Evolutionary Rate Heterogeneity of Primary and Secondary Metabolic Pathway Genes in *Arabidopsis thaliana*

Dola Mukherjee[1], Ashutosh Mukherjee[2,*], and Tapash Chandra Ghosh[1,*]

[1]Bioinformatics Centre, Bose Institute, Kolkata, West Bengal, India

[2]Department of Botany, Vivekananda College, Thakurpukur, Kolkata, West Bengal, India

*Corresponding author: E-mail: ashutoshcaluniv@gmail.com; tapash@boseinst.ernet.in.

Accepted: November 5, 2015

Abstract

Primary metabolism is essential to plants for growth and development, and secondary metabolism helps plants to interact with the environment. Many plant metabolites are industrially important. These metabolites are produced by plants through complex metabolic pathways. Lack of knowledge about these pathways is hindering the successful breeding practices for these metabolites. For a better knowledge of the metabolism in plants as a whole, evolutionary rate variation of primary and secondary metabolic pathway genes is a prerequisite. In this study, evolutionary rate variation of primary and secondary metabolic pathway genes has been analyzed in the model plant *Arabidopsis thaliana*. Primary metabolic pathway genes were found to be more conserved than secondary metabolic pathway genes. Several factors such as gene structure, expression level, tissue specificity, multifunctionality, and domain number are the key factors behind this evolutionary rate variation. This study will help to better understand the evolutionary dynamics of plant metabolism.

Key words: metabolic pathway genes, effective number of codons, multifunctionality, principal component analysis.

Introduction

Being sessile organisms, plants tolerate constantly changing environments over their whole lifespan (Milo and Last 2012). To combat this, plants produce an enormous array of chemicals as unique adaptive strategies (Weng and Noel 2012; Weng 2014). These chemicals are synthesized or deconstructed by a collection of enzyme-catalyzed chemical reactions called metabolism (Weng and Noel 2012). A set of enzymes that catalyze sequential reactions in a highly concerned manner form the metabolic pathway (Weng 2014). Metabolism meets two apparently conflicting requirements: to maintain the homoeostasis necessary for a living organism and to respond dynamically to the constantly changing environment (Milo and Last 2012). Primary metabolic pathways required for the survival of plants are conserved in all living organisms (Mullins et al. 2008; Weng 2014), whereas secondary or specialized metabolic pathways have a multitude of roles in plant interaction with the environment (Zhao et al. 2013; Weng 2014). The majority of plant metabolites is secondary metabolites and has a direct effect on plant fitness (Zhao et al. 2013).

There are several differences in primary and secondary metabolic pathway enzymes. In primary metabolism, the demand for high metabolic flux forces a major selective pressure on the evolution of its enzymes (Nam et al. 2012). As a result, these enzymes show low levels of catalytic promiscuity (Bar-Even et al. 2011). On the other hand, for secondary metabolic pathway enzymes, the selection pressure is to optimize the regulation, control, and localization with the fluctuating environmental conditions rather than to increase the metabolic flux (Bar-Even et al. 2011; Weng and Noel 2012). Under such circumstances, efficiency and precision are traded for synthesis of a wider diversity of products to cope up with the spatially and temporally changing environments (Weng and Noel 2012). This makes the secondary metabolic pathway enzymes more promiscuous than primary metabolic pathway enzymes (Weng and Noel 2012). Moreover, specialized metabolic enzymes are approximately 30 times less active than primary metabolic enzymes (Bar-Even et al. 2011). Evolutionary selection pressures as well as physicochemical constrains affect enzymes (Bar-Even et al. 2011), and functional promiscuity can potentiate adaptive evolution (DePristo 2007). As chemical diversity shapes biological diversity, the principles of evolution must be relevant to chemical diversity (Firn and Jones 2009). Thus, understanding the effect of selection on genes involved in

pathways has received ample attention in the study of molecular evolution recently (Clotault et al. 2012). According to Weng (2014), metabolism offers an attractive platform to investigate evolutionary processes that lead to biological complexity. At present, the challenge in studying metabolism is to understand how evolution worked and shaped the characteristics of extant plants (Milo and Last 2012). Additionally, various plant metabolites are used for the production of dyes, medicines, flavors, insecticides, and fragrances (Verpoorte and Memelink 2002) and are thus of industrial, pharmaceutical, and agricultural interest (Zhao et al. 2013). Several primary metabolites such as starch, vitamins, and amino acids are potential candidates for metabolic engineering (Trethewey 2004). However, there is a scarcity in the comparative studies of primary and secondary metabolism (Weng and Noel 2012). Few studies have been conducted on the evolutionary rate variation of specific metabolic pathways in plants (Rausher et al. 2008; Ramsay et al. 2009) and thus involve a relatively small number of genes. Poor characterization of plant secondary metabolic pathways is a major constraint for successful molecular breeding practices (Verpoorte and Memelink 2002). Elucidation of plant metabolite biosynthesis will thus provide an expanded knowledge base and molecular tools for the genetic manipulation of biochemical pathways (Zhao et al. 2013).

It will also be of immense interest to study whether rapid enzyme evolution in plants is facilitated by other molecular machineries encoded by the plant genome (Weng 2014). Different attributes shape the evolutionary dynamics of a gene (Yang and Gaut 2011). It has been previously reported that the characteristics of gene such as gene length (Marais and Duret 2001) and intron number (Larracuente et al. 2008) correlate with both synonymous and nonsynonymous evolution. Other factors such as GC content, untranslated region (UTR) length, expression level, tissue specificity, and multifunctionality also correlate with the evolutionary rate of genes of *A. thaliana* (Yang and Gaut 2011). Protein domains are basic evolutionary units (Fong et al. 2007), and they are likely to have a highly conserved location within proteins (Pils et al. 2005). Effective number of codons (EN_c) has also been showed to modulate evolutionary rates in *Drosophila* (Han et al. 2013). Domains typically cover a majority of a protein sequence and play a crucial role in protein evolution (Toll-Riera and Albà 2013).

Primary metabolic pathways are well established (Castillo et al. 2013). Although a significant portion of the plant genome is involved in specialized metabolic pathways (Castillo et al. 2013), genomic analysis generally fails (Bocobza et al. 2012) because the taxonomically narrowly distributed pathways lack true orthologs (Castillo et al. 2013). Complete genome sequences of the model plant *Arabidopsis thaliana* (L.) Heynh. (Arabidopsis Genome Initiative 2000) have helped to resolve many problems regarding gene regulation and functional compensation (Hanada et al. 2011). However, lack of studies regarding variation of evolutionary rate between primary and secondary metabolic pathway genes was due to the absence of a closely related genome that allows accurate ortholog identification (Gaut and Ross-Ibarra 2008). However, the availability of *A. lyrata* genome (Hu et al. 2011) made it possible to correctly identify the orthologs for *A. thaliana* (Yang and Gaut 2011). Additionally, these two species have diverged approximately 13 Ma (Beilstein et al. 2010) and have approximately 80% sequence identity over whole-genome alignments (Hu et al. 2011). Thus, the study of *A. thaliana* genes with the help of *A. lyrata* orthologs will give us an accurate measure of evolutionary rate variation in primary and secondary metabolic pathway genes. Previously, genome-wide patterns of evolutionary rate variation among *A. thaliana* nuclear genes and its correlates have been studied (Yang and Gaut 2011). However, primary and secondary metabolic pathway genes should show difference in evolutionary rates as they act under different selective pressures. Hence, we have analyzed the difference in evolutionary rates and the factors that shape this variation in *A. thaliana*. We have addressed three questions. First, what is the difference in evolutionary rates between primary and secondary metabolic pathway genes? Second, what are the correlates of the evolutionary rate? Third, what is the relative contribution of these correlates in the evolutionary rate variation? Our study suggests that primary metabolic pathway genes are more conserved than secondary metabolic pathway genes of *A. thaliana*. This variation is mainly governed by gene structure, expression level, tissue specificity, multifunctionality, and domain number. The differences in nonsynonymous substitutions in the two types of pathways are mainly due to factors related to gene expression, whereas the differences in synonymous substitutions are mainly due to gene-level variations. This information is valuable for further biotechnological studies.

