and SMG singletons (Spearman's $\rho = -\ 0.486$; $p = 0.03$), but not in SMG duplicates ($p > 0.05$). So, SMG duplications makes them 'specialized' in terms of the biological process they are involved in, but that is not dependent on the paralog count.

Effect of duplication on cellular compartmentalization of metabolic genes

In addition to the molecular function and biological process, the function of a gene is also associated with the subcellular localization of its encoded protein (Acharya and Ghosh 2016). To assess this, we have considered the cellular compartmentalization of their encoded proteins in addition to tissue specificity. The localization of the encoded proteins was obtained by using the Gene Ontology (GO) terms under the GO domain 'Cellular Component' against the gene identifier. In both PMGs and SMGs, duplicates are expressed in less cellular components than singletons (Table 2). However, TDs of PMGs showed more 'Cellular Components' than SMGs, a trend absent in WGDs and LDs (Table 1). Thus, WGDs and local duplicates of PMGs and SMGs expressed in more or less similar number of cellular Components. For local duplicates, this may be due to the fact that these local duplicates took place in *Arabidopsis* by more recent duplication process, both appeared probably by adaptive pressure in both PMGs and SMGs. However, WGDs should be studied further regarding this parameter. Paralog count was found to be significantly correlated with number of GO term 'Cellular Components' in both PMG duplicates (Spearman's $\rho = -\ 0.067$, $p = 0.015$) and in SMG duplicates (Spearman's $\rho = -\ 0.357$, $p = 10^{-6}$). Here, it is noteworthy that the correlation is far stronger in SMGs than PMGs. Thus, in SMGs, duplication actually profoundly compartmentalized the expression of SMGs.

Regarding all types of GO terms, the most frequent GO terms in all the categories i.e. 'molecular function', 'biological process' and 'cellular compartment' showed that some GO terms overlap in PMG duplicates and singletons as well as in SMG duplicates. Notably, in all three categories, the ten most frequent GO terms for SMG singletons were almost exclusive (Fig. 4). This showed that SMG singletons perform some unique functions and expressed in some unique cellular compartments.

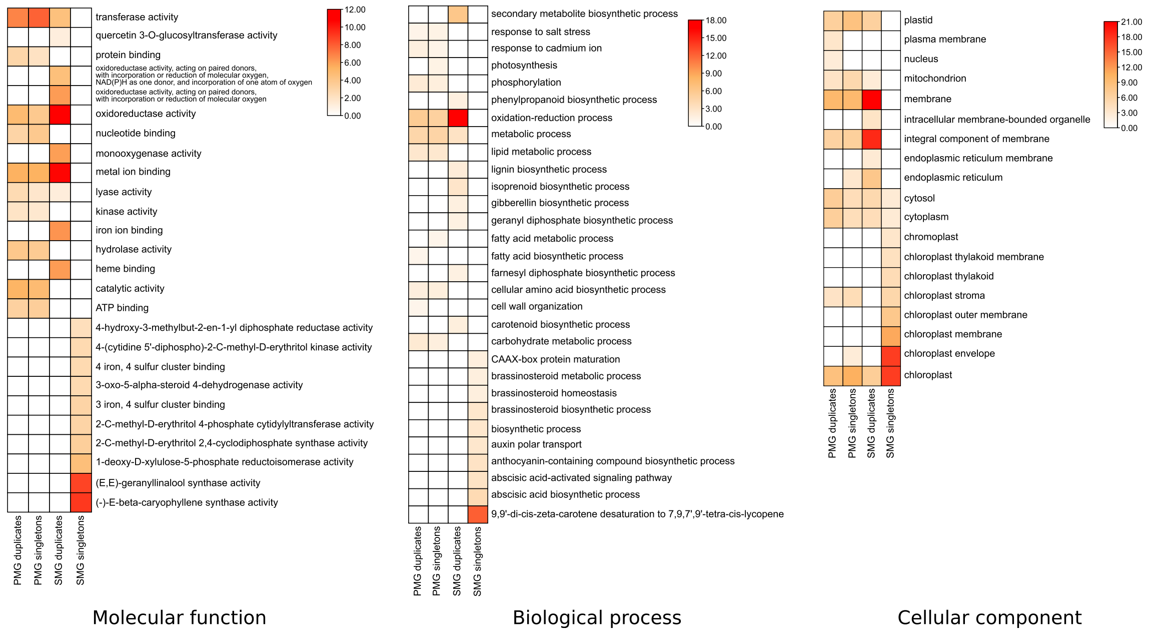

Fig. 4 Heatmaps showing the percentages of ten most abundant GO terms in each of the four categories of genes (darker shades represent increasing value). Note uniqueness of molecular function, biological process and cellular components in SMG singletons

Role of duplication in altering catalytic properties

Since, our dataset mostly comprises of metabolic enzymes, we decided to analyze whether duplicability somehow alters catalytic properties of these enzymes. To study this, we chose product diversity of the PMG and SMG duplicates as a measure of enzyme's functionality. We simply counted the number of unique products yielded by the enzyme in a given reaction. We have observed that in SMGs, a significant difference was observed between duplicates and singletons regarding product diversity. However, in PMGs, this trend was absent (Table 2). We have also observed that in the SMG duplicates, product diversity correlated negatively with paralog number (Spearman's $\rho = -0.310$, $P < 10^{-6}$). However, product diversity did not correlate with duplicability in PMGs (Spearman's $\rho = -0.030$, $P = 0.366$).

