Molecular Phylogenetics and Evolution 120 (2018) 33–42

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Evolutionary history of the cobalamin-independent methionine synthase gene family across the land plants

Hugo Vianna Silva Rody, Luiz Orlando de Oliveira*

Department of Biochemistry and Molecular Biology, Universidade Federal de Viçosa, Viçosa, Minas Gerais 36570-900, Brazil

ARTICLE INFO

Keywords:
Biased gene retention
Molecular evolution
Polyploidy
Positive selection
Whole-genome duplication

ABSTRACT

Plants are successful paleopolyploids. The wide diversity of land plants is driven strongly by their gene duplicates undergoing distinct evolutionary fates after duplication. We used genomic resources from 35 model plant species to unravel the evolutionary fate of gene copies (paralogs) of the cobalamin-independent methionine synthase (metE) gene family across the land plants. To explore genealogical relationships and characterize positive selection as a driving force in the evolution of metE paralogs within a single species, we carried out complementary analyses on genomic data of 32 genotypes of soybean. The size of the metE gene family remained small across the land plants; most of the studied species possessed 1–6 paralogs. Gene products were either cytosolic or chloroplastic; this dual subcellular distribution arose early during the divergence of the land plants and reached all extant lineages. Biased gene loss and gene retention events took place multiple times; recurrent evolution remodeled redundant metE paralogs to recover and maintain the dual subcellular distribution of MetE. Shared whole-genome duplication events gave rise to the metE paralogs of both soybean and *Medicago truncatula*. In soybean, the ancestral paralog pair GlymaPP2A encoded a cytosolic isoform of MetE, was under strong purifying selection, and retained high levels of expression across eight RNA-seq expression libraries. The daughters GlymaPP1 and GlymaPP2B showed accelerated rates of evolution, accumulated many sites predicted to be under positive selection, and possessed low levels of expression. Our results suggest that the metE paralogs of soybean follow Ohno's neofunctionalization model of gene duplicate evolution.

1. Introduction

Polyploidy is a remarkable mechanism of duplication for promoting the diversification of plant species (Simillion et al., 2002; Sémon and Wolfe, 2007). Indeed, higher plants are paleopolyploids; that is, they are descended from an ancient polyploid ancestor (Bowers et al., 2003; Jaillon et al., 2007). The polyploids are intriguing; they survived the naturally deleterious process of chromosome doubling, have overcome their rarity in the population, and outcompeted their parental species (Otto and Whitton, 2000). After an event of whole-genome duplication (WGD), the resulting gene copies (paralogs) may undergo distinct fates depending on the direction and intensity of the several evolutionary forces that operate after duplication (Edger and Pires, 2009; Freeling, 2009). According to Ohno's (1970) model, a single gene copy is sufficient to maintain proper gene function in an organism; the other paralogs would be redundant and might undergo relaxed purifying selection, which in turn would favor neutral evolution and the accumulation of random mutations. Occasionally, these random mutations could provide the daughter paralog with a new, advantageous role (neofunctionalization); eventually, the daughter paralog would be fixed in the population through positive selection. In contrast, if random mutations hamper gene function in such a way that the daughter paralog becomes disadvantageous, the outcome would be pseudogenization; the malfunctional, daughter paralog would be lost from the genome. In another possible scenario, both parental and daughter paralogs of a given gene would cooperate and share the ancestral function (subfunctionalization). Beyond the work of Ohno, substantial efforts have been made to unravel the fate of duplicated gene pairs, especially after the recent advances in next-generation sequencing (NGS) technology. Underlying biological attributes, such as the mechanism of duplication (Birchler et al., 2001; Papp et al., 2003; Veitia, 2002), increase in dosage benefits (Kondrashov and Koonin, 2004), metabolic flux relevance (Hudson et al., 2011), and domestication (Corbi et al., 2011) may shape the fate of gene duplicates. Investigating the evolution of a small gene family may provide insights into the fate of gene duplicates and may corroborate the validity of current gene models within the current genomic framework.