Materials and Methods

Data Set Preparation and Evolutionary Rate Estimation

The KEGG Orthology for *A. thaliana* (ath00001.keg) was downloaded from KEGG database (Kanehisa and Goto 2000). We have chosen all the nuclear genes of primary and secondary metabolic pathways of *A. thaliana*. A list of all these pathways is given in supplementary material S1, Supplementary Material online. We have obtained a total of 2,030 genes for primary metabolism and 482 genes for secondary metabolism. We extracted the corresponding *Arabidopsis lyrata* orthologs (with 1:1 orthology and at least 80% sequence similarity) of *A. thaliana* genes from Ensembl Plants database (Kersey et al. 2014) using Biomart (Kinsella et al. 2011) as well as obtained their pairwise nonsynonymous (d_N) and synonymous (d_S) substitution rates to compute gene-specific evolutionary rate (d_N/d_S).

These d_N and d_S values have been calculated using codeml from the PAML package (Yang 1997). Protein coding sequences of these genes were also acquired from Ensembl database. For genes with more than one isoform, the longest isoform was considered. The final data set comprised 2,030 primary metabolic and 273 secondary metabolic genes with available evolutionary rate for further analysis (supplementary material S2, Supplementary Material online).

Determination of Potential Factors of Evolutionary Rates

Several factors such as gene structure, GC content, expression profile, multifunctionality, and protein domain organization have been analyzed to determine their potential to modulate the evolutionary rate. These are studied as follows.

Gene Structure and GC Content

Gene, UTR, and coding DNA sequences (CDS) have been downloaded from Ensembl Plants database. A CDS that did not begin with an ATG start codon or did not end with the stop codon (TAG/TAA/TGA) or did not occur in multiples of three nucleotides has been discarded. We have calculated gene length, 5′- and 3′-UTR length, intron number, and average intron length as well as GC content of genes, 5′- and 3′-UTRs. GC_3 has been calculated from CDS of the genes using CodonW (http://codonw.sourceforge.net/, last accessed June 15, 2015). To measure the state of codon usage bias of the genes, we have measured EN_c (Wright 1990) using CodonW. Pfam (Finn et al. 2014) domain annotations were obtained from Ensembl Biomart against the longest peptide and number of domains per gene was counted.

Gene Expression Level and Pattern

The expression data were obtained from using Genevestigator (Hruz et al. 2008). Various microarray expression data of the *A. thaliana* (ATH1:22 k array) were obtained from Genevestigator plant biology version (https://www.genevestigator.com/gv/plant.jsp). The expression level of a gene was estimated by the average value of all the samples. The tissue specificity index τ was measured following Yanai et al. (2005) as follows:

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

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 a higher value signifying higher specificity. The index τ has been used because of its advantage over using expression breadth as reported previously (Liao and Zhang 2006). We have also collected Plant Ontology (PO) (Avraham et al. 2008) data for each gene to better understand the expression of a particular gene in different plant structures and developmental stages. The PO data have been obtained from Ensembl Plants (http://plants.ensembl.org/index.html).

Function

According to Gene Ontology (GO) Slim annotations that classify proteins to obtain a high-level view of functions (Prachumwat and Li 2006), the multifunctionality of a gene has been assessed by counting the number of biological processes in which a gene takes part. The GO slim accessions were obtained from Ensembl Biomart (Kinsella et al. 2011).

Statistical Analyses

Statistical analyses were performed using SPSS v.13. Mann–Whitney U test (Mann and Whitney 1947) was used to compare the average values of different variables between two classes of genes as the values were not normally distributed in our data set. For correlation analysis, we performed the Spearman's rank correlation coefficient ρ (Spearman 1904), where the significant correlations were denoted by $P < 0.05$. For relative contribution analysis of each factor to evolutionary rate, a principal component analysis (PCA) was performed.

Results and Discussion

The Variation of Evolutionary Rates between Primary and Secondary Metabolic Pathway Genes

This study clearly showed that primary metabolic pathway genes are more conserved than secondary metabolic pathway genes in *A. thaliana*. d_N, d_S, and d_N/d_S were calculated for 1,035 primary and 241 secondary metabolic pathway genes. Frequency distributions of these three parameters are shown in figure 1. The average values of d_N, d_S, and d_N/d_S were significantly (Mann–Whitney U test, $P < 0.01$ in all cases) different in primary and secondary metabolic pathway genes (table 1). The frequency distribution of d_N, d_S, and d_N/d_S was also different in two types of pathways (fig. 1). Highest frequency (around 37%) of genes in primary metabolic pathways showed a d_N value of approximately 0.01, whereas the highest frequency (around 31%) of genes in secondary metabolic pathways showed a d_N value of approximately 0.02. Considering d_S values, around 10% of primary metabolic pathway genes showed a d_S value of approximately 0.14, whereas around 12% of secondary metabolic pathway genes showed a d_S value of approximately 0.14. Highest frequency (around 5.5%) of genes in primary metabolic pathways showed a d_N/d_S value of approximately 0.08, whereas a highest frequency (around 7.46%) of genes in secondary metabolic pathways showed a d_N/d_S value of approximately 0.13. The synonymous (d_N) and nonsynonymous (d_S) substitution rates were 1.3 and 1.06 times greater in secondary metabolic pathway genes, respectively. We also found that no gene in our data set showed d_N/d_S value >1, which indicate that, on average, genes involved in primary and secondary metabolic

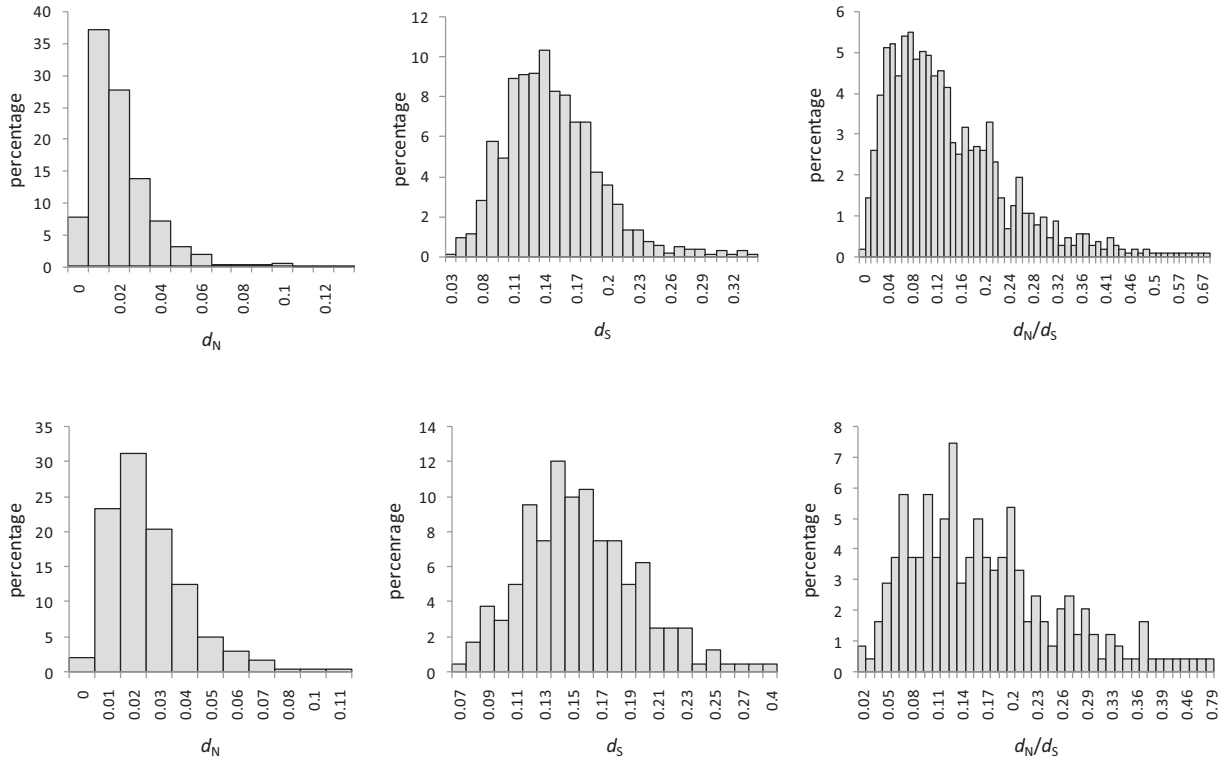

Fig. 1.—Distribution of d_N, d_S, and d_N/d_S in primary (upper panel) and secondary (lower panel) metabolic pathway genes in *A. thaliana*.