Relation of duplication and gene essentiality in PMGs and SMGs

It would be interesting to investigate whether this duplication pattern influence the overall importance of the genes concerning organism's fitness. It is a pertinent question that how far duplication changes the essentiality of the metabolic genes. It has been observed that in both PMGs and SMGs, singletons are more essential than duplicates (Table 2). It was also observed that WGD and TD duplicates are more essential in PMG than SMG duplicates, while LDs showed no such difference (Table 1). In PMG duplicates, the essentiality score was found to be significantly negatively correlated with paralog count, a trend absent in SMGs. Among PMGs, all three types of duplicates (WGDs, LDs and TDs) showed significant negative correlation between essentiality score and paralog count, while in SMGs, only WGD duplicates showed this correlation with a significance level much lower than that of PMG (Table 3). Further, we have observed that in PMGs, essentiality was significantly correlated with both expression level (PMG_duplicates:Spearman's $\rho_{\text{PMGduplicates}} = 0.381$, $P = 1 \times 10^{-6}$; Spearman's $\rho_{\text{PMGsingletons}} = 0.302$, $P = 1.24 \times 10^{-6}$) and tissue specificity (Spearman's $\rho_{\text{PMGduplicates}} = -0.525$, $P = 1 \times 10^{-6}$; Spearman's $\rho_{\text{PMGsingletons}} = -0.340$, $P = 1 \times 10^{-6}$) but only with tissue specificity in SMGs (Spearman's $\rho_{\text{SMGduplicates}} = -0.260$, $P = 0.005$; Spearman's $\rho_{\text{SMGsingletons}}- 0.85$, $P = 0.004$). All these results suggest that SMGs became essential based on their tissue specificity while PMGs became essential based on their expression as well as tissue specificity. Moreover, in SMG duplicates, essentiality was significantly negatively correlated with product diversity (Spearman $\rho = -0.211$, $P < 0.039$). In PMGs, these two were not correlated. This is probably due to the fact that genes of the secondary metabolism are often promiscuous (Weng and Noel 2012).

Epigenetic modification of metabolic genes following duplication

We have observed the effect of all three types (CG, CHG and CHH) of methylation in both promoter and gene-body on expressed traits of metabolic genes. It was observed that in PMGs, singleton genes are more methylated in their promoters for CG and CHG methylation, while regarding gene body methylation, PMG singletons were more CG methylated but less CHG and CHH methylated (Table 2). However, the picture is more complicated in SMGs. Here, gene body of singletons are more CG methylated and less CHH methylated and promoters of singletons are less CHG methylated but slightly more CHH methylated (Table 2). Notably, WGD, TD as well as LD duplicates of PMGs and SMGs showed no significant difference regarding promoter methylation (Table 4). However, regarding gene body methylation, WGDs and TDs of PMGs are more CG methylated than WGDs and TDs of SMGs, respectively, while LDs showed no such difference. WGDs of PMGs, however, showed less CHG methylation than WGDs of SMGs. CHH methylation showed no difference between different types of duplicates of PMGs and SMGs (Table 4).

Duplicates of PMGs showed strong negative correlation of all types of promoter and CG gene body methylation with paralog count, while gene body CHG and CHH methylation showed positive correlation. SMGs, however, showed no correlation at all (Table 5). Moreover, paralog count was significantly negatively correlated with CG gene body methylation in all types of duplicates (WGDs, TDs and LDs) of PMGs. However, regarding other types of methylation, significance varied between different types of duplicates. This showed that gene body CG methylation has been strongly reduced after all types of duplication events in PMGs. Thus, gene regulation by methylation (especially gene body CG methylation) is a characteristic of PMGs while methylation of SMGs is not affected by duplication. Notably, transcript count also showed strong positive correlation with gene body GC methylation, while strong negative correlation was observed with gene body CHG and CHH methylation. Expression level showed strong positive correlation with gene body CG methylation while tau showed strong negative correlation. However, in PMG singletons, no correlation was found between gene body CG methylation and expression as well as tau. All these results showed that after duplication of PMGs,

Table 4 Difference between different types of methylation of different types of duplicated genes in PMGs and SMGs (*p*-values obtained with Mann–Whitney *U* test with Bonferroni correction)

Parameters	PMGs (Mean ± SE)	SMGs (Mean ± SE)	*P*-value
Promoter methylation			
CG methylation			
WGD	0.088 ± 0.006	0.065 ± 0.019	1.0^{NS}
TD	0.121 ± 0.012	0.100 ± 0.027	0.34^{NS}
LD	0.125 ± 0.018	0.079 ± 0.015	0.88^{NS}
CHG methylation			
WGD	0.045 ± 0.004	0.031 ± 0.010	1.0^{NS}
TD	0.064 ± 0.007	0.056 ± 0.018	0.18^{NS}
LD	0.066 ± 0.012	0.033 ± 0.008	0.42^{NS}
CHH methylation			
WGD	0.032 ± 0.002	0.022 ± 0.006	0.44^{NS}
TD	0.038 ± 0.004	0.034 ± 0.008	0.72^{NS}
LD	0.042 ± 0.006	0.022 ± 0.003	1.0^{NS}
Gene body methylation			
CG methylation			
WGD	0.132 ± 0.005	0.081 ± 0.013	0.018*
TD	0.163 ± 0.009	0.061 ± 0.009	1.64×10^{-6}**
LD	0.108 ± 0.012	0.078 ± 0.012	0.38^{NS}
CHG methylation			
WGD	0.022 ± 0.001	0.029 ± 0.003	0.028*
TD	0.034 ± 0.004	0.022 ± 0.001	1.0^{NS}
LD	0.038 ± 0.008	0.028 ± 0.006	0.74^{NS}
CHH methylation			
WGD	0.022 ± 0.0004	0.025 ± 0.002	0.22^{NS}
TD	0.025 ± 0.001	0.025 ± 0.002	0.68^{NS}
LD	0.028 ± 0.003	0.023 ± 0.002	0.76^{NS}