The methionine synthase gene family is an interesting model for

* Corresponding author at: Department of Biochemistry and Molecular Biology, Av. P. H. Rolfs s/n, Universidade Federal de Viçosa, 36570-000 Viçosa, MG, Brazil.
E-mail address: luiz.ufv@hotmail.com (L.O.d. Oliveira).

https://doi.org/10.1016/j.ympev.2017.12.003
Received 10 November 2017; Accepted 4 December 2017
Available online 06 December 2017

investigating the fate of duplicated gene pairs across the land plants. Over the course of evolution, two nonhomologous enzymes, cobalamin-dependent methionine synthase (MetH) and cobalamin-independent methionine synthase (MetE), acquired methionine synthase activity; these enzymes share no sequence similarity and possess divergent dependence on cobalamin (vitamin B_{12}) (Helliwell et al., 2014). Currently, mammals, protists, and most bacteria retain MetH, while green plants, fungi, and some bacteria possess MetE (Bromke and Hesse, 2015). Although some extant species of green algae, such as *Chlamydomonas reinhardtii*, possess both MetH and MetE, multiple, independent gene losses have been postulated to explain the absence of metE from the genome of many lineages of green algae (Helliwell et al., 2011). Methionine (Met) is a sulfur-containing amino acid. As a building block, Met drives the metabolism of proteins; as a component of the co-factor S-adenosylmethionine (SAM), it plays a crucial role in many essential metabolic pathways as a donor of C1 units to SAM-dependent methyltransferases (Hesse et al., 2004). In plants, the *de novo* synthesis of Met from the precursor *O*-phosphohomoserine requires only three enzymes: cystathionine γ-synthase, cystathionine β-lyase, and MetE (Ravanel et al., 2004). MetE (5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase; EC 2.1.1.14) catalyzes Met biosynthesis by the direct transfer of a methyl group from N5-methyl-5,6,7,8-tetrahydrofolate to *L*-homocysteine (Hcy) in a reaction that does not require vitamin B_{12} (cobalamin) as the cofactor (Ravanel et al., 2004). MetE also participates in regeneration of the methyl group of the co-factor SAM after a methylation reaction (Ravanel et al., 2004).

The evolutionary history of the metE gene family in higher plants remains unexplored. In *Arabidopsis thaliana*, three variants of metE have been reported (Hesse et al., 2004; Ravanel et al., 2004) and doublets have been found in *Ammi majus* and *Petroselinum crispum* (Eichel et al., 1995). Other studies have suggested that metE is a low-copy-number gene in potato (Zeh et al., 2002) and soybean (Hesse et al., 2004). Early studies have suggested that *Arabidopsis thaliana* possesses MetE isoforms with dual subcellular localization; that is, each isoform is present in either the cytosol or the chloroplast (Ravanel et al., 2004). Crystal structures of MetE of phylogenetically unrelated organisms, such as higher plants (*Arabidopsis thaliana*; Ferrer et al., 2004), thermophilic bacteria (*Thermotoga maritima*; Pejchal and Ludwig, 2005), and fungi (*Neurospora crassa*; Wheatley et al., 2016), reveal striking structural and sequence similarities: MetE is a monomeric protein with two domains, each containing an $(\alpha\beta)_8$ barrel.

Genome-wide sequence data of high quality from model plant species are now available in public databanks; these include data for species representative of all major clades of the land plants. Moreover, the resequencing of several genomes of a single plant species, such as soybean (*Glycine max*) (Lam et al., 2010), may provide crucial information for understanding the late evolution of duplicate gene pairs. Soybean is a paleopolyploid species; the soybean genome is the result of a hexaploidization event that occurred at the origin of the eudicots (Jaillon et al., 2007), in addition to at least two more recent WGD events (Shoemaker et al., 2006). About 4,500 years ago, soybean was domesticated in China. Since then, this crop species has been under intense, artificial selection and plant breeding to combine superior agronomic traits possessed by different genotypes. A collection of 31 accessions of soybean, including both wild and domesticated genotypes, had their entire genome resequenced (Lam et al., 2010). Soybeans and its close relative *Medicago truncatula* began to diverge during a WGD event that took place ~44 million years ago (Pfeil et al., 2005); its complete genome has also been sequenced (Young et al., 2011).