Table 1
Evolutionary Rates for Primary and Secondary Metabolic Pathway Genes (*P* Values Were Obtained by Mann–Whitney *U* Test)

	Primary	Secondary	*P* value
d_N			
Mean (SD)	0.02 (0.016)	0.026 (0.016)	1.73×10^{-10}
CV	0.789	0.629	
Range	0.0008–0.135	0.002–0.111	
d_S			
Mean (SD)	0.147 (0.044)	0.157 (0.043)	2.5×10^{-4}
CV	0.305	0.276	
Range	0.031–0.371	0.072–0.399	
d_N/d_S			
Mean (SD)	0.142 (0.103)	0.170 (0.102)	1.19×10^{-6}
CV	0.725	0.599	
Range	0.003–0.855	0.018–0.791	

pathways in *A. thaliana* are not under positive selection. In general, genes under positive selection are rare in *A. thaliana* genome (Yang and Gaut 2011). It is also noteworthy that d_N and d_S values were highly positively correlated (Spearman's rank correlation $\rho = 0.246 \times 10^{-6}$). It suggests the effect of common evolutionary mechanisms on both synonymous and nonsynonymous sites of which at least one is shared mutation rates (Yang and Gaut 2011). The positive correlation of synonymous and nonsynonymous substitutions has also been found in *Drosophila* (Comeron and Kreitman 1998), indicating that synonymous substitutions are not independent of selective constraints acting on the amino acid level (Dunn et al. 2001).

The higher rate of evolution confers an advantage to genes involved in secondary metabolism. As the enzymes of secondary metabolism exhibit high plasticity (Khersonsky et al. 2006), a few mutations can increase the promiscuous activity. It has been shown that 10^4- to 10^6-fold improvement in enzyme plasticity has been achieved in response to a single mutation (Khersonsky et al. 2006), and this functional promiscuity may potentiate adaptive evolution (DePristo 2007). As primary metabolic pathway enzymes are directly involved in the core functioning of a plant, they show vanishingly low levels of promiscuity (Weng and Noel 2012). This explains that higher evolutionary rate of the genes involved in secondary metabolism gives the plant a selective advantage in the ever changing environment, whereas conserved nature of primary metabolic pathway genes assures the integrity of the core functioning. This finding also has implications in biotechnology, especially protein engineering. Enzyme-catalyzed industrial processes are increasing in various fields ranging from food processing to produce small molecule pharmaceuticals (Gustafsson et al. 2012). Our results suggest that, in case of secondary metabolic pathway genes of *A. thaliana*, protein engineering can result in the formation of new and novel metabolites that may be advantageous for the plant (as they accumulate more

synonymous and nonsynonymous substitutions as well as more promiscuous than primary metabolic pathway genes). Indeed, several carotenoid biosynthetic enzymes such as synthases, desaturases, cyclases, and oxygenases have been altered for both substrate specificity and reaction selectivity by single amino acid substitutions to produce a plethora of novel carotenoids (Umeno et al. 2005; Tracewell and Arnold 2009). On the other hand, altering the highly conserved primary metabolic pathway genes can disrupt the normal functioning of the enzyme. Chen et al. (2010) showed that in a α-amylase (an enzyme from primary metabolic pathway) from *Bacillus* sp. strain TS-23, a mutation of Asp-234 and Asp-236 of conserved sequence region V (CSR-V) resulted in 33% and 86% reduction in specific activity (Chen et al. 2010). However, in another α-amylase from *Anoxybacillus* species, replacement of Ala-161 of CSR-V with an aspartic acid increased the specific activity (Ranjani et al. 2014). Thus, it is advisable that creating any mutation in primary metabolic genes should be performed with great care to retain the functionality of the protein.

Effect of Different Factors on the Evolutionary Rate Variation among the Primary and Secondary Metabolic Pathway Genes in *A. thaliana*

One of the most important objectives of molecular evolution studies is to understand the factors that influence genetic variation in the genome (Clotault et al. 2012). Effect of mutation and selection has different effect on synonymous and nonsynonymous substitutions (Yang and Nielsen 1998). Although the occurrence of a synonymous mutation is assumed to have no effect on the fitness of the individual (Zhou et al. 2012), selection on synonymous sites has been shown to be associated with mRNA secondary structure and stability (Duan et al. 2003; Chamary and Hurst 2005; Stoletzki 2008; Gu et al. 2010) as well as protein expression (Zhou et al. 2010). Moreover, several gene properties can affect the mutation rate or the local selection environment, both of which encompass protein evolutionary rates (Chang and Liao 2013). If the influence on protein evolutionary rates was at the selection level, its correlation to d_N and d_N/d_S should not differ considerably (Chang and Liao 2013). Considering all the above factors, we have measured the effect of several factors on d_N, d_S, and d_N/d_S. Many factors such as gene length (Stoletzki and Eyre-Walker 2007), gene compactness (Liao et al. 2006; Chang and Liao 2013), GC content (Ratnakumar et al. 2010), codon usage bias (Sharp and Li 1987; Urrutia and Hurst 2001; Yang and Gaut 2011), expression level (Pál et al. 2001; Subramanian and Kumar 2004; Rocha 2006), and tissue specificity (Larracuente et al. 2008; Slotte et al. 2011) have been shown to influence evolutionary rate. However, the relative importance of determinants for protein evolutionary rates varies widely among various taxa. For example, in yeasts, predominant factor determining the rate of protein evolution was found to be mRNA abundance (Drummond et al. 2006), whereas in mammals, gene compactness was found to have a stronger influence on protein evolutionary rates compared with the abundance of mRNA (Liao et al. 2006, 2010). Besides, coding sequence length showed the strongest positive correlation with protein evolutionary rates in flagellated algae (Chang and Liao 2013). We have thus focused on various parameters that could have affected the evolutionary rate variation in genes of primary and secondary metabolic pathways in *A. thaliana*.

Effect of Gene Length and Gene Compactness

Gene length has been previously shown to have a negative correlation with evolutionary rate in *A. thaliana* (Yang and Gaut 2011). In this study, the average length of genes in the primary metabolic pathways (2,762.75 bp, $N = 1,035$) is significantly (Mann–Whitney U test, $P = 3.1 \times 10^{-4}$) higher than secondary metabolic pathway genes (2,439.63 bp, $N = 241$). We have found a significant negative correlation between gene length and d_N, d_S, and d_N/d_S (table 2). Previously, Slotte et al. (2011) reported no correlation between d_N/d_S and gene length when analyzing all genes in the *A. thaliana* genome. However, Yang and Gaut (2011) reported a strong negative correlation between gene length and d_N, d_S, and d_N/d_S. We have also studied the effect of different attributes of gene compactness, such as UTR length, intron length, and intron number on the evolutionary rate of primary and secondary metabolic pathway genes. Average length of 5′-UTR in the primary metabolic pathways (128.60 bp, $N = 1,035$) is significantly (Mann–Whitney U test, $P = 8.9 \times 10^{-8}$) higher than secondary metabolic pathway genes (96.49 bp, $N = 241$). Average length of 3′-UTR in the primary metabolic pathways (231.06 bp, $N = 1,035$) is significantly (Mann–Whitney U test, $P = 7.5 \times 10^{-6}$) higher than secondary metabolic pathway genes (192.87 bp, $N = 241$). Both 5′- and 3′-UTR lengths were found to be negatively significantly correlated with d_N, d_S, and d_N/d_S (Spearman's rank correlation, $P < 0.001$) (table 2). This is in accordance with the results obtained for all genes in *A. thaliana* (Yang and Gaut 2011). Our results showed that UTR length plays a significant role in the evolutionary rate variation in *A. thaliana*.