^{NS}indicates nonsignificant variation while * and ** indicate significant variation at $P < 0.05$ and $P < 0.001$, respectively

Table 5 Correlation between different types of methylation patterns with paralog count, transcript count, expression level and tau as well as gene architectural parameters (*P*-values obtained with Spearman's Rank correlation test)

Parameters	PMG duplicatess						SMG duplicates					
	Promoter			Gene body			Promoter			Gene body		
	CG	CHG	CHH	CG	CHG	CHH	CG	CHG	CHH	CG	CHG	CHH
Paralog count	–0.084**	–0.084**	–0.091***	–0.213***	0.114***	0.100***	0.013^{NS}	0.060^{NS}	0.049^{NS}	–0.011^{NS}	–0.072^{NS}	–0.051^{NS}
WGD	–0.124**	–0.097*	–0.092*	–0.211***	0.053^{NS}	0.032^{NS}	–0.025^{NS}	0.009^{NS}	–0.024^{NS}	–0.322*	0.178^{NS}	0.207^{NS}
TD	0.010^{NS}	–0.016^{NS}	–0.059^{NS}	–0.118*	0.187**	0.194**	0.012^{NS}	0.083^{NS}	–0.227^{NS}	–0.081^{NS}	–0.039^{NS}	–0.089^{NS}
LD	–0.107^{NS}	–0.157^{NS}	–0.112^{NS}	–0.185*	0.114^{NS}	0.116^{NS}	0.054^{NS}	0.262*	0.230*	0.157^{NS}	–0.056^{NS}	0.013^{NS}
Transcript count	0.005^{NS}	–0.035^{NS}	–0.015^{NS}	0.110***	–0.086**	–0.109***	–0.072^{NS}	0.0003^{NS}	–0.102^{NS}	0.056^{NS}	–0.103^{NS}	–0.115^{NS}
WGD	0.061^{NS}	0.009^{NS}	0.018^{NS}	0.095*	–0.015^{NS}	–0.074^{NS}	–0.299*	–0.122^{NS}	–0.228^{NS}	0.185^{NS}	–0.180^{NS}	–0.191^{NS}
TD	–0.048^{NS}	–0.042^{NS}	0.014^{NS}	0.056^{NS}	–0.245***	–0.202***	0.043^{NS}	0.082^{NS}	–0.023^{NS}	–0.072^{NS}	–0.258^{NS}	–0.103^{NS}
LD	0.051^{NS}	0.026^{NS}	0.014^{NS}	0.187*	–0.072^{NS}	–0.039^{NS}	0.090^{NS}	0.076^{NS}	0.058^{NS}	0.193^{NS}	0.078^{NS}	0.001^{NS}
Expression level	0.00005^{NS}	0.050^{NS}	0.077*	0.193***	–0.048^{NS}	–0.038^{NS}	–0.005^{NS}	0.048^{NS}	0.079^{NS}	0.008^{NS}	–0.078^{NS}	–0.036^{NS}
WGD	0.001^{NS}	0.048^{NS}	0.062^{NS}	0.179***	–0.037^{NS}	–0.077^{NS}	0.220^{NS}	0.003^{NS}	0.268^{NS}	–0.033^{NS}	–0.402*	–0.374*
TD	–0.019^{NS}	0.022^{NS}	0.048^{NS}	0.195**	–0.092^{NS}	–0.011^{NS}	–0.065^{NS}	0.077^{NS}	0.100^{NS}	0.389*	0.169^{NS}	0.071^{NS}
LD	–0.044^{NS}	0.089^{NS}	0.110^{NS}	0.136^{NS}	–0.052^{NS}	0.020^{NS}	–0.078^{NS}	–0.046^{NS}	0.011^{NS}	–0.334*	–0.121^{NS}	0.010^{NS}
Tissue specificity	–0.042^{NS}	–0.077*	–0.109**	–0.284***	0.108**	0.105**	–0.152^{NS}	–0.210*	–0.212*	–0.187*	0.013^{NS}	–0.029^{NS}
WGD	–0.055^{NS}	–0.065^{NS}	–0.112*	–0.256***	0.075^{NS}	0.099^{NS}	–0.405*	–0.200^{NS}	–0.340*	–0.220^{NS}	0.222^{NS}	0.275^{NS}
TD	–0.023^{NS}	–0.078^{NS}	–0.095^{NS}	–0.277***	0.169*	0.096^{NS}	–0.001^{NS}	–0.122^{NS}	–0.324^{NS}	–0.490**	0–111^{NS}	–0.062^{NS}
LD	0.014^{NS}	–0.061^{NS}	–0.034^{NS}	–0.159^{NS}	0.108^{NS}	0.014^{NS}	–0.185^{NS}	–.0359*	–0.155^{NS}	0.002^{NS}	0.026^{NS}	–0.153^{NS}

^{NS}indicates nonsignificant correlation while *, ** and *** indicate significant correlation at $P < 0.05$, $P < 0.01$ and $P < 0.001$, respectively methylation, especially gene body CG methylation, controls the expression and tissue specificity while in PMG singletons, expression and tau is not correlated with gene body CG methylation. In SMGs, on the other hand, expression is not, as a whole, dependent on methylation. Tissue specificity in SMG duplicates is, however, to some extent, correlated with some methylation types (Table 5).