The study reported herein explored the genomic resources of 34 species of land plants, the green alga *Chlamydomonas reinhardtii* (Chlorophyta), together with NGS data from the resequencing of 31 genotypes of soybean to investigate the evolutionary history of duplicate gene pairs of the metE gene family across the land plants. The following four questions were addressed. (1) How ubiquitous is the dual (cytosolic/chloroplastic) subcellular distribution of MetE? (2) How large is the metE gene family? (3) What was the likely duplication mechanism that gave rise to the metE gene family? (4) What evolutionary forces shaped the evolution of the metE paralogs? Answering these questions should shed light on the evolutionary history of the metE gene family and provide new insights into how biased gene loss and gene retention events shaped the diversification of land plants.

2. Materials and methods

2.1. Assembling orthologs and paralogs of metE

Protein sequences and coding-DNA sequences (CDSs) of *Chlamydomonas reinhardtii* (Chlorophyta), *Physcomitrella patens* (Bryophyta), *Amborella trichopoda* (Amborellales), seven species of monocots, and 25 species of eudicots (Table S1) were obtained from PLAZA (Proost et al., 2015). Gene annotations for soybean were downloaded from Phytozome v10.3 (Goodstein et al., 2012). The sequence of MetE of *Arabidopsis thaliana* (Protein Data Bank ID # 1U1H) was used as a query in BLAST searches (Altschul et al., 1990) for orthologous genes against each of the protein databases we had obtained from PLAZA. During searches, BLAST retained only primary non-redundant sequences that showed similarity > 75% to the query sequence, with a minimum alignment length of 480 bp, which is ~60% of the length of metE of *Arabidopsis thaliana*. The CDS of each metE ortholog was retrieved from PLAZA and aligned using MUSCLE (Edgar, 2004) to create DatasetA (n = 104; 6585 bp), which contained 104 aligned CDSs of metE orthologs across 34 species of land plants, in addition to a single sequence of *Chlamydomonas reinhardtii*. We also assembled DatasetB (n = 9; 2443 bp), which included nine CDSs of two species (soybean and *Medicago truncatula*) of Papilionoideae, a subfamily within the Fabaceae.

We investigated the predicted subcellular location of each metE ortholog using TargetP v1.1 (Emanuelsson et al., 2007), with DatasetA as the input file. The software calculates scores to predict the presence of N-terminal pre-sequences, such as chloroplast transit peptide, mitochondrial targeting peptide, or secretory pathway signal peptide. According to TargetP, the highest score indicates the most likely location. In TargetP, the parameters were set as follows: plant organism group, no cut-offs, and cleavage site prediction enabled.

2.2. Bayesian phylogeny

Bayesian phylogeny was used to predict/ the phylogenetic relationships among orthologs and paralogs of metE. DatasetA was input into MRMODELTEST v2 (Nylander, 2004) and the Akaike Information Criterion suggested GTR + I + G as the best-fit model among 24 models of molecular evolution. The Bayesian phylogenetic analysis was performed in MRBAYES v3.2 (Huelsenbeck and Ronquist, 2001) using two simultaneous runs of 2 million generations each, with one cold and seven heated chains in each run; the temperature parameter was set to 0.25. Trees were sampled once every 2,000 generations. The sequence of the orthologous metE of *Chlamydomonas reinhardtii* was used as an outgroup. TRACER 1.5 (Drummond et al., 2012) was used to evaluate the effective sample size values for the combined simultaneous runs; these values were well above 200 for most of the statistics. MRBAYES results showed that the average standard deviation of split frequencies at the end of each run was < 0.01. Thus, the selected settings for the Bayesian phylogeny ensured that sufficient sampling of the posterior occurred. The first 250 trees were discarded as burn-in samples. A 50%-majority-rule consensus tree of the two independent runs was obtained with posterior probabilities (PPs) that were equal to bipartition frequencies. The consensus tree was visualized using FigTree v1.4 (http://tree.bio.ed.ac.uk/, last accessed 23/02/16).