Average intron length was not found to be significantly different in these two types of pathways (Mann–Whitney U test, $P > 0.05$). However, intron number was found to be significantly different between the two types of pathways (6.17, $N = 976$ and 4.30, $N = 222$, respectively, for primary and secondary metabolic pathways, Mann–Whitney U test, $P = 1.86 \times 10^{-6}$). Intron number was significantly correlated with d_N and d_S, but not with d_N/d_S (table 2). Thus, it was found that primary metabolic pathway genes were longer and less compact than secondary metabolic pathway genes, and these features are responsible for their evolutionary rate heterogeneity in *A. thaliana*. Earlier reports showed that shorter and

intron-poor genes have either more variable (Jeffares et al. 2008; Lawniczak et al. 2008) or stronger (Castillo-Davis et al. 2002; Chiaromonte et al. 2003; Ren et al. 2006) expression levels. Regulatory responses can be delayed by introns and are selected against genes whose transcripts require rapid adjustment for survival of environmental challenges (Jeffares et al. 2008). Indeed, secondary metabolic genes that are generally expressed in response of environmental challenges are shorter as determined in our study. It is noteworthy that correlation of intron number was higher with d_S than d_N. To better understand this variation, we have divided the data set into three groups: Intronless genes, genes with upto ten introns, and genes with more than ten introns. We have found no significant difference in d_S between intronless genes of primary and secondary metabolic pathway genes (fig. 2). This was also found in genes with more than ten introns. However, d_N was found to be significantly different in all three groups. However, when intron number increases to more than 10, the rate of synonymous substitutions became similar. This later observation is unclear at the moment. However, one reason behind this may be the very small number of secondary metabolic genes with more than ten introns (only 26), which probably gave a biased result here. However, in each group, nonsynonymous substitutions were significantly higher in primary than secondary metabolic genes. All these results show that introns significantly accumulate more synonymous substitutions in primary than secondary metabolic genes. Primary metabolic pathway genes contain significantly higher domain number (1.53 per gene, $N = 1{,}034$; Mann–Whitney U test, $P = -5.51 \times 10^{-12}$) than secondary metabolic pathway genes (1.26, $N = 240$). We have also analyzed the proportion of single, double, and multidomain proteins in the two types of pathways. In total, 62.41% and 75.98% of all proteins were single domain proteins in primary and secondary metabolic pathways, respectively, and they are significantly different (Z score $= -5.62$, $P < 0.01$). However, primary metabolic pathways significantly contain more number of double and multidomain proteins (fig. 3). There is a strong negative correlation between domain number per gene and evolutionary rates (table 2). Domain number was found to show higher correlation coefficient with d_N than d_S.

Effect of GC Content Variation

GC content of the genes in the two types of pathways was not significantly different (Mann–Whitney U test, $P = 0.675$). However, GC content of UTRs was significantly different. Average GC content of 5′-UTR in the primary metabolic pathways (34.27%, $N = 1{,}035$) is significantly (Mann–Whitney U test, $P = 9.3 \times 10^{-9}$) higher than that in the secondary metabolic pathway genes (31.00%, $N = 241$). Average GC content of 3′-UTR in the primary metabolic pathways (29.58 bp, $N = 1{,}035$) is significantly (Mann–Whitney U test, $P = 8.8 \times 10^{-7}$) higher than secondary metabolic pathway genes (27.42 bp, $N = 241$). Both 5′- and 3′-UTR GC contents were significantly negatively correlated with d_N, d_S, and d_N/d_S (Spearman's rank correlation, $P < 0.001$) (table 2). Notably, we have also found that UTR GC content is significantly positively correlated with UTR length (Spearman's $\rho_{5'\text{-UTR length vs.}}$

Table 2
Spearman's Rank Correlations of Evolutionary Rates with Potentially Contributing Factors

Variables	Spearman's ρ (P value)		
	d_N	d_S	d_N/d_S
Gene structure and compactness			
Gene length	−0.170 (1.0×10^{-6})***	−0.199 (1.0×10^{-6})***	−0.96 (9.9×10^{-4})***
Intron number	−0.08 (4.3×10^{-3})**	−0.248 (1.0×10^{-6})***	0.011 (NS)
5′-UTR length	−0.277 (1.0×10^{-6})***	−0.176 (1.0×10^{-6})***	−0.214 (1.0×10^{-6})***
3′-UTR length	−0.326 (1.0×10^{-6})***	−0.158 (1.0×10^{-6})***	−0.277 (1.0×10^{-6})***
GC content			
5′-UTR	−0.170 (1.0×10^{-6})***	−0.167 (1.0×10^{-6})***	−0.103 (4.1×10^{-4})***
3′-UTR	−0.204 (1.0×10^{-6})***	−0.145 (1.0×10^{-6})***	−0.151 (1.0×10^{-6})***
GC_3	−0.092 (1.6×10^{-3})**	0.05 (NS)	−0.111 (1.4×10^{-4})***
Domain number	−0.1 (6.1×10^{-4})***	−0.073 (1.2×10^{-2})*	−0. 073 (1.2×10^{-2})*
Expression			
Expression level	−0.461 (1.0×10^{-6})***	−0.151 (1.0×10^{-6})***	−0.405 (1.0×10^{-6})***
Tissue specificity	0.241 (1.0×10^{-6})***	0.198 (1.0×10^{-6})***	0.170 (1.0×10^{-6})***
ENc	0.151 (1.0×10^{-6})***	−0.031 (NS)	0.168 (1.0×10^{-6})***
PO	−0.392 (1.0×10^{-6})***	−0.163 (1.0×10^{-6})***	−0.338 (1.0×10^{-6})***
Multifunctionality			
GO slim	−0.359 (1.0×10^{-6})***	−0.142 (1.09×10^{-6})***	−0.308 (1.0×10^{-6})***

NOTE.—NS, not significant. * $P<0.05$, ** $P<0.01$, *** $P < 0.001$.