As our work is mainly focused on the relationship between gene architecture and expression pattern, we have studied the relationship between methylation and gene length as well as average exon length. It was found that gene body CG methylation is strongly positively correlated with gene length in PMG duplicates (Spearman's $\rho = 0.432$, $p = 10^{-6}$), PMG singletons (Spearman's $\rho = 0.558$, $p = 10^{-6}$) and SMG singletons (Spearman's $\rho = 0.434$, $p = 0.001$), but not in SMG duplicates. Moreover, methylation is strongly negatively correlated with average exon length in PMG duplicates (Spearman's $\rho = -0.188$, $p = 10^{-6}$), PMG singletons (Spearman's $\rho = -0.144$, $p = 0.002$) and SMG singletons (Spearman's $\rho = -0.473$, $p = 0.0005$), but not in SMG duplicates. Promoter CG methylation, on the other hand, is negatively correlated with average exon length in PMG duplicates (Spearman's $\rho = -0.068$, $p = 0.014$) and singletons (Spearman's $\rho = -0.094$, $p = 0.037$), but not in SMGs. In PMG duplicates, promoter CG methylation is significantly correlated with expression level (Spearman's $\rho = -0.097$, $p = 0.03$).

Discussion

Metabolism is a complex system as enzymes are part of organized pathways and this makes metabolism an attractive platform to study evolutionary processes. Moreover, it has been shown that in plants, metabolic novelty has been created by gene and genome duplication (Moghe and Last 2015). New branches of metabolism continuously arose throughout land-plant evolution (Weng et al. 2012). Gene duplication gave rise to generally conserved primary metabolism and lineage-specific secondary metabolism (Moore et al. 2019). As Kliebenstein (2008) stated that at least in metabolic pathways, the maintenance and evolution of duplicated genes is determined by the biology of the pathway, we here concentrated on primary and secondary metabolic pathways of *A. thaliana* for a deeper insight into the matter. As we have previously shown that although primary metabolic pathway genes (PMGs) are highly expressed and more conserved than secondary metabolic pathway genes (SMGs) during their evolution, they are longer than SMGs (Mukherjee et al. 2016). As this observation contradicts with the selection hypothesis (Ellegren and Sheldon 2008) as well as the genomic design hypothesis (Carmel and Koonin 2009), we have tried to study the relationship between gene architecture and gene expression of metabolic genes in the context of gene and genome duplication. Our study showed that in both PMGs and SMGs, singletons are longer and with higher expression than duplicates in both PMGs and SMGs. However, our results showed that this length variation is due to higher intron number in singletons than duplicates as no significant difference was observed between singletons and duplicates regarding coding sequence length. This indicates that against which factor introns are selected in duplicated genes. As introns make an enormous contribution to the genome size, it is reasonable that selective force for genome reduction acted as a selective force against the accumulation of introns (Fawcett et al. 2012; Yang et al. 2013) in *A. thaliana* as this species has a smaller genome size (Bekaert et al. 2011). Moreover, *A. thaliana* lost half of its genome and lost more introns than *A. lyrata* after their divergence (Yang et al. 2013). According to mutational-hazard hypothesis (Lynch 2006, 2007; Lynch et al. 2006), more noncoding DNA are more likely to accumulate deleterious mutations and thus, genes with higher mutation rates are prone to intron loss. Indeed, our analysis showed that PMG duplicates are more conserved than SMG duplicates during their evolution and thus, SMG duplicates showed lesser number of introns than PMG duplicates (also in WGDs, TDs and LDs), as well as singletons which are also much conserved. Moreover, our analysis showed that LDs of PMGs are more conserved than LDs of SMGs and thus, lower introns of SMG LDs can be explained by mutational-hazard hypothesis. However, evolutionary rates of WGDs and TDs of PMG and SMG duplicates are similar although SMGs of these two categories contain significantly less introns. So, this cannot be explained by mutational-hazard hypothesis.