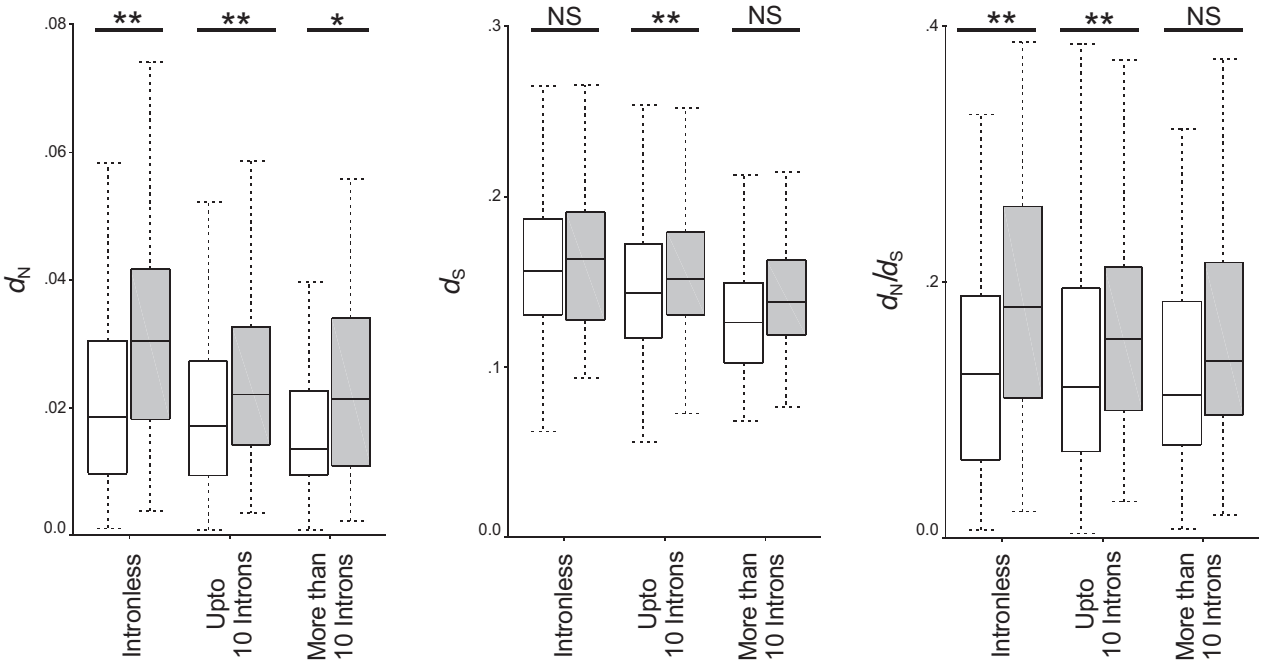

Fig. 2.—Boxplots showing distribution of d_N, d_S, and d_N/d_S in primary (white boxes) and secondary (gray boxes) metabolic pathway genes in three groups according to the number of introns in *A. thaliana*. NS, not significant, * $P<0.05$, ** $P<0.01$.

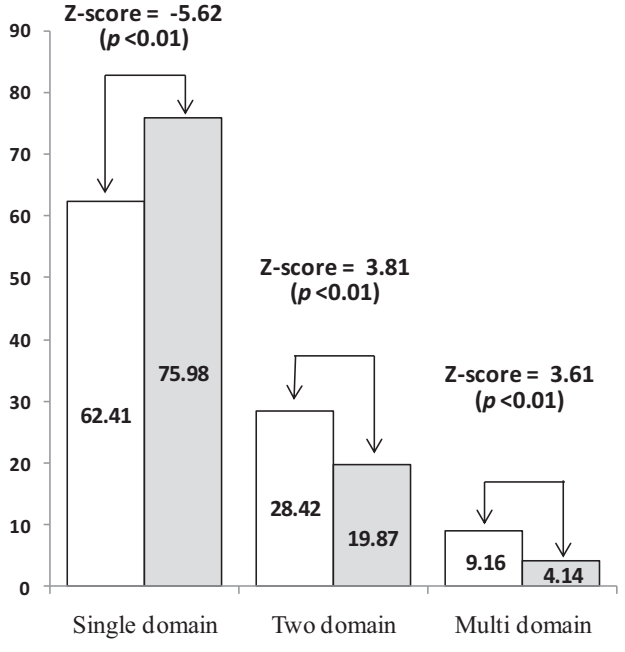

Fig. 3.—The bar diagram depicts the percentage of single domain, two domain, and multidomain proteins within primary and secondary metabolic pathway genes in *A. thaliana*. In each group, the white bar represents primary metabolic pathway genes whereas the gray bar belongs to secondary metabolic pathway genes.

$_{GC\ content} = 0.371$, $P = 1.0 \times 10^{-6}$ and Spearman's $\rho_{3'\text{-UTR length vs. GC content}} = 0.608$, $P = 1.0 \times 10^{-6}$). It is known that UTR regions are crucial for posttranscriptional regulation of gene expression (Mignone et al. 2002). In human, it was found that genes located in large GC-rich regions of a chromosome (heavy isochores) have shorter UTRs than genes located in GC-poor isochores (Mignone et al. 2002). Moreover, in vertebrates, it has been proposed that most housekeeping genes should be located in GC-rich isochores, whereas tissue-specific genes should be located in GC-poor isochores (Bernardi 2000). However, secondary metabolic pathway genes are more tissue specific and have shorter UTRs than primary metabolic pathway genes, as found in this study, which contradicts the situation in humans and vertebrates. Average GC_3 content of primary metabolic pathway genes (0.412, $N = 1{,}024$) was significantly lower (Mann–Whitney U test, $P = 3 \times 10^{-3}$) than secondary metabolic pathway genes (0.423, $N = 237$). GC_3 was significantly negatively correlated with d_N and d_N/d_S but not d_S (table 2). It has been shown that recombination is a driving force for the increase in GC_3 in many organisms (Tatarinova et al. 2010; Elhaik and Tatarinova 2012, p. 3). Although self-pollination in *Arabidopsis* keeps its recombination rates low, which ultimately results in reduced GC_3 content, evolutionary pressure would selectively keep high recombination rates for some genes (Elhaik and Tatarinova 2012). After GC_3 richness evolves in those genes under selective pressure, its additional transcriptional advantage is achieved (Elhaik and Tatarinova

2012). Moreover, Straussman et al. (2009) described a correlation between methylation and GC_3. GC_3-rich genes provide more targets for de novo methylation that can serve as an additional mechanism of transcriptional regulation that ultimately increases adaptability to a species under external stresses (Elhaik and Tatarinova 2012). The higher GC_3 content of secondary metabolic genes in our study supports this view. As these genes provide a selective advantage to the plant under changing environmental conditions, the GC_3 content became higher than primary metabolic pathway genes.

Effect of Gene Expression

Average expression level of primary metabolic pathway genes (9,025.77, $N = 1{,}004$) was approximately 2.03 fold of secondary metabolic pathway genes (4,430.29, $N = 235$) and the difference was significant (Mann–Whitney U test, $P = 1.79 \times 10^{-7}$). The tissue specificity ($\tau = 0.287$, $N = 235$) of secondary metabolic pathway genes was significantly (Mann–Whitney U test, $P = 1.17 \times 10^{-12}$) higher than primary metabolic pathway genes ($\tau = 0.236$, $N = 1{,}008$). Both expression level and tissue specificity were significantly correlated with d_N, d_S, and d_N/d_S (table 2). EN_c of primary metabolic pathway genes (53.18, $N = 1{,}024$) was also significantly lower (Mann–Whitney U test, $P = 2.97 \times 10^{-7}$) than secondary metabolic pathway genes (54.47, $N = 237$), indicating that primary metabolic pathway genes show more codon biasness than secondary metabolic pathway genes. We have also studied the relationship of expression level and tissue specificity with GC_3 and EN_c. Both GC_3 and EN_c showed a significant correlation with expression level (Spearman's $\rho_{\text{expression level vs. ENc}} = -0.184$, $P = 1.0 \times 10^{-6}$; Spearman's $\rho_{\text{expression level vs. GC3}} = 0.138$, $P = 3.3 \times 10^{-6}$). GC_3 and EN_c were significantly correlated with tissue specificity (Spearman's $\rho_{\tau \text{ vs. ENc}} = -0.121$, $P = 4.6 \times 10^{-5}$; Spearman's $\rho_{\tau \text{ vs. GC3}} = 0.105$, $P = 3.81 \times 10^{-4}$). EN_c was significantly positively correlated with d_N and d_N/d_S but not d_S (table 2). There is a strong positive correlation between GC_3 and EN_c (Spearman's $\rho_{\text{ENc vs. GC3}} = 0.322$, $P = 1.0 \times 10^{-6}$). It was reported that mRNA secondary-structure stability is correlated with both GC content and codon usage (Gu et al. 2010). Moreover, the common causes of heterologous gene expression are mainly associated with the disparities in codon bias, mRNA secondary structure and stability, gene product toxicity, and product solubility (Makrides 1996; Gustafsson et al. 2004). Hence, it is clear from this study that expressing secondary metabolic genes in another host such as bacteria or simple eukaryotes is easier than primary metabolic pathway genes. For heterologous expression of the latter, additional methods such as codon optimization (Angov 2011) or codon harmonization (Angov et al. 2008) may be required. For a better knowledge about the expression of genes with respect to different structures and developmental stages of the plant, we have also studied the PO terms (Avraham et al. 2008). The PO database uses 71 plant structure development stages and 326 plant anatomical entities to describe *Arabidopsis* gene expression patterns (Cooper et al. 2013). We have counted the number of PO terms per gene and correlated that with evolutionary rate. Primary metabolic pathway genes are involved in significantly more PO terms (28.25 per gene, $N = 1{,}035$; Mann–Whitney U test, $P = 7.44 \times 10^{-12}$) than secondary metabolic pathway genes (22.48 per gene, $N = 241$). There is a strong negative correlation between PO terms per gene and evolutionary rates (table 2).

Multifunctionality

Primary metabolic pathway genes were found to be significantly more multifunctional (12.95 per gene, $N = 1{,}035$; Mann–Whitney U test, $P = 1.72 \times 10^{-12}$) than secondary metabolic pathway genes (10.91 per gene, $N = 241$) (fig. 4). There is a strong negative correlation between GO slim terms per gene and evolutionary rates (table 2). It shows that primary metabolic genes are evolutionarily more conserved due to

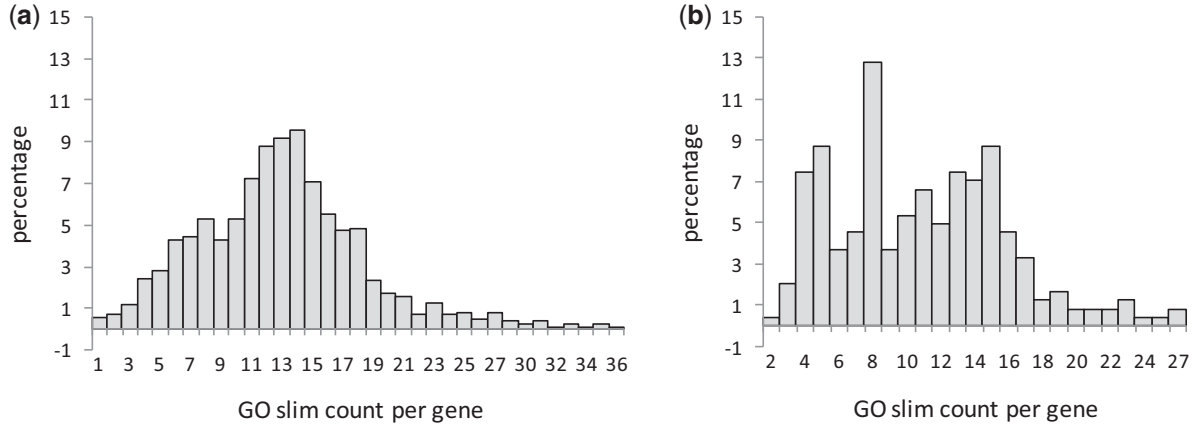

Fig. 4.—Distribution of GO slim numbers per gene in (*a*) primary and (*b*) secondary metabolic pathway genes in *A. thaliana*.

their higher multifunctionality. It was also found that highest number (12.86%) of genes of the secondary metabolic pathways take part in eight biological processes whereas 9.56% genes of the primary metabolic pathways take part in 14 biological processes. Multifunctionality, that is, involvement of a protein in several processes has been shown by several enzymes such as hexokinase, triose-phosphate isomerase, enolase, etc. and the presence of multifunctional proteins increases the metabolic efficiency of a cell (Schwab 2003). This study shows that primary metabolic pathway genes are more multifunctional than genes involved in secondary metabolism. Hence, metabolic efficiency is primarily maintained by genes involved in primary metabolism. In yeast, it has been found that multifunctional proteins evolve at a slower rate (Salathé et al. 2006). It was also revealed in this study that multifunctionality has greater effect on d_N than d_S. However, our analysis shows that the magnitude of the effect of multifunctionality on d_N, d_S, and d_N/d_S of metabolic pathways is higher than its effect on whole genome of *A. thaliana* as studied by Yang and Gaut (2011). This shows that multifunctionality has greater effect on metabolic genes than other proteins of the genome. It is also notable that both expression level and tissue specificity are significantly correlated with multifunctionality (Spearman's $\rho_{\text{expression level vs. multifunctionality}} = 0.547$, $P = 1.0 \times 10^{-6}$; Spearman's $\rho_{\text{tissue specificity vs. multifunctionality}} = -0.308$, $P = 1.0 \times 10^{-6}$). In PCA, the first principal component includes expression parameters and multifunctionality. This contradicts with the result of Salathé et al. (2006) who have not found any significant correlation of multifunctionality with expression level. However, the reason behind the higher expression and lower tissue specificity of multifunctional genes of metabolic pathways is not clear yet. One possible explanation is that secondary metabolic pathway genes evolve in response to specific environmental factors and thus less multifunctional and more tissue specific. On the other hand, genes involved in primary metabolism are more multifunctional as this enhances metabolic efficiency. They are also ubiquitously expressed in the plant body at a higher level to successfully maintain the core functioning of the plant body. Number of domains per gene was significantly higher in genes involved in primary metabolic pathway genes than secondary metabolic pathway genes. It was also found that domain number is negatively correlated with evolutionary rate. The reason behind this correlation was not clear. Then, we correlated domain number with other parameters to investigate whether it is correlated with other factors. Domain number was significantly correlated with multifunctionality (Spearman's $\rho_{\text{domain numbers multifunctionality}} = 0.153$, $P = 1.0 \times 10^{-6}$) as well as gene length (Spearman's $\rho_{\text{domain numbers gene length}} = 0.358$, $P = 1.0 \times 10^{-6}$), intron length (Spearman's $\rho_{\text{domain numbers intron length}} = 0.188$, $P = 1.0 \times 10^{-6}$), and GC_3 (Spearman's $\rho_{\text{domain numbers GC3}} = -0.123$, $P = 7.0 \times 10^{-5}$). In PCA, domain number along with gene length, intron number, and GC_3 was included in principal component 2. Thus, it seems that domain number is more of a function of gene character rather than multifunctionality.

Relative Involvement of the Factors in Shaping Evolutionary Rate Variation among Primary and Secondary Metabolic Pathway Genes

To elucidate the covariance structure of different factors, we have performed PCA. Results of PCA analysis are given in table 3. Components with values more than 0.5 have been retained. It has been observed that the first two principal components have explained 43.312% of the total variance. The major contributors of the first component were PO, 3′-UTR GC content, 5′-UTR GC content, expression level, tissue specificity, 3′-UTR length, and multifunctionality (table 3). The major contributors of the second principal component were gene length, intron number, GC_3, and domain number. It is noteworthy that the first component showed higher correlation coefficient with d_N than d_S and the second component showed the reverse. The contributors of the first coefficient mainly comprised factors related to gene expression and the second coefficient mainly comprised gene-level factors. Hence, it is apparent that the difference in nonsynonymous substitutions in the two types of pathways is mainly due to factors related to gene expression, whereas the difference in