It is known that introns do confer some benefits either by expanding protein diversity through alternative splicing (Kalsotra and Cooper 2011). As PMG duplicates have more introns than SMG duplicates, these factors might have been the selective forces for retaining some introns. To examine this, we have performed correlation of transcript count with intron number. The fact that intron number of PMG and SMG duplicates as well as PMG singletons are positively correlated with transcript count showed that alternative splicing are a selective force for retaining some introns. However, PMG duplicates showed that paralog count is significantly negatively correlated with intron number while SMG duplicates do not show this trend. It has also been observed that transcript count is highly negatively correlated with paralog count in PMG duplicates (except TDs) while this trend is absent in SMGs. Kopelman et al. (2005) reported that fewer genes of larger gene families tend to be affected by alternative splicing compared to singletons. As many enzyme families involved in secondary metabolism arose through tandem gene duplication (Moore et al. 2019), our observation that paralog number is not correlated with transcript count in SMG duplicates supports the view of Kopelman et al. (2005). Also, d_N/d_S of SMG duplicates are significantly correlated with paralog number showing that duplication of SMGs have adaptive advantage. Indeed, Chae et al. (2014) suggested that in plants, unlike primary metabolism, secondary metabolism is under selection which drives the expansion of gene families coding for specific enzymic processes. Our analysis also showed that average exon length i.e. exons are significantly longer in SMG duplicates than PMG duplicates but singletons showed similar average exon lengths. This supports the view of Wang et al. (2013) who showed that average exon length is positively correlated with evolutionary rate in *A. thaliana*. Thus, as a whole, PMG duplicates increased protein diversity by alternative splicing while SMG duplicates relied more on gene duplication. Now, the question arises, what is the reason behind this discrepancy between PMGs and SMGs. In *A. thaliana*, it has been shown that unlike animals, majority of alternatively-spliced transcripts are not translated into proteins or low in translation (Yu et al. 2016). The authors suggested that many of them are subject to NMD (nonsense-mediated mRNA decay). NMD actually downregulates the expression of a gene by shunting a portion of its pre-mRNAs into a decay pathway (Lareau and Brenner 2015). Thus, while alternative splicing does increase protein diversity (Syed et al. 2012), it has a role in gene expression regulation (Smith et al. 1989). Thus, in PMGs, alternative splicing may regulate their expression in addition to increasing the protein diversity to some extent. On the other hand, gene duplication with higher evolutionary rates is a mechanism by which novel functions can arise in secondary metabolism. It has been shown that a few mutations can readily increase promiscuous activity of an enzyme by 10–1000 fold (Schmidt et al. 2003; Varadarajan et al. 2005). Thus, duplication followed by high evolutionary rate proved to be beneficial for increasing the chemical weapons by secondary metabolism. Indeed, in our analysis, we have shown that SMG duplicates have significant product diversity than SMG singletons. Moreover, these compounds do not affect the fitness of the plant as they are not immediately required for the survival of the plant (Weng and Noel 2012). Actually, while secondary metabolism produces enormous chemical diversity, they represent a small fraction of the total mass of plant tissues (Shih and Morgan 2020). However, no significant difference was observed between duplicates and singletons in PMGs regarding product diversity. This is probably due to the fact that the high flux nature of primary metabolism (Almass et al. 2004) imposes a major selective pressure on the evolution of PMGs (Nam et al. 2012). Thus, promiscuity of these enzymes can negatively impact their catalytic efficiency (Weng and Noel 2012). As these enzymes perform conserved core metabolic functions (Bar-Even et al. 2011), reactions are highly specific with low mechanistic elasticity (Weng et al. 2012). Moreover, primary and secondary metabolic pathways are also different in their pathway structure. The dosage balance hypothesis states that in eukaryotes, selection acts against duplications of central networks genes as this will imbalance the stoichiometry of the network (Conant et al. 2014). However, in context of WGD, this situation is reversed as the loss of the second copy of a particular duplicated gene would introduce imbalance in relation to the other duplicates (Bekaert et al. 2011). As SMGs have limited network-wide interconnections, LDs are prevalent in secondary metabolism where the increased dosage would be beneficial which also supports the view of Hudson et al. (2011). However, as primary metabolism is highly interconnected, WGD is a potential route for increased flux. This also explains our results which showed that number of LDs and TDs are much closer to number of WGDs in PMGs but not in SMGs. This happened since relative expression level is tuned in these high flux primary metabolic enzymes in view of the whole pathway maintaining relative dosage balance after a WGD, supporting the view of Bekaert et al. (2011).

The expression heterogeneity of PMGs and SMGs are also the result of the selection pressure that acts on them. As PMGs are highly interconnected and multiple layers of regulation act on them, their expression must be fine tuned. It is well documented that introns possess some regulatory elements which regulate gene expression (Parenteau et al. 2011; Rose et al. 2008). Moreover, as we have discussed, NMD also help in expression regulation. Thus, higher number of introns gives some advantage to PMGs relative to SMGs. It certainly increases the burden of carrying surplus DNA, but as Ren et al.(2006) showed, introns in *A. thaliana* are much smaller (average length is 152 bp) than humans (average length is $\sim$ 5.5 kb). Thus, it is obvious that at least in PMGs, selection force for the retention of introns is more than the selection force against their retention. This is the reason behind the fact that WGDs and TDs of PMGs have significantly more introns than WGDs and TDs of SMGs, although they showed similar evolutionary rates. These introns of PMGs must have played a role in fine tuning the expression of their genes. In fact, the length variation of these metabolic genes has variable effects on the gene expression patterns. It was found that gene length is negatively correlated with expression level in PMG singletons and positively correlated with both PMG and SMG duplicates. Thus, PMG singletons behave like metazoan housekeeping genes and follows the selection hypothesis (Ellegren and Sheldon 2008) as well as the genomic design hypothesis (Carmel and Koonin 2009). However, this trend has not been observed in SMG singletons. In fact, other genic characters like intron number, transcript count or average exon length was found to be correlated with expression level or tissue specificity in SMG singletons. This is probably because they are highly conserved (as d_N/d_S values of SMG singletons are not significantly different to PMG singletons as well as PMG duplicates) and the end products are of lower flux than PMGs. The highly conserved nature is probably due to their similar essentiality score as of PMG singletons. Moreover, their expression levels are significantly lower than PMG singletons ($p < 0.05$). As genes with lower expression level are not transcriptionally demanding, selection for miniaturization becomes economically irrelevant (Woody and Shoemaker 2011). So, probably neither selection hypothesis nor mutational bias hypothesis shaped their correlation between genic architecture and expression. Thus, the type of metabolic genes as well as the duplication status determines the expression pattern of the genes and expression pattern is not always dependent on gene architecture. This again supports the view of Kliebenstein (2008) who stated that at least in metabolic pathways, the maintenance and evolution of duplicated genes is determined by the biology of the pathway.