Table 3

PCA on d_N, d_S, and d_N/d_S and Major Contributors of the Principal Components

	Principal Component 1	Principal Component 2
Percent of the total variance	26.164	17.148
Correlation coefficient (Spearman's ρ) with d_N	−0.439 ($P = 1.0 \times 10^{-6}$)	−0.013 ($P = 6.8 \times 10^{-1}$)
Correlation coefficient (Spearman's ρ) with d_S	−0.207 ($P = 1.0 \times 10^{-6}$)	−0.158 ($P = 1.0 \times 10^{-6}$)
Correlation coefficient (Spearman's ρ) with d_N/d_S	−0.361 ($P = 1.0 \times 10^{-6}$)	−0.044 ($P = 1.5 \times 10^{-1}$)
Major contributors		
PO	0.740	
3′-UTR GC content	0.754	
5′-UTR GC content	0.689	
3′-UTR length	0.586	
GO slim	0.581	
Tissue specificity	−0.660	
Expression level	0.533	
Gene length		0.869
Intron number		0.857
GC_3		−0.616
Domain number		0.567

synonymous substitutions is due to gene-level variations such as length, intron number, and domain number. However, the inclusion of GC_3 content in the second component seems to be strange as GC_3 has effect on the regulation of gene expression by methylation as discussed earlier. We have performed Spearman correlation of GC_3 with gene length, intron number, and domain number. Indeed, GC_3 content is significantly negatively correlated with intron number (Spearman's $\rho = -0.499$, $P = 1.0 \times 10^{-6}$), gene length (Spearman's $\rho = -0.326$, $P = 1.0 \times 10^{-6}$), and domain number (Spearman's $\rho = -0.122$, $P = 7.1 \times 10^{-5}$). Moreover, it has been found in corn that genes with high GC_3 tend to be mono-exonic (Alexandrov et al. 2009). We have also found a similar trend in both primary and secondary metabolic pathway genes in *A. thaliana* (fig. 5). Thus, it is clear that intron number is highly correlated with GC_3 content. Probably, the correlation is due to the accumulation of more synonymous mutations in intronic regions, which also explains the significant negative correlation with gene length. As genes of the primary metabolic pathways are longer than secondary metabolic pathway genes and their coding sequence length is not significantly different, it is, thus, evident that the increased gene length in primary metabolic genes is mainly contributed by introns, and thus gene length shows a significant negative correlation with GC_3. However, the correlation between GC_3 and domain number is not very clear. The d_N/d_S ratio showed higher correlation coefficient with component 1 than component 2. Thus, it is clear that the evolutionary rate difference between primary and secondary metabolic pathway genes is chiefly due to the factors related to expression than gene-level predictors in *A. thaliana*.

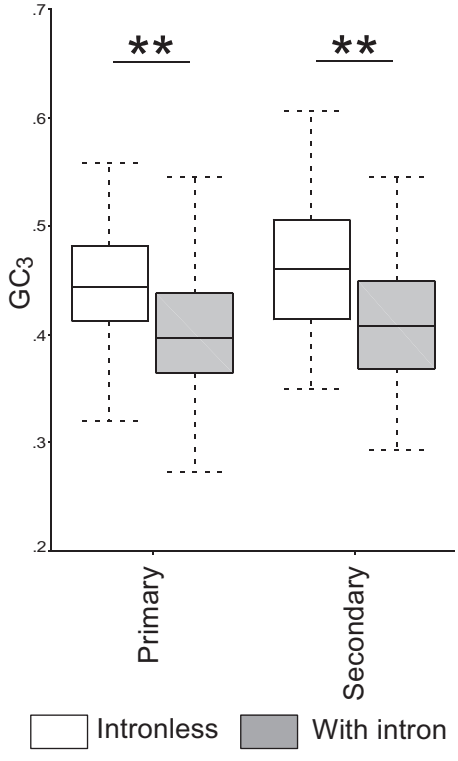

Fig. 5.—Comparison of GC_3 content between intronless genes and genes with introns in primary and secondary metabolic pathway genes of *A. thalaiana*. The intronless genes showed significantly higher GC_3 content than genes with intron (Mann–Whitney *U* test, $^{**}P < 0.01$).

Conclusion

This study showed that primary metabolic pathway genes are evolutionary more conserved than secondary metabolic pathway genes in *A. thaliana*. The effect of different gene level expression level and protein level factors showed that gene length, gene compactness, expression level, tissue specificity, multifunctionality, and domain number are the major contributors for the evolutionary rate difference of these primary and secondary metabolic pathway genes. To the best of our knowledge, this is the first extensive comparison of primary and secondary metabolic pathway genes from an evolutionary perspective. Improving the agronomic quality of a crop by altering its metabolic signature by targeted breeding can be an important tool for a breeder. The knowledge gathered from this study can play a pivotal role for this kind of breeding practices. As secondary metabolic pathway genes are less conserved, their intra- or interspecific variation should be greater, and this variation can be a starting point for transgenic manipulation or targeted breeding. Moreover, as these genes tend to accumulate more substitutions, protein engineering by site-directed mutagenesis can lead to the formation of a plethora of new economically important metabolites. On the other hand, when targeting a primary metabolic gene, emphasis should not be to alter its coding sequence as this can disrupt the function of these highly conserved genes that will ultimately affect the plant phenotype. Rather, for improved production of primary metabolites, factors that affect gene expression such as UTRs or other regulatory elements may be altered. Alternately, for successful heterologous expression, codon optimization or codon harmonization may be beneficial. Our study, thus, provides valuable information on the evolutionary aspects of primary and secondary metabolism in *A. thaliana* which, along with further laboratory-based experimental studies, can be helpful for metabolic engineering and production of improved plant varieties in the near future.

Supplementary Material

Supplementary materials S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors are thankful to Dr Maria Costantini, associate editor, and the two anonymous referees for their valuable comments in improving the manuscript. This study was financially supported by the Department of Biotechnology, Government of India to D.M.

Literature Cited

Alexandrov NN, et al. 2009. Insights into corn genes derived from large-scale cDNA sequencing. Plant Mol Biol. 69:179–194.
Angov E. 2011. Codon usage: nature's roadmap to expression and folding of proteins. Biotechnol J. 6:650–659.
Angov E, Hillier CJ, Kincaid RL, Lyon JA. 2008. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. PLoS One 3:e2189.
Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815.
Avraham S, et al. 2008. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res. 36:D449–D454.
Bar-Even A, et al. 2011. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry (Mosc.) 50:4402–4410.
Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. 2010. Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 107:18724–18728.
Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17.
Bocobza S, Willmitzer L, Raikhel NV, Aharoni A. 2012. Discovery of new modules in metabolic biology using chemometabolomics. Plant Physiol. 160:1160–1163.
Castillo DA, Kolesnikova MD, Matsuda SPT. 2013. An effective strategy for exploring unknown metabolic pathways by genome mining. J Am Chem Soc. 135:5885–5894.
Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. Nat Genet. 31:415–418.
Chamary J-V, Hurst LD. 2005. Biased codon usage near intron-exon junctions: selection on splicing enhancers, splice-site recognition or something else? Trends Genet 21:256–259.
Chang T-Y, Liao B-Y. 2013. Flagellated algae protein evolution suggests the prevalence of lineage-specific rules governing evolutionary rates of eukaryotic proteins. Genome Biol Evol. 5:913–922.
Chen Y-H, et al. 2010. Mutational analysis of the proposed calcium-binding aspartates of a truncated α-amylase from *Bacillus* sp. strain TS-23. Ann Microbiol. 60:307–315.
Chiaromonte F, Miller W, Bouhassira EE. 2003. Gene length and proximity to neighbors affect genome-wide expression levels. Genome Res. 13:2602–2608.
Clotault J, Peltier D, Soufflet-Freslon V, Briard M, Geoffriau E. 2012. Differential selection on carotenoid biosynthesis genes as a function of gene position in the metabolic pathway: a study on the carrot and dicots. PLoS One 7:e38724.
Comeron JM, Kreitman M. 1998. The correlation between synonymous and nonsynonymous substitutions in *Drosophila*: mutation, selection or relaxed constraints? Genetics 150:767–775
Cooper L, et al. 2013. The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol. 54 (2):e1.
DePristo MA. 2007. The subtle benefits of being promiscuous: adaptive evolution potentiated by enzyme promiscuity. HFSP J. 1:94–98.
Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327–337.
Duan J, et al. 2003. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet. 12:205–216.
Dunn KA, Bielawski JP, Yang Z. 2001. Substitution rates in *Drosophila* nuclear genes: implications for translational selection. Genetics 157:295–305.
Elhaik E, Tatarinova T. 2012. GC_3 Biology in Eukaryotes and Prokaryotes. In: DNA methylation—from genomics to technology. In: Tatarinova T, editor. InTech. Available from: http://www.intechopen.com/books/dna-methylation-from-genomics-to-technology/gc3-biology-in-eukaryotes-and-prokaryotes.
Finn RD, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–D230.
Firn RD, Jones CG. 2009. A Darwinian view of metabolism: molecular properties determine fitness. J Exp Bot. 60:719–726.
Fong JH, Geer LY, Panchenko AR, Bryant SH. 2007. Modeling the evolution of protein domain architectures using maximum parsimony. J Mol Biol. 366:307–315.
Gaut BS, Ross-Ibarra J. 2008. Selection on major components of angiosperm genomes. Science 320:484–486.
Gustafsson C, et al. 2012. Engineering genes for predictable protein expression. Protein Expr Purif. 83:37–46.
Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. Trends Biotechnol. 22:346–353.
Gu W, Zhou T, Wilke CO. 2010. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol. 6 (2):e1000664.
Hanada K, et al. 2011. Functional compensation of primary and secondary metabolites by duplicate genes in *Arabidopsis thaliana*. Mol Biol Evol. 28:377–382.
Han M, et al. 2013. Evolutionary rate patterns of genes involved in the *Drosophila* Toll and Imd signaling pathway. BMC Evol Biol. 13:245.
Hruz T, et al. 2008. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinforma. 2008:420747.
Hu TT, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43:476–481.
Jeffares DC, Penkett CJ, Bähler J. 2008. Rapidly regulated genes are intron poor. Trends Genet. 24:375–378.
Kanehisa M, Goto S. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.
Kersey PJ, et al. 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res. 42:D546–D552.
Khersonsky O, Roodveldt C, Tawfik D. 2006. Enzyme promiscuity: evolutionary and mechanistic aspects. Curr Opin Chem Biol. 10:498–508.
Kinsella RJ, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030.
Larracuente AM, et al. 2008. Evolution of protein-coding genes in *Drosophila*. Trends Genet. 24:114–123.
Lawniczak MK, Holloway AK, Begun DJ, Jones CD. 2008. Genomic analysis of the relationship between gene expression variation and DNA polymorphism in *Drosophila simulans*. Genome Biol. 9:R125.
Liao B-Y, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol. 23:2072–2080.
Liao B-Y, Weng M-P, Zhang J. 2010. Impact of extracellularity on the evolutionary rate of mammalian proteins. Genome Biol Evol. 2:39–43.
Liao B-Y, Zhang J. 2006. Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution. Mol Biol Evol. 23:1119–1128.
Makrides SC. 1996. Strategies for achieving high-level expression of genes in *Escherichia coli*. Microbiol Rev. 60:512–538.
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.

Marais G, Duret L. 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J Mol Evol. 52:275–280.
Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. Genome Biol. 3:1–10.
Milo R, Last RL. 2012. Achieving diversity in the face of constraints: lessons from metabolism. Science 336:1663–1667.
Mullins EA, Francois JA, Kappock TJ. 2008. A specialized citric acid cycle requiring succinyl-coenzyme A (CoA):acetate CoA-transferase (AarC) confers acetic acid resistance on the acidophile *Acetobacter aceti*. J Bacteriol. 190:4933–4940.
Nam H, et al. 2012. Network context and selection in the evolution to enzyme specificity. Science 337:1101–1104.
Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.
Pils B, Copley RR, Schultz J. 2005. Variation in structural location and amino acid conservation of functional sites in protein domain families. BMC Bioinformatics 6:1–10.
Prachumwat A, Li W-H. 2006. Protein function, connectivity, and duplicability in yeast. Mol Biol Evol. 23:30–39.
Ramsay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. Mol Biol Evol. 26:1045–1053.
Ranjani V, et al. 2014. Protein engineering of selected residues from conserved sequence regions of a novel *Anoxybacillus* a-amylase. Sci Rep. 4:5850.
Ratnakumar A, et al. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans R Soc B Biol Sci. 365:2571–2580.
Rausher MD, Lu Y, Meyer K. 2008. Variation in constraint versus positive selection as an explanation for evolutionary rate variation among anthocyanin genes. J Mol Evol. 67:137–144.
Ren X-Y, Vorst O, Fiers MW, Stiekema WJ, Nap J-P. 2006. In plants, highly expressed genes are the least compact. Trends Genet. 22:528–532.
Rocha EP. 2006. The quest for the universals of protein evolution. Trends Genet. 22:412–416.
Salathé M, Ackermann M, Bonhoeffer S. 2006. The effect of multifunctionality on the rate of evolution in yeast. Mol Biol Evol. 23:721–722.
Schwab W. 2003. Metabolome diversity: too few genes, too many metabolites? Phytochemistry 62:837–849.
Sharp PM, Li W-H. 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol. 4:222–230.
Slotte T, et al. 2011. Genomic determinants of protein evolution and polymorphism in *Arabidopsis*. Genome Biol Evol. 3:1210–1219.
Spearman C. 1904. "General Intelligence," objectively determined and measured. Am J Psychol. 15:201–292.
Stoletzki N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC Evol Biol. 8:224.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. Mol Biol Evol. 24:374–381.
Straussman R, et al. 2009. Developmental programming of CpG island methylation profiles in the human genome. Nat Struct Mol Biol. 16:564–571.
Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373–381.
Tatarinova TV, Alexandrov NN, Bouck JB, Feldmann KA. 2010. GC3 biology in corn, rice, sorghum and other grasses. BMC Genomics 11:308.
Toll-Riera M, Albà MM. 2013. Emergence of novel domains in proteins. BMC Evol Biol. 13:47.
Tracewell CA, Arnold FH. 2009. Directed enzyme evolution: climbing fitness peaks one amino acid at a time. Curr Opin Chem Biol. 13:3–9.
Trethewey RN. 2004. Metabolite profiling as an aid to metabolic engineering in plants. Curr Opin Plant Biol. 7:196–201.
Umeno D, Tobias AV, Arnold FH. 2005. Diversifying carotenoid biosynthetic pathways by directed evolution. Microbiol Mol Biol Rev. 69: 51–78.
Urrutia AO, Hurst LD. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. Genetics 159:1191–1199.
Verpoorte R, Memelink J. 2002. Engineering secondary metabolite production in plants. Curr Opin Biotechnol. 13:181–187.
Weng J-K. 2014. The evolutionary paths towards complexity: a metabolic perspective. New Phytol. 201:1141–1149.
Weng J-K, Noel JP. 2012. The remarkable pliability and promiscuity of specialized metabolism. Cold Spring Harb Symp Quant Biol. 77: 309–320.
Wright F. 1990. The "effective number of codons" used in a gene. Gene 87:23–29.
Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21:650–659.
Yang L, Gaut BS. 2011. Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. Mol Biol Evol. 28:2359–2369.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46:409–418.
Zhao N, Wang G, Norris A, Chen X, Chen F. 2013. Studying plant secondary metabolism in the age of genomics. Crit Rev Plant Sci. 32: 369–382.
Zhou T, et al. 2012. Non-silent story on synonymous sites in voltage-gated ion channel genes. PLoS One 7(10):e48541.
Zhou T, Gu W, Wilke CO. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol. 27:1912–1922.

Associate editor: Maria Costantini