Expression level of both PMG and SMG duplicates was found to be lower in singletons in our study. Qian et al. (2010) proposed that the expression is reduced in daughter genes to prevent the loss of any one of the copies as this requires the retention of functions of the duplicated copies. We also propose that these duplicate copies also become more tissue specific without changing their molecular function (in PMGs and SMGs) as well as biological processes they are involved in (PMGs). In SMG duplicates, significantly less GO term “biological processes” were found than SMG singletons. As these genes are involved in environmental interaction, we propose that streamlining of their involvement after gene duplication occurred after duplication. However, in PMGs, no significant change in molecular functions and involvement in biological processes occurred as they perform core metabolic activities. This is also supported by the fact that they are as conserved during their evolution as their singleton counterparts. Qian et al. (2010) proposed that expression reduction after duplication facilitates their long-term maintenance without changing their functional redundancy. However, after duplication, both PMG and SMG duplicates not only became tissue specific, but also more compartmentalized. So, change in location of expression is a major fate of the duplicates of metabolic genes. As intron number of PMG singletons are negatively correlated, this again supports the selection hypothesis. However, this does not explain the positive correlation between expression level and gene length. It was also observed that expression level is also positively correlated with intron number in PMG duplicates. Vinogradov (2004) proposed that broadly expressed genes require simple regulation and thus, fewer regulatory elements. However, Yang (2009) proposed that broadly expressed genes in plants contain longer non-coding sequences as these may be needed for their complex regulations. Woody and Shoemaker (2011) proposed that in lowly expressed genes with higher tissue specificity, introns and intergenic regions are increased as it has been hypothesized that intron and intergenic regions are involved in chromatin-mediated suppression as well as higher-order regulation. As PMG duplicates have low expression and with higher tissue specificity compared to PMG singletons, their intron number is positively correlated with their expression level. However, this is not applicable in SMG duplicates. In fact, both expression level and tau were not significantly correlated with gene length as well intron number in SMG duplicates and singletons.

From the above discussion, it is clear that gene regulation is a crucial factor for gene expression of plant genes. To better understand that, we have studied the methylation pattern of these genes for their role of epigenetic factors in plant metabolic genes. Cytosine DNA methylation is an epigenetic mark important in various gene regulatory systems including expression of endogenous genes (Feng et al. 2010). According to the sequence context of the cytosines, three types of DNA methylation are known: CG, CHG and CHH (H = A, C or T) (Law and Jacobsen 2010). In plants, non-CG methylation (CHG and CHH) are abundant compared to mammals and this can partly be explained by plant specific CMT (*CHROMOMETHYLASE*) genes (Stroud et al. 2014). However, in plants, level of methylation is highest at CG sites, medium at CHG and lowest at CHH sites (Feng et al. 2010). Singletons of both PMGs and SMGs are with shorter exons, more introns and with more gene body CG methylation than duplicates indicating that introns are selected in singletons for epigenetic regulations. Divergence of DNA methylation (primarily CG) is common for duplicated copies of given gene pair which ultimately is correlated with differential expression (Wang et al. 2017). Gene body methylation has been proposed to suppress spurious transcription from cryptic promoters that might otherwise interfere with gene expression (Tran et al. 2005). Moreover, gene body CG methylation is more conserved than methylation of other genomic regions in plants (Takuno and Gaut 2013; vonHoldt et al. 2012). Paralog count was negatively correlated with promoter and gene body CG methylation in PMGs showing that duplication decreases these types of methylations. More strikingly, paralog count, transcript count, expression level were not, in general, correlated with methylation pattern of SMGs. As a whole, gene body CG methylation plays a role in regulation of PMGs. Moreover, expression level was positively and tissue specificity was negatively correlated with gene body CG methylation pattern in PMGs showing that the effect of epigenetic regulation on PMG duplicates is greater than SMG duplicates. The controversial functional role of the gene body methylation has been reported earlier (Lee et al. 2010; Su et al. 2011). Although our study showed highly heterogeneous pattern of both promoter and gene methylation, it is clear that PMG duplicates are more epigenetically controlled than SMG duplicates. This again shows the highly regulated nature of PMGs than SMGs after duplication, which ultimately resulted in difference in their genic architecture as well as expression patterns.

Conclusion

In conclusion, our study suggests that expression heterogeneity and gene structure variation of PMGs and SMGs in *Arabidopsis thaliana* are partly results of different duplication events. WGDs are prevalent in PMGs while LDs are prevalent in SMGs. It has been found that PMGs are more interconnected as part of a broader metabolic network. This, along with their high flux nature shaped their genic architecture as well as expression pattern. On the other hand, as SMGs, in general, take part in ecological interactions, they demand rapid expression. This, along with their low flux nature and higher enzymic promiscuity shaped their evolution of genic structure and expression pattern. The results have direct ecological and agronomic significance, too. For example, in Brassicales, defence against herbivores involves specialized metabolites such as the glucosinolates (Demain and Fang 2000; Halkier and Gershenzon 2006) and the evolution of the glucosinolate pathway involves the retention of the core Trp pathway gene duplicates derived from the β-WGD event (Edgeret al. 2015). Similarly, Brenchley et al. (2012) hypothesized that polyploidization has contributed to the expansion of wheat storage protein genes. It is also opined that selection in *Brassica napus* has led to the preservation of duplicate oil biosynthesis genes (Chalhoub et al. 2014). Our work on *Arabidopsis thaliana* may serve as a foundation for genome-wide study of the role of duplication on metabolism in plants as the plant is a wild species and devoid of any anthropogenic activities. Further studies on different crop plants can be compared with this study to decode the relationship of duplication and evolution of plant metabolism with human intervention as a factor.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12298-022-01188-2.

Acknowledgements DM is thankful to the Department of Biotechnology, Govt. of India for financial support. DS and DA are thankful to Bose Institute, Kolkata for financial support

Author contributions DM, DS and AM conceptualized the work. DM and DS carried out the experiments. Manuscript is prepared and edited by DM, DS, AM, DA and TCG.

Data availability All the data are tabulated in this manuscript.

Declarations

Conflict of interest All authors unanimously declare of no conflict of interest over the submission of the manuscript.

References

Acharya D, Ghosh TC (2016) Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution. BMC Genomics 17:71

Bekaert M, Edger PP, Pires JC, Conant GC (2011) Two-phase resolution of polyploidy in the *Arabidopsis* metabolic network gives rise to relative and absolute dosage constraints. Plant Cell 23:1719–1728

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679–1691

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438

Brenchley R, Spannagl M, Pfeifer M, Barker GLA, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710

Carmel L, Koonin EV (2009) A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. Genome Biol Evol 1:382–390

Chae L, Kim T, Nilo-Poyanco R, Rhee SY (2014) Genomic signatures of specialized metabolism in plants. Science 344:510–513

Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B et al (2014) Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 345:950–953

Conant GC, Birchler JA, Pires JC (2014) Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Curr Opin Plant Biol 19:91–98

Crow KD, Wagner GP (2006) What is the role of genome duplication in the evolution of complexity and diversity? Mol Biol Evol 23:887–892

Dash S, Van Hemert J, Hong L, Wise RP, Dickerson JA (2012) PLEXdb: gene expression resources for plants and plant pathogens. Nucleic Acids Res 40:D1194–D1201

Demain AL, Fang A (2000) The natural functions of secondary metabolites. His Mod Biotechnol I:1–39

Edger PP, Heidel-Fischer HM, Bekaert M, Rota J, Glöckner G, Platts AE, Heckel DG, Der JP, Wafula EK, Tang M et al (2015) The butterfly plant arms-race escalated by gene and genome duplications. Proc Natl Acad Sci USA 112:8362–8366

Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural populations. Nature 452:169–175

Fawcett JA, Rouzé P, Van de Peer Y (2012) Higher intron loss rate in Arabidopsis thaliana than *A. lyrata*is consistent with stronger selection for a smaller genome. Mol Biol Evol 29:849–859

Feng S, Jacobsen SE, Reik W (2010) Epigenetic reprogramming in plant and animal development. Science 330:622–627

Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants. New Phytol 183:557–564

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol 60:433–453

Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D (2008) Many or most genes in *Arabidopsis* transposed after the origin of the order Brassicales. Genome Res 18:1924–1937

Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL (2007) All duplicates are not equal: the difference between small-scale and genome duplication. Genome Biol 8:R209

Halkier BA, Gershenzon J (2006) Biology and biochemistry of glucosinolates. Annu Rev Plant Biol 57:303–333

Hanada K, Kuromori T, Myouga F, Toyoda T, Li W-H, Shinozaki K (2009) Evolutionary persistence of functional compensation by duplicate genes in *Arabidopsis*. Genome Biol Evol 1:409–414

Hudson CM, Puckett EE, Bekaert M, Pires JC, Conant GC (2011) Selection for higher gene copy number after different types of plant gene duplications. Genome Biol Evol 3:1369–1380

Kalsotra A, Cooper TA (2011) Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet 12:715–729

Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P (2011) EnsemblBioMarts: a hub for data retrieval across taxonomic space. Database

Kliebenstein DJ (2008) A role for gene duplication and natural variation of gene expression in the evolution of metabolism. PLoS ONE 3:e1838

Kopelman NM, Lancet D, Yanai I (2005) Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat Genet 37:588–589

Lareau LF, Brenner SE (2015) Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible. Mol Biol Evol 32:1072–1079

Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11:204–220

Lee TF, Zhai J, Meyers BC (2010) Conservation and divergence in eukaryoticDNA methylation. Proc Natl Acad Sci USA 107:9027–9028

Lloyd JP, Seddon AE, Moghe GD, Simenc MC, Shiu S-H (2015) Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. Plant Cell 27:2133–2147

Lynch M (2006) The origins of eukaryotic gene structure. Mol Biol Evol 23:450–468

Lynch M (2007) The origins of genome architecture. Sinauer Associates Sunderland, Massachusetts

Lynch M, Koskella B, Schaack S (2006) Mutation pressure and the evolution of organelle genomic architecture. Science 311:1727–1730

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA 102:5454–5459

Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R (2013) Gene duplication as a major force in evolution. J Genet 92:155–161

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18:50–60

Moghe GD, Last RL (2015) Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. Plant Physiol 169:1512–1523

Moore BM, Wang P, Fan P, Leong B, Schenck CA, Lloyd JP, Lehti-Shiu MD, Last RL, Pichersky E, Shiu S-H (2019) Robust predictions of specialized metabolism genes through machine learning. Proc Natl Acad Sci USA 116:2344–2353

Mueller LA, Zhang P, Rhee SY (2003) AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol 132:453–460

Mukherjee D, Mukherjee A, Ghosh TC (2016) Evolutionary rate heterogeneity of primary and secondary metabolic pathway genes in Arabidopsis thaliana. Genome Biol Evol 8:17–28

Mukherjee D, Saha D, Acharya D, Mukherjee A, Chakraborty S, Ghosh TC (2018) The role of introns in the conservation of the metabolic genes of *Arabidopsis thaliana*. Genomics 110:310–317

Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, Palsson BO (2012) Network context and selection in the evolution to enzyme specificity. Science 337:1101–1104

Parenteau J, Durand M, Morin G, Gagnon J, Lucier J-F, Wellinger RJ, Chabot B, Elela SA (2011) Introns within ribosomal protein genes regulate the production and function of yeast ribosomes. Cell 147:320–331

Pichersky E, Gang DR (2000) Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends Plant Sci 5:439–445

Qian W, Liao B-Y, Chang AY-F, Zhang J (2010) Maintenance of duplicate genes and their functional redundancy by reduced expression. Trends Genet 26:425–430

Ren X-Y, Vorst O, Fiers MWEJ, Stiekema WJ, Nap J-P (2006) In plants, highly expressed genes are the least compact. Trends Genet 22:528–532

Rizzon C, Ponger L, Gaut BS (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol 2:e115

Rose AB, Elfersi T, Parra G, Korf I (2008) Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. Plant Cell 20:543–551

Schmidt S, Sunyaev S, Bork P, Dandekar T (2003) Metabolites: a helping hand for pathway evolution? Trends Biochem Sci 28:336–341

Smith CWJ, Patton JG, Nadal-Ginard B (1989) Alternative splicing in the control of gene expression. Annu Rev Genet 23:527–577

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15:72–101

Su Z, Han L, Zhao Z (2011) Conservation and divergence of DNA methylation in eukaryotes: new insights from single base-resolution DNA methylomes. Epigenetics 6:134–140

Syed NH, Kalyna M, Marquez Y, Barta A, Brown JWS (2012) Alternative splicing in plants–coming of age. Trends Plant Sci 17:616–623

Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc Natl Acad Sci USA 110:1797–1802

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, Henikoff S (2005) DNA methylation profiling identifies CG methylation clusters in *Arabidopsis* genes. Curr Biol 15:154–159

Varadarajan N, Gam J, Olsen MJ, Georgiou G, Iverson BL (2005) Engineering of protease variants exhibiting high catalytic activity and exquisite substrate selectivity. Proc Natl Acad Sci USA 102:6855–6860

Vinogradov AE (2004) Compactness of human housekeeping genes: selection for economy or genomic design? Trends Genet 20:248–253

vonHoldt BM, Takuno S, Gaut BS (2012) Recent retrotransposon insertions are methylated and phylogenetically clustered in japonica rice (*Oryza sativa* spp. *japonica*). Mol Biol Evol 29:3193–3203

Wang Y, Wang X, Tang H, Tan X, Ficklin SP, Feltus FA, Paterson AH (2011) Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms. PLoS ONE 6:e28150

Wang Y, Tan X, Paterson AH (2013) Different patterns of gene structure divergence following gene duplication in Arabidopsis. BMC Genomics 14:652

Wang X, Zhang Z, Fu T, Hu L, Xu C, Gong L, Wendel JF, Liu B (2017) Gene-body CG methylation and divergent expression of duplicate genes in rice. Sci Rep 7:2675

Weng J-K, Noel JP (2012) The remarkable pliability and promiscuity of specialized metabolism. Cold Spring Harb Symp Quant Biol 77:309–320

Weng J-K, Philippe RN, Noel JP (2012) The rise of chemodiversity in plants. Science 336:1667–1670

Woody JL, Shoemaker RC (2011) Gene expression: sizing it all up. Front Genet 2:70

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E et al (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659

Yang H (2009) In plants, expression breadth and expression level distinctly and non-linearly correlate with gene structure. Biol Direct 4:45

Yang Y-F, Zhu T, Niu D-K (2013) Association of intron loss with high mutation rate in *Arabidopsis*: implications for genome size evolution. Genome Biol Evol 5:723–733

Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, Bhai J, Campbell LI, Martinez MC, Chakiachvili M et al (2022) Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. Nucleic Acids Res 50:D996–D1003

Yu H, Tian C, Yu Y, Jiao Y (2016) Transcriptome survey of the contribution of alternative splicing to proteome diversity in *Arabidopsis thaliana*. Mol Plant 9:749–752

Yu J, Tehrim S, Wang L, Dossa K, Zhang X, Ke T, Liao B (2017) Evolutionary history and functional divergence of the cytochrome P450 gene superfamily between *Arabidopsisthaliana* and *Brassica* species uncover effects of whole genome and tandem duplications. BMC Genomics 18:733

Zhang J (2003) Evolution by gene duplication: an update. TrendsEcolEvol 18:292–298